“Frontmatter”
Mechanical Engineering Handbook
Ed. Frank Kreith
Boca Raton: CRC Press LLC, 1999

Contents

Sandor, B.I.; Roloff, R; et. al. "Mechanics of Solids"
Mechanical Engineering Handbook
Ed. Frank Kreith
Boca Raton: CRC Press LLC, 1999

Mechanics of Solids

Bela I. Sandor
University of Wisconsin-Madison

Ryan Roloff
Allied Signal Aerospace

Stephen M. Birn
Allied Signal Aerospace

Maan H. Jawad
Nooter Consulting Services

Michael L. Brown
A.O. Smith Corp.

1.1 Introduction

Bela I. Sandor

Engineers use the concepts and methods of mechanics of solids in designing and evaluating tools, machines, and structures, ranging from wrenches to cars to spacecraft. The required educational background for these includes courses in statics, dynamics, mechanics of materials, and related subjects. For example, dynamics of rigid bodies is needed in generalizing the spectrum of service loads on a car, which is essential in defining the vehicle's deformations and long-term durability. In regard to structural

FIGURE 1.1.1 Artist's concept of a moving stainless steel roadway to drive the suspension system through a spinning, articulated wheel, simulating three-dimensional motions and forces. (MTS Systems Corp., Minneapolis, MN. With permission.) Notes: Flat-Trac® Roadway Simulator, R&D100 Award-winning system in 1993. See also Color Plate 1.*

integrity and durability, the designer should think not only about preventing the catastrophic failures of products, but also of customer satisfaction. For example, a car with gradually loosening bolts (which is difficult to prevent in a corrosive and thermal and mechanical cyclic loading environment) is a poor product because of safety, vibration, and noise problems. There are sophisticated methods to assure a product's performance and reliability, as exemplified in Figure 1.1.1. A similar but even more realistic test setup is shown in Color Plate 1.

It is common experience among engineers that they have to review some old knowledge or learn something new, but what is needed at the moment is not at their fingertips. This chapter may help the reader in such a situation. Within the constraints of a single book on mechanical engineering, it provides overviews of topics with modern perspectives, illustrations of typical applications, modeling to solve problems quantitatively with realistic simplifications, equations and procedures, useful hints and reminders of common errors, trends of relevant material and mechanical system behaviors, and references to additional information.

The chapter is like an emergency toolbox. It includes a coherent assortment of basic tools, such as vector expressions useful for calculating bending stresses caused by a three-dimensional force system on a shaft, and sophisticated methods, such as life prediction of components using fracture mechanics and modern measurement techniques. In many cases much more information should be considered than is covered in this chapter.

* Color Plates 1 to 16 follow page 1-131.

1.2 Statics

Bela I. Sandor

Vectors. Equilibrium of Particles. Free-Body Diagrams

Two kinds of quantities are used in engineering mechanics. A scalar quantity has only magnitude (mass, time, temperature, …). A vector quantity has magnitude and direction (force, velocity, ...). Vectors are represented here by arrows and bold-face symbols, and are used in analysis according to universally applicable rules that facilitate calculations in a variety of problems. The vector methods are indispensable in three-dimensional mechanics analyses, but in simple cases equivalent scalar calculations are sufficient.

Vector Components and Resultants. Parallelogram Law

A given vector **F** may be replaced by two or three other vectors that have the same net effect and representation. This is illustrated for the chosen directions *m* and *n* for the components of **F** in two dimensions (Figure 1.2.1). Conversely, two concurrent vectors **F** and **P** of the same units may be combined to get a resultant **R** (Figure 1.2.2).

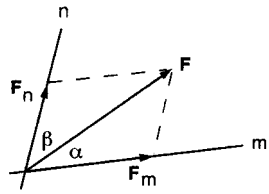


FIGURE 1.2.1 Addition of concurrent vectors **F** and **P**.

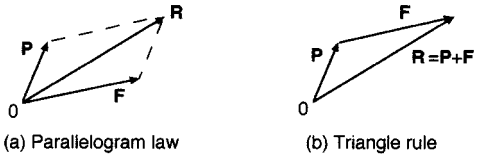


FIGURE 1.2.2 Addition of concurrent, coplanar vectors **A**, **B**, and **C**.

Any set of components of a vector **F** must satisfy the *parallelogram law*. According to Figure 1.2.1, the law of sines and law of cosines may be useful.

$$\frac{F_n}{\sin\alpha} = \frac{F_m}{\sin\beta} = \frac{F}{\sin\left[180^\circ - (\alpha+\beta)\right]}$$
$$F^2 = F_n^2 + F_m^2 - 2F_nF_m\cos\left[180^\circ - (\alpha+\beta)\right] \qquad (1.2.1)$$

Any number of concurrent vectors may be summed, mathematically or graphically, and in any order, using the above concepts as illustrated in Figure 1.2.3.

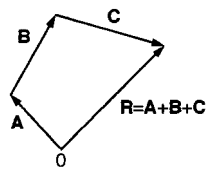


FIGURE 1.2.3 Addition of concurrent, coplanar vectors **A**, **B**, and **C**.

Unit Vectors

Mathematical manipulations of vectors are greatly facilitated by the use of unit vectors. A unit vector **n** has a magnitude of unity and a defined direction. The most useful of these are the unit coordinate vectors **i**, **j**, and **k** as shown in Figure 1.2.4.

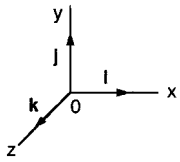


FIGURE 1.2.4 Unit vectors in Cartesian coordinates (the same **i**, **j**, and **k** set applies in a parallel $x'y'z'$ system of axes).

The three-dimensional components and associated quantities of a vector **F** are shown in Figure 1.2.5. The unit vector **n** is collinear with **F**.

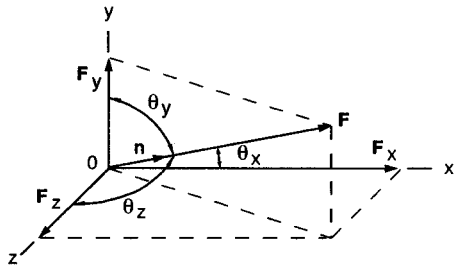


FIGURE 1.2.5 Three-dimensional components of a vector **F**.

The vector **F** is written in terms of its scalar components and the unit coordinate vectors,

$$\mathbf{F} = F_x\mathbf{i} + F_y\mathbf{j} + F_z\mathbf{k} = F\mathbf{n} \tag{1.2.2}$$

where

$$F_x = F\cos\theta_x \quad F_y = F\cos\theta_y \quad F_z = F\cos\theta_z$$

$$F = \sqrt{F_x^2 + F_y^2 + F_z^2}$$

$$n_x = \cos\theta_x \quad n_y = \cos\theta_y \quad n_z = \cos\theta_z$$

$$n_x^2 + n_y^2 + n_z^2 = 1$$

$$\frac{n_x}{F_x} = \frac{n_y}{F_y} = \frac{n_z}{F_z} = \frac{1}{F}$$

The unit vector notation is convenient for the summation of concurrent vectors in terms of scalar or vector components:

Scalar components of the resultant **R**:

$$R_x = \sum F_x \qquad R_y = \sum F_y \qquad R_z = \sum F_z \tag{1.2.3}$$

Vector components:

$$\mathbf{R}_x = \sum \mathbf{F}_x = \sum F_x \mathbf{i} \qquad \mathbf{R}_y = \sum \mathbf{F}_y = \sum F_y \mathbf{j} \qquad \mathbf{R}_z = \sum \mathbf{F}_z = \sum F_z \mathbf{k} \tag{1.2.4}$$

Vector Determination from Scalar Information

A force, for example, may be given in terms of its magnitude F, its sense of direction, and its line of action. Such a force can be expressed in vector form using the coordinates of any two points on its line of action. The vector sought is

$$\mathbf{F} = F_x \mathbf{i} + F_y \mathbf{j} + F_z \mathbf{k} = F\mathbf{n}$$

The method is to find $\mathbf{n}$ on the line of points $A(x_1, y_1, z_1)$ and $B(x_2, y_2, z_2)$:

$$\mathbf{n} = \frac{\text{vector A to B}}{\text{distance A to B}} = \frac{d_x \mathbf{i} + d_y \mathbf{j} + d_z \mathbf{k}}{\sqrt{d_x^2 + d_y^2 + d_z^2}}$$

where $d_x = x_2 - x_1$, $d_y = y_2 - y_1$, $d_z = z_2 - z_1$.

Scalar Product of Two Vectors. Angles and Projections of Vectors

The scalar product, or dot product, of two concurrent vectors $\mathbf{A}$ and $\mathbf{B}$ is defined by

$$\mathbf{A} \cdot \mathbf{B} = AB\cos\phi \tag{1.2.5}$$

where A and B are the magnitudes of the vectors and ϕ is the angle between them. Some useful expressions are

$$\mathbf{A} \cdot \mathbf{B} = \mathbf{B} \cdot \mathbf{A} = A_x B_x + A_y B_y + A_z B_z$$

$$\phi = \arccos \frac{A_x B_x + A_y B_y + A_z B_z}{AB}$$

The projection F' of a vector $\mathbf{F}$ on an arbitrary line of interest is determined by placing a unit vector $\mathbf{n}$ on that line of interest, so that

$$F' = \mathbf{F} \cdot \mathbf{n} = F_x n_x + F_y n_y + F_z n_z$$

Equilibrium of a Particle

A particle is in **equilibrium** when the resultant of all forces acting on it is zero. In such cases the algebraic summation of rectangular scalar components of forces is valid and convenient:

$$\sum F_x = 0 \qquad \sum F_y = 0 \qquad \sum F_z = 0 \tag{1.2.6}$$

Free-Body Diagrams

Unknown forces may be determined readily if a body is in equilibrium and can be modeled as a particle. The method involves **free-body diagrams**, which are simple representations of the actual bodies. The appropriate model is imagined to be isolated from all other bodies, with the significant effects of other bodies shown as force vectors on the free-body diagram.

Example 1

A mast has three guy wires. The initial tension in each wire is planned to be 200 lb. Determine whether this is feasible to hold the mast vertical (Figure 1.2.6).

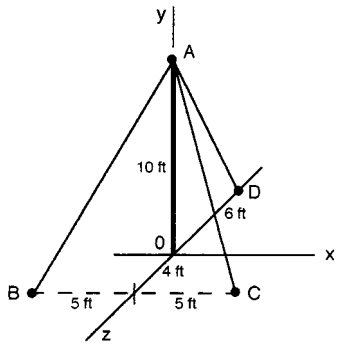


FIGURE 1.2.6 A mast with guy wires.

Solution.

$$\mathbf{R} = \mathbf{T}_{AB} + \mathbf{T}_{AC} + \mathbf{T}_{AD}$$

The three tensions of known magnitude (200 lb) must be written as vectors.

$$\mathbf{T}_{AB} = (\text{tension } AB)(\text{unit vector } A \text{ to } B) = 200 \text{ lb } \mathbf{n}_{AB} = 200 \text{ lb } \frac{\left(d_x\mathbf{i} + d_y\mathbf{j} + d_z\mathbf{k}\right)}{d}$$

$$= \frac{200 \text{ lb}}{\sqrt{5^2 + 10^2 + 4^2}}\left(-5\mathbf{i} - 10\mathbf{j} + 4\mathbf{k}\right)\frac{\text{ft}}{\text{ft}} = -84.2 \text{ lb } \mathbf{i} - 168.4 \text{ lb } \mathbf{j} + 67.4 \text{ lb } \mathbf{k}$$

$$\mathbf{T}_{AC} = \frac{200 \text{ lb}}{11.87 \text{ ft}}\left(5\mathbf{i} - 10\mathbf{j} + 4\mathbf{k}\right) \text{ ft} = 84.2 \text{ lb } \mathbf{i} + 168.4 \text{ lb } \mathbf{j} + 67.4 \text{ lb } \mathbf{k}$$

$$\mathbf{T}_{AD} = \frac{200 \text{ lb}}{11.66 \text{ ft}}\left(0\mathbf{i} - 10\mathbf{j} + 6\mathbf{k}\right) \text{ ft} = -171.5 \text{ lb } \mathbf{j} - 102.9 \text{ lb } \mathbf{k}$$

The resultant of the three tensions is

$$\mathbf{R} = \sum F_x\mathbf{i} + \sum F_y\mathbf{j} + \sum F_z\mathbf{k} = (-84.2 + 84.2 + 0) \text{ lb } \mathbf{i} + (-168.4 - 168.4 - 171.5) \text{ lb } \mathbf{j}$$

$$+ (67.4 + 67.4 - 102.9) \text{ lb } \mathbf{k} = 0 \text{ lb } \mathbf{i} - 508 \text{ lb } \mathbf{j} + 31.9 \text{ lb } \mathbf{k}$$

There is a horizontal resultant of 31.9 lb at *A*, so the mast would not remain vertical.

Forces on Rigid Bodies

All solid materials deform when forces are applied to them, but often it is reasonable to model components and structures as rigid bodies, at least in the early part of the analysis. The forces on a rigid body are generally not concurrent at the center of mass of the body, which cannot be modeled as a particle if the force system tends to cause a rotation of the body.

Moment of a Force

The turning effect of a force on a body is called the moment of the force, or torque. The moment M_A of a force **F** about a point A is defined as a scalar quantity

$$M_A = Fd \tag{1.2.7}$$

where d (the moment arm or lever arm) is the nearest distance from A to the line of action of **F**. This nearest distance may be difficult to determine in a three-dimensional scalar analysis; a vector method is needed in that case.

Equivalent Forces

Sometimes the equivalence of two forces must be established for simplifying the solution of a problem. The necessary and sufficient conditions for the equivalence of forces **F** and **F′** are that they have the same magnitude, direction, line of action, and moment on a given rigid body in static equilibrium. Thus,

$$\mathbf{F} = \mathbf{F}' \quad \text{and} \quad M_A = M'_A$$

For example, the ball joint A in Figure 1.2.7 experiences the same moment whether the vertical force is pushing or pulling downward on the yoke pin.

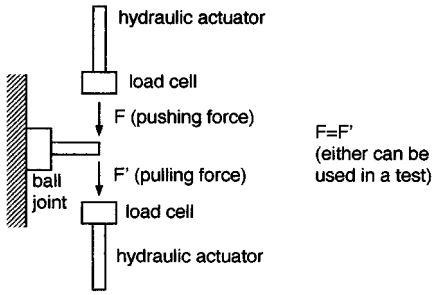


FIGURE 1.2.7 Schematic of testing a ball joint of a car.

Vector Product of Two Vectors

A powerful method of vector mechanics is available for solving complex problems, such as the moment of a force in three dimensions. The *vector product* (or cross product) of two concurrent vectors **A** and **B** is defined as the vector $\mathbf{V} = \mathbf{A} \times \mathbf{B}$ with the following properties:

1. **V** is perpendicular to the plane of vectors **A** and **B**.
2. The sense of **V** is given by the right-hand rule (Figure 1.2.8).
3. The magnitude of **V** is $V = AB \sin\theta$, where θ is the angle between **A** and **B**.
4. $\mathbf{A} \times \mathbf{B} \neq \mathbf{B} \times \mathbf{A}$, but $\mathbf{A} \times \mathbf{B} = -(\mathbf{B} \times \mathbf{A})$.
5. For three vectors, $\mathbf{A} \times (\mathbf{B} + \mathbf{C}) = \mathbf{A} \times \mathbf{B} + \mathbf{A} \times \mathbf{C}$.

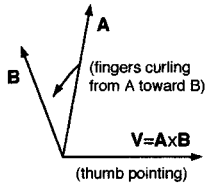


FIGURE 1.2.8 Right-hand rule for vector products.

The vector product is calculated using a determinant,

$$\mathbf{V} = \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ A_x & A_y & A_z \\ B_x & B_y & B_z \end{vmatrix} = A_y B_z \mathbf{i} + A_z B_x \mathbf{j} + A_x B_y \mathbf{k} - A_y B_x \mathbf{k} - A_x B_z \mathbf{j} - A_z B_y \mathbf{i} \tag{1.2.8}$$

Moment of a Force about a Point

The vector product is very useful in determining the moment of a force **F** about an arbitrary point *O*. The vector definition of moment is

$$\mathbf{M}_O = \mathbf{r} \times \mathbf{F} \tag{1.2.9}$$

where **r** is the position vector from point *O* to any point on the line of action of **F**. A double arrow is often used to denote a moment vector in graphics.

The moment $\mathbf{M}_O$ may have three scalar components, M_x, M_y, M_z, which represent the turning effect of the force **F** about the corresponding coordinate axes. In other words, a single force has only one moment about a given point, but this moment may have up to three components with respect to a coordinate system,

$$\mathbf{M}_O = M_x \mathbf{i} + M_y \mathbf{j} + M_z \mathbf{k}$$

Triple Products of Three Vectors

Two kinds of products of three vectors are used in engineering mechanics. The *mixed triple product* (or scalar product) is used in calculating moments. It is the dot product of vector **A** with the vector product of vectors **B** and **C**,

$$\mathbf{A} \cdot (\mathbf{B} \times \mathbf{C}) = \begin{vmatrix} A_x & A_y & A_z \\ B_x & B_y & B_z \\ C_x & C_y & C_z \end{vmatrix} = A_x \left(B_y C_z - B_z C_y \right) + A_y \left(B_z C_x - B_x C_z \right) + A_z \left(B_x C_y - B_y C_x \right) \tag{1.2.10}$$

The *vector triple product* $(\mathbf{A} \times \mathbf{B}) \times \mathbf{C} = \mathbf{V} \times \mathbf{C}$ is easily calculated (for use in dynamics), but note that

$$(\mathbf{A} \times \mathbf{B}) \times \mathbf{C} \neq \mathbf{A} \times (\mathbf{B} \times \mathbf{C})$$

Moment of a Force about a Line

It is common that a body rotates about an axis. In that case the moment M_ℓ of a force F about the axis, say line ℓ, is usefully expressed as

$$\mathbf{M}_\ell = \mathbf{n} \cdot \mathbf{M}_O = \mathbf{n} \cdot (\mathbf{r} \times \mathbf{F}) = \begin{vmatrix} n_x & n_y & n_z \\ r_x & r_y & r_z \\ F_x & F_y & F_z \end{vmatrix} \tag{1.2.11}$$

where **n** is a unit vector along the line ℓ, and **r** is a position vector from point *O* on ℓ to a point on the line of action of **F**. Note that M_ℓ is the projection of $\mathbf{M}_O$ on line ℓ.

Special Cases

1. The moment about a line ℓ is zero when the line of action of **F** intersects ℓ (the moment arm is zero).
2. The moment about a line ℓ is zero when the line of action of **F** is parallel to ℓ (the projection of $\mathbf{M}_O$ on ℓ is zero).

Moment of a Couple

A pair of forces equal in magnitude, parallel in lines of action, and opposite in direction is called a *couple*. The magnitude of the moment of a couple is

$$M = Fd$$

where d is the distance between the lines of action of the forces of magnitude F. The moment of a couple is a free vector **M** that can be applied anywhere to a rigid body with the same turning effect, as long as the direction and magnitude of **M** are the same. In other words, a couple vector can be moved to any other location on a given rigid body if it remains parallel to its original position (equivalent couples). Sometimes a curled arrow in the plane of the two forces is used to denote a couple, instead of the couple vector **M**, which is perpendicular to the plane of the two forces.

Force-Couple Transformations

Sometimes it is advantageous to transform a force to a force system acting at another point, or vice versa. The method is illustrated in Figure 1.2.9.

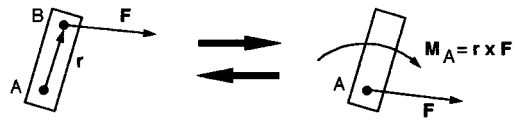


FIGURE 1.2.9 Force-couple transformations.

1. A force **F** acting at B on a rigid body can be replaced by the same force **F** acting at A and a moment $\mathbf{M}_A = \mathbf{r} \times \mathbf{F}$ about A.
2. A force **F** and moment $\mathbf{M}_A$ acting at A can be replaced by a force **F** acting at B for the same total effect on the rigid body.

Simplification of Force Systems

Any force system on a rigid body can be reduced to an equivalent system of a resultant force **R** and a resultant moment $\mathbf{M}_R$. The **equivalent force-couple system** is formally stated as

$$\mathbf{R} = \sum_{i=1}^{n} \mathbf{F}_i \quad \text{and} \quad \mathbf{M}_R = \sum_{i=1}^{n} \mathbf{M}_i = \sum_{i=1}^{n} \left(\mathbf{r}_i \times \mathbf{F}_i \right) \tag{1.2.12}$$

where $\mathbf{M}_R$ depends on the chosen reference point.

Common Cases

1. The resultant force is zero, but there is a resultant moment: $\mathbf{R} = 0$, $\mathbf{M}_R \neq 0$.
2. Concurrent forces (all forces act at one point): $\mathbf{R} \neq 0$, $\mathbf{M}_R = 0$.
3. Coplanar forces: $\mathbf{R} \neq 0$, $\mathbf{M}_R \neq 0$. $\mathbf{M}_R$ is perpendicular to the plane of the forces.
4. Parallel forces: $\mathbf{R} \neq 0$, $\mathbf{M}_R \neq 0$. $\mathbf{M}_R$ is perpendicular to **R**.

Example 2

The torque wrench in Figure 1.2.10 has an arm of constant length L but a variable socket length $d = OA$ because of interchangeable tool sizes. Determine how the moment applied at point O depends on the length d for a constant force $\mathbf{F}$ from the hand.

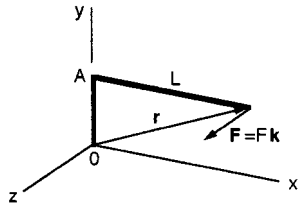


FIGURE 1.2.10 Model of a torque wrench.

Solution. Using $\mathbf{M}_O = \mathbf{r} \times \mathbf{F}$ with $\mathbf{r} = L\mathbf{i} + d\mathbf{j}$ and $\mathbf{F} = F\mathbf{k}$ in Figure 1.2.10,

$$\mathbf{M}_O = (L\mathbf{i} + d\mathbf{j}) \times F\mathbf{k} = Fd\mathbf{i} - FL\mathbf{j}$$

Judgment of the Result

According to a visual analysis the wrench should turn clockwise, so the –**j** component of the moment is justified. Looking at the wrench from the positive x direction, point A has a tendency to rotate counterclockwise. Thus, the **i** component is correct using the right-hand rule.

Equilibrium of Rigid Bodies

The concept of equilibrium is used for determining unknown forces and moments of forces that act on or within a rigid body or system of rigid bodies. The equations of equilibrium are the most useful equations in the area of statics, and they are also important in dynamics and mechanics of materials. The drawing of appropriate free-body diagrams is essential for the application of these equations.

Conditions of Equilibrium

A rigid body is in static equilibrium when the equivalent force-couple system of the external forces acting on it is zero. In vector notation, this condition is expressed as

$$\sum \mathbf{F} = 0$$
$$\sum \mathbf{M}_O = \sum (\mathbf{r} \times \mathbf{F}) = 0 \tag{1.2.13}$$

where O is an arbitrary point of reference.

In practice it is often most convenient to write Equation 1.2.13 in terms of rectangular scalar components,

$$\sum F_x = 0 \qquad \sum M_x = 0$$
$$\sum F_y = 0 \qquad \sum M_y = 0$$
$$\sum F_z = 0 \qquad \sum M_z = 0$$

Maximum Number of Independent Equations for One Body

1. One-dimensional problem: $\sum F = 0$
2. Two-dimensional problem:

$$\sum F_x = 0 \quad \sum F_y = 0 \quad \sum M_A = 0$$

$$\text{or} \quad \sum F_x = 0 \quad \sum M_A = 0 \quad \sum M_B = 0 \quad (x \text{ axis not} \perp \text{AB})$$

$$\text{or} \quad \sum M_A = 0 \quad \sum M_B = 0 \quad \sum M_C = 0 \quad (AB \text{ not} \parallel \text{BC})$$

3. Three-dimensional problem:

$$\sum F_x = 0 \quad \sum F_y = 0 \quad \sum F_z = 0$$

$$\sum M_x = 0 \quad \sum M_y = 0 \quad \sum M_z = 0$$

where *xyz* are orthogonal coordinate axes, and *A*, *B*, *C* are particular points of reference.

Calculation of Unknown Forces and Moments

In solving for unknown forces and moments, always draw the free-body diagram first. Unknown external forces and moments must be shown at the appropriate places of action on the diagram. The directions of unknowns may be assumed arbitrarily, but should be done consistently for systems of rigid bodies. A negative answer indicates that the initial assumption of the direction was opposite to the actual direction. Modeling for problem solving is illustrated in Figures 1.2.11 and 1.2.12.

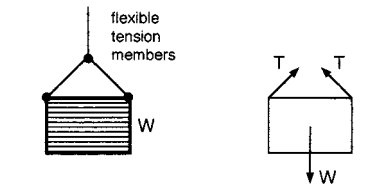


FIGURE 1.2.11 Example of two-dimensional modeling.

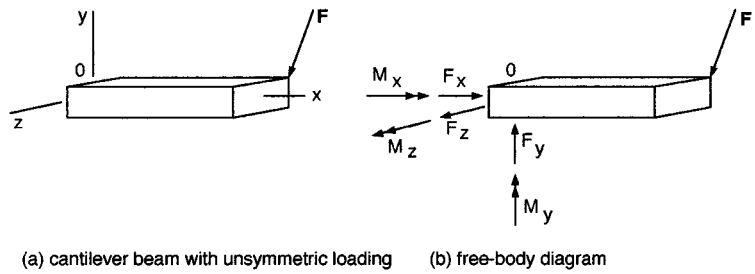


FIGURE 1.2.12 Example of three-dimensional modeling.

Notes on Three-Dimensional Forces and Supports

Each case should be analyzed carefully. Sometimes a particular force or moment is possible in a device, but it must be neglected for most practical purposes. For example, a very short sleeve bearing cannot

support significant moments. A roller bearing may be designed to carry much larger loads perpendicular to the shaft than along the shaft.

Related Free-Body Diagrams

When two or more bodies are in contact, separate free-body diagrams may be drawn for each body. The mutual forces and moments between the bodies are related according to Newton's third law (action and reaction). The directions of unknown forces and moments may be arbitrarily assumed in one diagram, but these initial choices affect the directions of unknowns in all other related diagrams. The number of unknowns and of usable equilibrium equations both increase with the number of related free-body diagrams.

Schematic Example in Two Dimensions (Figure 1.2.13)

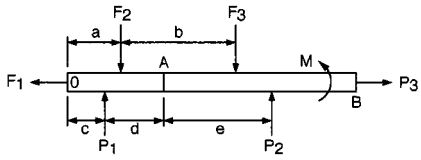


FIGURE 1.2.13 Free-body diagram.

Given: F_1, F_2, F_3, M
Unknowns: P_1, P_2, P_3, and forces and moments at joint A (rigid connection)

Equilibrium Equations

$$\sum F_x = -F_1 + P_3 = 0$$

$$\sum F_y = P_1 + P_2 - F_2 - F_3 = 0$$

$$\sum M_O = P_1 c + P_2 (c + d + e) + M - F_2 a - F_3 (a + b) = 0$$

Three unknowns (P_1, P_2, P_3) are in three equations.

Related Free-Body Diagrams (Figure 1.2.14)

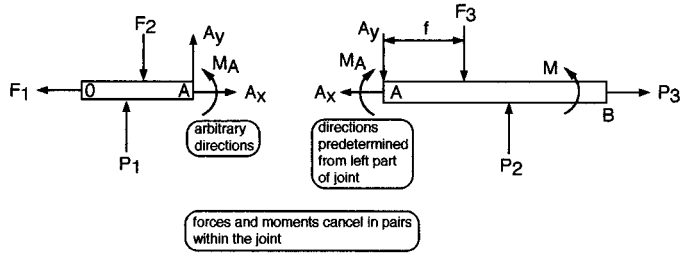


FIGURE 1.2.14 Related free-body diagrams.

Dimensions *a*, *b*, *c*, *d*, and *e* of Figure 1.2.13 are also valid here.

New Set of Equilibrium Equations

Left part: (OA)

$$\sum F_x = -F_1 + A_x = 0$$

$$\sum F_y = P_1 + A_y - F_2 = 0$$

$$\sum M_O = P_1 c + A_y (c + d) + M_A - F_2 a = 0$$

Right side: (AB)

$$\sum F_x = -A_x + P_3 = 0$$

$$\sum F_y = P_2 - A_y - F_3 = 0$$

$$\sum M_A = -M_A + P_2 e + M - F_3 f = 0$$

Six unknowns (P_1, P_2, P_3, A_x, A_y, M_A) are in six equations.

Note: In the first diagram (Figure 1.2.13) the couple M may be moved anywhere from O to B. M is not shown in the second diagram (O to A) because it is shown in the third diagram (in which it may be moved anywhere from A to B).

Example 3

The arm of a factory robot is modeled as three bars (Figure 1.2.15) with coordinates A: (0.6, –0.3, 0.4) m; B: (1, –0.2, 0) m; and C: (0.9, 0.1, –0.25) m. The weight of the arm is represented by $\mathbf{W}_A$ = –60 N**j** at A, and $\mathbf{W}_B$ = –40 N**j** at B. A moment $\mathbf{M}_C$ = (100**i** – 20**j** + 50**k**) N · m is applied to the arm at C. Determine the force and moment reactions at O, assuming that all joints are temporarily fixed.

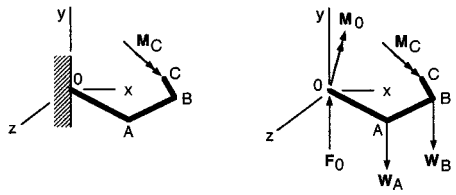


FIGURE 1.2.15 Model of a factory robot.

Solution. The free-body diagram is drawn in Figure 1.2.15b, showing the unknown force and moment reactions at O. From Equation 1.2.13,

$$\sum \mathbf{F} = 0$$

$$\mathbf{F}_O + \mathbf{W}_A + \mathbf{W}_B = 0$$

$$\mathbf{F}_O - 60 \text{ N } \mathbf{j} - 40 \text{ N } \mathbf{j} = 0$$

$$\mathbf{F}_O = 100\ \text{N}\,\mathbf{j}$$

$$\sum M_O = 0$$

$$\mathbf{M}_O + \mathbf{M}_C + (\mathbf{r}_{OA} \times \mathbf{W}_A) + (\mathbf{r}_{OB} \times \mathbf{W}_B) = 0$$

$$\mathbf{M}_O + (100\mathbf{i} - 20\,\mathbf{j} + 50\mathbf{k})\ \text{N}\cdot\text{m} + (0.6\mathbf{i} - 0.3\mathbf{j} + 0.4\mathbf{k})\ \text{m} \times (-60\ \text{N}\,\mathbf{j}) + (\mathbf{i} - 0.2\,\mathbf{j})\ \text{m} \times (-40\ \text{N}\,\mathbf{j}) = 0$$

$$\mathbf{M}_O + 100\ \text{N}\cdot\text{m}\,\mathbf{i} - 20\ \text{N}\cdot\text{m}\,\mathbf{j} + 50\ \text{N}\cdot\text{m}\,\mathbf{k} - 36\ \text{N}\cdot\text{m}\,\mathbf{k} + 24\ \text{N}\cdot\text{m}\,\mathbf{i} - 40\ \text{N}\cdot\text{m}\,\mathbf{k} = 0$$

$$\mathbf{M}_O = (-124\mathbf{i} + 20\,\mathbf{j} + 26\mathbf{k})\ \text{N}\cdot\text{m}$$

Example 4

A load of 7 kN may be placed anywhere within *A* and *B* in the trailer of negligible weight. Determine the reactions at the wheels at *D*, *E*, and *F*, and the force on the hitch *H* that is mounted on the car, for the extreme positions *A* and *B* of the load. The mass of the car is 1500 kg, and its weight is acting at *C* (see Figure 1.2.16).

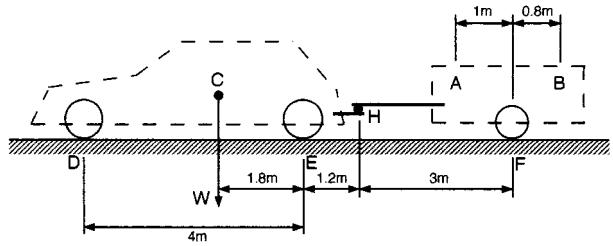


FIGURE 1.2.16 Analysis of a car with trailer.

Solution. The scalar method is best here.

Put the load at position *A* first
For the trailer alone, with *y* as the vertical axis
$\sum M_F = 7(1) - H_y(3) = 0$, $H_y = 2.33$ kN
On the car
$H_y = 2.33$ kN ↓Ans.
$\sum F_y = 2.33 - 7 + F_y = 0$, $F_y = 4.67$ kN ↑Ans.
For the car alone
$\sum M_E = -2.33(1.2) - D_y(4) + 14.72(1.8) = 0$
$D_y = 5.93$ kN ↑Ans.
$\sum F_y = 5.93 + E_y - 14.72 - 2.33 = 0$
$E_y = 11.12$ kN ↑Ans.

Put the load at position *B* next
For the trailer alone
$\sum M_F = 0.8(7) - H_y(3) = 0$, $H_y = -1.87$ kN
On the car
$H_y = 1.87$ kN ↓Ans.
$\sum F_y = -1.87 - 7 + E_y = 0$
$E_y = 8.87$ kN ↑Ans.
For the car alone
$\sum M_E = -(1.87)(1.2) - D_y(4) + 14.72(1.8) = 0$
$D_y = 7.19$ kN ↑Ans.
$\sum F_y = 7.19 + E_y - 14.72 - (-1.87) = 0$
$E_y = 5.66$ kN ↑Ans.

Forces and Moments in Beams

Beams are common structural members whose main function is to resist bending. The geometric changes and safety aspects of beams are analyzed by first assuming that they are rigid. The preceding sections enable one to determine (1) the external (supporting) reactions acting on a statically determinate beam, and (2) the internal forces and moments at any cross section in a beam.

Classification of Supports

Common supports and external reactions for two-dimensional loading of beams are shown in Figure 1.2.17.

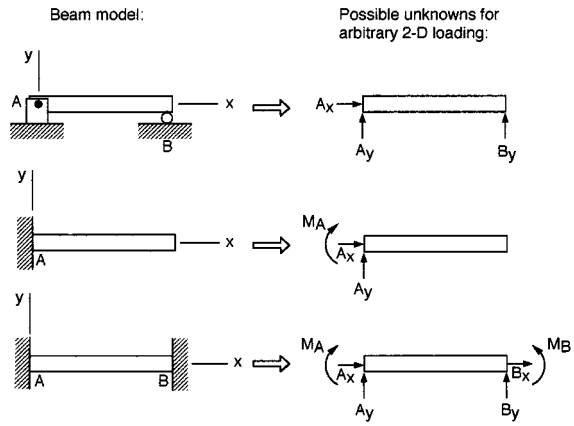


FIGURE 1.2.17 Common beam supports.

Internal Forces and Moments

The internal force and moment reactions in a beam caused by external loading must be determined for evaluating the strength of the beam. If there is no torsion of the beam, three kinds of internal reactions are possible: a horizontal normal force H on a cross section, vertical (transverse) shear force V, and bending moment M. These reactions are calculated from the equilibrium equations applied to the left or right part of the beam from the cross section considered. The process involves free-body diagrams of the beam and a consistently applied system of signs. The modeling is illustrated for a cantilever beam in Figure 1.2.18.

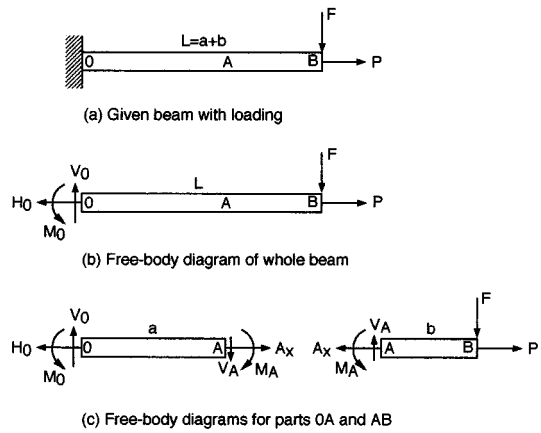


FIGURE 1.2.18 Internal forces and moments in a cantilever beam.

Sign Conventions. Consistent sign conventions should be used in any given problem. These could be arbitrarily set up, but the following is slightly advantageous. It makes the signs of the answers to the equilibrium equations correct for the directions of the shear force and bending moment.

A moment that makes a beam concave upward is taken as positive. Thus, a clockwise moment is positive on the left side of a section, and a counterclockwise moment is positive on the right side. A

shear force that acts upward on the left side of a section, or downward on the right side, is positive (Figure 1.2.19).

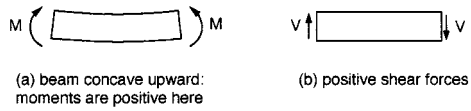


FIGURE 1.2.19 Preferred sign conventions.

Shear Force and Bending Moment Diagrams

The critical locations in a beam are determined from shear force and bending moment diagrams for the whole length of the beam. The construction of these diagrams is facilitated by following the steps illustrated for a cantilever beam in Figure 1.2.20.

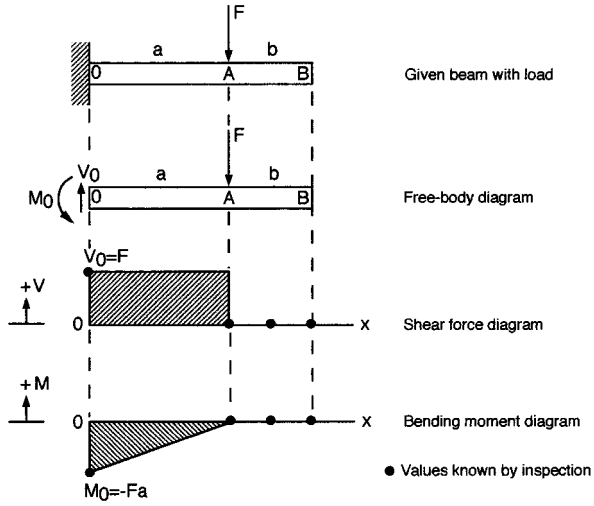


FIGURE 1.2.20 Construction of shear force and bending moment diagrams.

1. Draw the free-body diagram of the whole beam and determine all reactions at the supports.
2. Draw the coordinate axes for the shear force (V) and bending moment (M) diagrams directly below the free-body diagram.
3. Immediately plot those values of V and M that can be determined by inspection (especially where they are zero), observing the sign conventions.
4. Calculate and plot as many additional values of V and M as are necessary for drawing reasonably accurate curves through the plotted points, or do it all by computer.

Example 5

A construction crane is modeled as a rigid bar AC which supports the boom by a pin at B and wire CD. The dimensions are AB = 10ℓ, BC = 2ℓ, BD = DE = 4ℓ. Draw the shear force and bending moment diagrams for bar AC (Figure 1.2.21).

Solution. From the free-body diagram of the entire crane,

$$\sum F_x = 0 \qquad \sum F_y = 0 \qquad \sum M_A = 0$$

$$A_x = 0 \qquad -P + A_y = 0 \qquad -P(8\ell) + M_A = 0$$

$$A_y = P \qquad M_A = 8P\ell$$

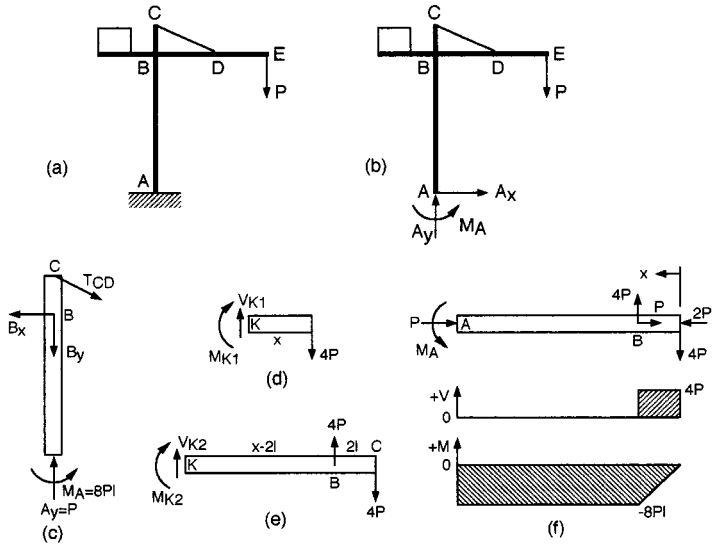


FIGURE 1.2.21 Shear force and bending moment diagrams of a component in a structure.

Now separate bar AC and determine the forces at B and C.

$$\sum F_x = 0 \qquad \sum F_y = 0 \qquad \sum M_A = 0$$

$$-B_x + T_{CD_x} = 0 \qquad P - B_y - T_{CD_y} = 0 \qquad -\frac{2}{\sqrt{5}} T_{CD}(12\ell) + B_x(10\ell) + M_A = 0$$

$$\text{(a)}\ B_x = \frac{2}{\sqrt{5}} T_{CD} \qquad \text{(b)}\ B_y = P - \frac{1}{\sqrt{5}} T_{CD} \qquad -\frac{24\ell}{\sqrt{5}} T_{CD} + \frac{20\ell}{\sqrt{5}} T_{CD} = -8P\ell$$

$$\text{(c)}\ T_{CD} = \frac{8\sqrt{5}}{4} P = 2\sqrt{5}P$$

From (a) and (c), $B_x = 4P$ and $T_{CD_x} = 4P$. From (b) and (c), $B_y = P - 2P = -P$ and $T_{CD_y} = 2P$.

Draw the free-body diagram of bar AC horizontally, with the shear force and bending moment diagram axes below it. Measure x from end C for convenience and analyze sections $0 \le x \le 2\ell$ and $2\ell \le x \le 12\ell$ (Figures 1.2.21b to 1.2.21f).

1. $0 \le x \le 2\ell$

$$\sum F_y = 0 \qquad \sum M_K = 0$$

$$-4P + V_{K_1} = 0 \qquad M_{K_1} + 4P(x) = 0$$

$$V_{K_1} = 4P \qquad M_{K_1} = -4Px$$

2. $2\ell \le x \le 12\ell$

$$\sum F_y = 0 \qquad \sum M_K = 0$$

$$4P - 4P + V_{K_2} = 0 \qquad M_{K_2} - 4P(x - 2\ell) + 4P(x) = 0$$

$$V_{K_2} = 0 \qquad M_{K_2} = -8P\ell$$

At point *B*, $x = 2\ell$, $M_{K_1} = -4P(2\ell) = -8P\ell = M_{K_2} = M_A$. The results for section *AB*, $2\ell \le x \le 12\ell$, show that the combined effect of the forces at *B* and *C* is to produce a couple of magnitude $8P\ell$ on the beam. Hence, the shear force is zero and the moment is constant in this section. These results are plotted on the axes below the free-body diagram of bar *A*-*B*-*C*.

Simple Structures and Machines

Ryan Roloff and Bela I. Sandor

Equilibrium equations are used to determine forces and moments acting on statically determinate simple structures and machines. A simple structure is composed solely of two-force members. A machine is composed of multiforce members. The method of joints and the method of sections are commonly used in such analysis.

Trusses

Trusses consist of straight, slender members whose ends are connected at joints. Two-dimensional *plane trusses* carry loads acting in their planes and are often connected to form three-dimensional *space trusses*. Two typical trusses are shown in Figure 1.2.22.

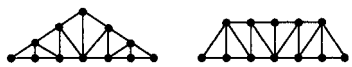


FIGURE 1.2.22 Schematic examples of trusses.

To simplify the analysis of trusses, assume frictionless pin connections at the joints. Thus, all members are two-force members with forces (and no moments) acting at the joints. Members may be assumed weightless or may have their weights evenly divided to the joints.

Method of Joints

Equilibrium equations based on the entire truss and its joints allow for determination of all internal forces and external reactions at the joints using the following procedure.

1. Determine the support reactions of the truss. This is done using force and moment equilibrium equations and a free-body diagram of the entire truss.
2. Select any arbitrary joint where only one or two unknown forces act. Draw the free-body diagram of the joint assuming unknown forces are tensions (arrows directed away from the joint).
3. Draw free-body diagrams for the other joints to be analyzed, using Newton's third law consistently with respect to the first diagram.
4. Write the equations of equilibrium, $\sum F_x = 0$ and $\sum F_y = 0$, for the forces acting at the joints and solve them. To simplify calculations, attempt to progress from joint to joint in such a way that each equation contains only one unknown. Positive answers indicate that the assumed directions of unknown forces were correct, and vice versa.

Example 6

Use the method of joints to determine the forces acting at A, B, C, H, and I of the truss in Figure 1.2.23a. The angles are $\alpha = 56.3°$, $\beta = 38.7°$, $\phi = 39.8°$, and $\theta = 36.9°$.

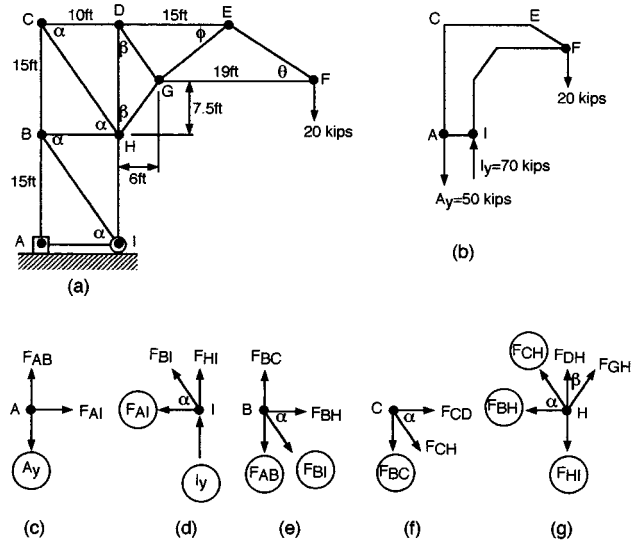


FIGURE 1.2.23 Method of joints in analyzing a truss.

Solution. First the reactions at the supports are determined and are shown in Figure 1.2.23b. A joint at which only two unknown forces act is the best starting point for the solution. Choosing joint A, the solution is progressively developed, always seeking the next joint with only two unknowns. In each diagram circles indicate the quantities that are known from the preceding analysis. Sample calculations show the approach and some of the results.

Joint A:

$$\sum F_x = 0 \qquad \sum F_y = 0$$

$$F_{AI} = 0 \qquad F_{AB} - A_y = 0$$

$$F_{AB} - 50 \text{ kips} = 0$$

$$F_{AB} = 50 \text{ kips (tension)}$$

Joint H:

$$\sum F_x = 0 \qquad \sum F_y = 0$$

$$F_{GH}\sin\beta - F_{CH}\cos\alpha - F_{BH} = 0 \qquad F_{CH}\sin\alpha + F_{DH} + F_{GH}\cos\beta - F_{HI} = 0$$

$$F_{GH}(0.625) + (60.1 \text{ kips})(0.555) - 0 = 0 \qquad -(60.1 \text{ kips})(0.832) + F_{DH} - (53.4 \text{ kips})(0.780) + 70 \text{ kips} = 0$$

$$F_{GH} = -53.4 \text{ kips (compression)} \qquad F_{DH} = 21.7 \text{ kips (tension)}$$

Method of Sections

The method of sections is useful when only a few forces in truss members need to be determined regardless of the size and complexity of the entire truss structure. This method employs any section of the truss as a free body in equilibrium. The chosen section may have any number of joints and members in it, but the number of unknown forces should not exceed three in most cases. Only three equations of equilibrium can be written for each section of a plane truss. The following procedure is recommended.

1. Determine the support reactions if the section used in the analysis includes the joints supported.

2. Section the truss by making an imaginary cut through the members of interest, preferably through only three members in which the forces are unknowns (assume tensions). The cut need not be a straight line. The sectioning is illustrated by lines *l-l*, *m-m*, and *n-n* in Figure 1.2.24.
3. Write equations of equilibrium. Choose a convenient point of reference for moments to simplify calculations such as the point of intersection of the lines of action for two or more of the unknown forces. If two unknown forces are parallel, sum the forces perpendicular to their lines of action.
4. Solve the equations. If necessary, use more than one cut in the vicinity of interest to allow writing more equilibrium equations. Positive answers indicate assumed directions of unknown forces were correct, and vice versa.

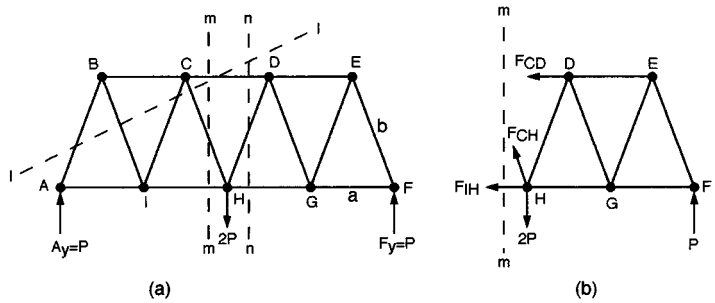


FIGURE 1.2.24 Method of sections in analyzing a truss.

Space Trusses

A space truss can be analyzed with the method of joints or with the method of sections. For each joint, there are three scalar equilibrium equations, $\sum F_x = 0$, $\sum F_y = 0$, and $\sum F_z = 0$. The analysis must begin at a joint where there are at least one known force and no more than three unknown forces. The solution must progress to other joints in a similar fashion.

There are six scalar equilibrium equations available when the method of sections is used: $\sum F_x = 0$, $\sum F_y = 0$, $\sum F_z = 0$, $\sum M_x = 0$, $\sum M_y = 0$, and $\sum M_z = 0$.

Frames and Machines

Multiforce members (with three or more forces acting on each member) are common in structures. In these cases the forces are not directed along the members, so they are a little more complex to analyze than the two-force members in simple trusses. Multiforce members are used in two kinds of structure. *Frames* are usually stationary and fully constrained. *Machines* have moving parts, so the forces acting on a member depend on the location and orientation of the member.

The analysis of multiforce members is based on the consistent use of related free-body diagrams. The solution is often facilitated by representing forces by their rectangular components. Scalar equilibrium equations are the most convenient for two-dimensional problems, and vector notation is advantageous in three-dimensional situations.

Often, an applied force acts at a pin joining two or more members, or a support or connection may exist at a joint between two or more members. In these cases, a choice should be made of a single member at the joint on which to assume the external force to be acting. This decision should be stated in the analysis. The following comprehensive procedure is recommended.

Three independent equations of equilibrium are available for each member or combination of members in two-dimensional loading; for example, $\sum F_x = 0$, $\sum F_y = 0$, $\sum M_A = 0$, where A is an arbitrary point of reference.

1. Determine the support reactions if necessary.
2. Determine all two-force members.

3. Draw the free-body diagram of the first member on which the unknown forces act assuming that the unknown forces are tensions.
4. Draw the free-body diagrams of the other members or groups of members using Newton's third law (action and reaction) consistently with respect to the first diagram. Proceed until the number of equilibrium equations available is no longer exceeded by the total number of unknowns.
5. Write the equilibrium equations for the members or combinations of members and solve them. Positive answers indicate that the assumed directions for unknown forces were correct, and vice versa.

Distributed Forces

The most common distributed forces acting on a body are parallel force systems, such as the force of gravity. These can be represented by one or more concentrated forces to facilitate the required analysis. Several basic cases of distributed forces are presented here. The important topic of stress analysis is covered in mechanics of materials.

Center of Gravity

The center of gravity of a body is the point where the equivalent resultant force caused by gravity is acting. Its coordinates are defined for an arbitrary set of axes as

$$\bar{x} = \frac{\int x\, dW}{W} \qquad \bar{y} = \frac{\int y\, dW}{W} \qquad \bar{z} = \frac{\int z\, dW}{W} \tag{1.2.14}$$

where x, y, z are the coordinates of an element of weight dW, and W is the total weight of the body. In the general case $dW = \gamma\, dV$, and $W = \int \gamma\, dV$, where γ = specific weight of the material and dV = elemental volume.

Centroids

If γ is a constant, the center of gravity coincides with the centroid, which is a geometrical property of a body. Centroids of lines L, areas A, and volumes V are defined analogously to the coordinates of the center of gravity,

$$\text{Lines:} \quad \bar{x} = \frac{\int x\, dL}{W} \qquad \bar{y} = \frac{\int y\, dL}{L} \qquad \bar{z} = \frac{\int z\, dL}{L} \tag{1.2.15}$$

$$\text{Areas:} \quad \bar{x} = \frac{\int x\, dA}{A} \qquad \bar{y} = \frac{\int y\, dA}{A} \qquad \bar{z} = \frac{\int z\, dA}{A} \tag{1.2.16}$$

$$\text{Volumes:} \quad \bar{x} = \frac{\int x\, dV}{V} \qquad \bar{y} = \frac{\int y\, dV}{V} \qquad \bar{z} = \frac{\int z\, dV}{V} \tag{1.2.17}$$

For example, an area A consists of discrete parts A_1, A_2, A_3, where the centroids x_1, x_2, x_3 of the three parts are located by inspection. The x coordinate of the centroid of the whole area A is $\bar{x}$ obtained from $A\bar{x} = A_1x_1 + A_2x_2 + A_3x_3$.

Surfaces of Revolution. The surface areas and volumes of bodies of revolution can be calculated using the concepts of centroids by the theorems of Pappus (see texts on Statics).

Distributed Loads on Beams

The distributed load on a member may be its own weight and/or some other loading such as from ice or wind. The external and internal reactions to the loading may be determined using the condition of equilibrium.

External Reactions. Replace the whole distributed load with a concentrated force equal in magnitude to the area under the load distribution curve and applied at the centroid of that area parallel to the original force system.

Internal Reactions. For a beam under a distributed load $w(x)$, where x is distance along the beam, the shear force V and bending moment M are related according to Figure 1.2.25 as

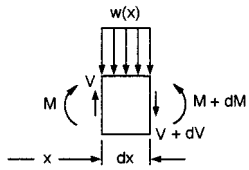


FIGURE 1.2.25 Internal reactions in a beam under distributed loading.

$$w(x) = -\frac{dV}{dx} \qquad V = \frac{dM}{dx} \tag{1.2.18}$$

Other useful expressions for any two cross sections A and B of a beam are

$$V_A - V_B = \int_{x_A}^{x_B} w(x)\,dx = \text{area under } w(x)$$
$$M_B - M_A = \int_{x_A}^{x_B} V\,dx = \text{area under shear force diagram} \tag{1.2.19}$$

Example 7 (Figure 1.2.26)

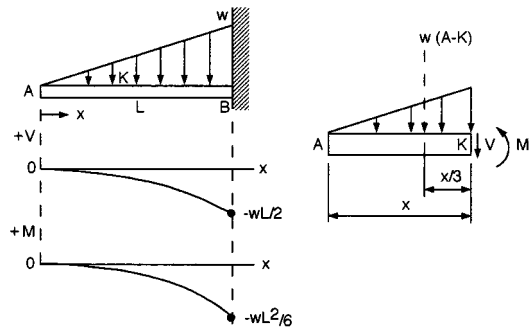


FIGURE 1.2.26 Shear force and bending moment diagrams for a cantilever beam.

Distributed Loads on Flexible Cables

The basic assumptions of simple analyses of cables are that there is no resistance to bending and that the internal force at any point is tangent to the cable at that point. The loading is denoted by $w(x)$, a

continuous but possibly variable load, in terms of force per unit length. The differential equation of a cable is

$$\frac{d^2y}{dx^2} = \frac{w(x)}{T_o} \tag{1.2.20}$$

where T_o = constant = horizontal component of the tension T in the cable.

Two special cases are common.

Parabolic Cables. The cable supports a load w which is uniformly distributed horizontally. The shape of the cable is a parabola given by

$$y = \frac{wx^2}{2T_o} \quad (x = 0 \text{ at lowest point}) \tag{1.2.21}$$

In a symmetric cable the tension is $T = \sqrt{T_o^2 + w^2x^2}$.

Catenary Cables. When the load w is uniformly distributed along the cable, the cable's shape is given by

$$y = \frac{T_o}{w}\left(\cosh\frac{wx}{T_o} - 1\right) \tag{1.2.22}$$

The tension in the cable is $T = T_o + wy$.

Friction

A friction force F (or $\mathcal{F}$, in typical other notation) acts between contacting bodies when they slide relative to one another, or when sliding tends to occur. This force is tangential to each body at the point of contact, and its magnitude depends on the normal force N pressing the bodies together and on the material and condition of the contacting surfaces. The material and surface properties are lumped together and represented by the coefficient of friction μ. The friction force opposes the force that tends to cause motion, as illustrated for two simple cases in Figure 1.2.27.

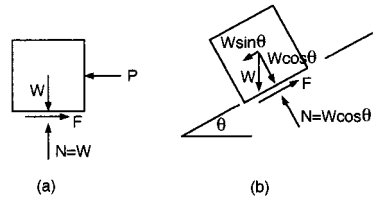


FIGURE 1.2.27 Models showing friction forces.

The friction forces F may vary from zero to a maximum value,

$$F_{\max} = \mu N \quad (0 \le F \le F_{\max}) \tag{1.2.23}$$

depending on the applied force that tends to cause relative motion of the bodies. The coefficient of kinetic friction μ_k (during sliding) is lower than the coefficient of static friction μ or μ_s; μ_k depends on the speed of sliding and is not easily quantified.

Angle of Repose

The critical angle θ_c at which motion is impending is the angle of repose, where the friction force is at its maximum for a given block on an incline.

$$\tan\theta_c = \frac{F}{N} = \mu_s \tag{1.2.24}$$

So θ_c is measured to obtain μ_s. Note that, even in the case of static, dry friction, μ_s depends on temperature, humidity, dust and other contaminants, oxide films, surface finish, and chemical reactions. The contact area and the normal force affect μ_s only when significant deformations of one or both bodies occur.

Classifications and Procedures for Solving Friction Problems

The directions of unknown friction forces are often, but not always, determined by inspection. The magnitude of the friction force is obtained from $F_{max} = \mu_s N$ when it is known that motion is impending. Note that F may be less than F_{max}. The major steps in solving problems of dry friction are organized in three categories as follows.

A. Given: Bodies, forces, or coefficients of friction are known. Impending motion is not assured: $F \neq \mu_s N$.

Procedure: To determine if equilibrium is possible:

1. Construct the free-body diagram.
2. Assume that the system is in equilibrium.
3. Determine the friction and normal forces necessary for equilibrium.
4. Results: (a) $F < \mu_s N$, the body is at rest.
   (b) $F > \mu_s N$, motion is occurring, static equilibrium is not possible. Since there is motion, $F = \mu_k N$. Complete solution requires principles of dynamics.

B. Given: Bodies, forces, or coefficients of friction are given. Impending motion is specified. $F = \mu_s N$ is valid.

Procedure: To determine the unknowns:

1. Construct the free-body diagram.
2. Write $F = \mu_s N$ for all surfaces where motion is impending.
3. Determine μ_s or the required forces from the equation of equilibrium.

C. Given: Bodies, forces, coefficients of friction are known. Impending motion is specified, but the exact motion is not given. The possible motions may be sliding, tipping or rolling, or relative motion if two or more bodies are involved. Alternatively, the forces or coefficients of friction may have to be determined to produce a particular motion from several possible motions.

Procedure: To determine the exact motion that may occur, or unknown quantities required:

1. Construct the free-body diagram.
2. Assume that motion is impending in one of the two or more possible ways. Repeat this for each possible motion and write the equation of equilibrium.
3. Compare the results for the possible motions and select the likely event. Determine the required unknowns for any preferred motion.

Wedges and Screws

A wedge may be used to raise or lower a body. Thus, two directions of motion must be considered in each situation, with the friction forces always opposing the impending or actual motion. The self-locking

aspect of a wedge may be of interest. The analysis is straightforward using interrelated free-body diagrams and equilibrium equations.

Screw threads are special applications of the concept of wedges. Square threads are the easiest to model and analyze. The magnitude M of the moment of a couple required to move a square-threaded screw against an axial load P is

$$M = Pr\tan(\alpha + \phi) \tag{1.2.25}$$

where r = radius of the screw
α = $\tan^{-1}(L/2\pi r) = \tan^{-1}(np/2\pi r)$
L = lead = advancement per revolution
n = multiplicity of threads
p = pitch = distance between similar points on adjacent threads
ϕ = $\tan^{-1}\mu$

The relative values of α and ϕ control whether a screw is self-locking; $\phi > \alpha$ is required for a screw to support an axial load without unwinding.

Disk Friction

Flat surfaces in relative rotary motion generate a friction moment M opposing the motion. For a hollow member with radii R_o and R_i, under an axial force P,

$$M = \frac{2}{3}\mu P\frac{R_o^3 - R_i^3}{R_o^2 - R_i^2} \tag{1.2.26}$$

The friction moment tends to decrease (down to about 75% of its original value) as the surfaces wear. Use the appropriate μ_s or μ_k value.

Axle Friction

The friction moment M of a rotating axle in a journal bearing (sliding bearing) is approximated (if μ is low) as

$$M = Pr\mu \tag{1.2.27}$$

where P = transverse load on the axle
r = radius of the axle

Use the appropriate μ_s or μ_k value.

Rolling Resistance

Rolling wheels and balls have relatively low resistance to motion compared to sliding. This resistance is caused by internal friction of the materials in contact, and it may be difficult to predict or measure.

A coefficient of rolling resistance a is defined with units of length,

$$a \cong \frac{Fr}{P} \tag{1.2.28}$$

where r = radius of a wheel rolling on a flat surface
F = minimum horizontal force to maintain constant speed of rolling
P = load on wheel

Values of a range upward from a low of about 0.005 mm for hardened steel elements.

Belt Friction

The tensions T_1 and T_2 of a belt, rope, or wire on a pulley or drum are related as

$$T_2 = T_1 e^{\mu\beta} \qquad (T_2 > T_1) \tag{1.2.29}$$

where β = total angle of belt contact, radians ($\beta = 2\pi n$ for a member wrapped around a drum n times). Use μ_s for impending slipping and μ_k for slipping.

For a V belt of belt angle 2ϕ,

$$T_2 = T_1 e^{\mu\beta/\sin\phi}$$

Work and Potential Energy

Work is a scalar quantity. It is the product of a force and the corresponding displacement. Potential energy is the capacity of a system to do work on another system. These concepts are advantageous in the analysis of equilibrium of complex systems, in dynamics, and in mechanics of materials.

Work of a Force

The work U of a constant force **F** is

$$U = Fs \tag{1.2.30}$$

where s = displacement of a body in the direction of the vector **F**.

For a displacement along an arbitrary path from point 1 to 2, with $d\mathbf{r}$ tangent to the path,

$$U = \int_1^2 \mathbf{F} \cdot d\mathbf{r} = \int_1^2 \left(F_x\, dx + F_y\, dy + F_z\, dz\right)$$

In theory, there is no work when:

1. A force is acting on a fixed, rigid body ($dr = 0$, $dU = 0$).
2. A force acts perpendicular to the displacement ($\mathbf{F} \cdot d\mathbf{r} = 0$).

Work of a Couple

A couple of magnitude M does work

$$U = M\theta \tag{1.2.31}$$

where θ = angular displacement (radians) in the same plane in which the couple is acting.

In a rotation from angular position α to β,

$$U = \int_\alpha^\beta \mathbf{M} \cdot d\theta = \int_\alpha^\beta \left(M_x\, d\theta_x + M_y\, d\theta_y + M_z\, d\theta_z\right)$$

Virtual Work

The concept of virtual work (through imaginary, infinitesimal displacements within the constraints of a system) is useful to analyze the equilibrium of complex systems. The virtual work of a force **F** or moment **M** is expressed as

$$\delta U = \mathbf{F} \cdot \delta \mathbf{r}$$

$$\delta U = \mathbf{M} \cdot \delta \theta$$

There is equilibrium if

$$\delta U = \sum_{i=1}^{m} \mathbf{F}_i \cdot \delta \mathbf{r}_i + \sum_{j=1}^{n} \mathbf{M}_i \cdot \delta \theta_j = 0 \tag{1.2.32}$$

where the subscripts refer to individual forces or couples and the corresponding displacements, ignoring frictional effects.

Mechanical Efficiency of Real Systems

Real mechanical systems operate with frictional losses, so

$$\underset{(\text{output work})}{\text{input work} = \text{useful work}} + \text{work of friction}$$

The mechanical efficiency η of a machine is

$$\eta = \frac{\text{output work}}{\text{input work}} = \frac{\text{useful work}}{\text{total work required}}$$

$$0 < \eta < 1$$

Gravitational Work and Potential Energy

The potential of a body of weight W to do work because of its relative height h with respect to an arbitrary level is defined as its potential energy. If h is the vertical (y) distance between level 1 and a lower level 2, the work of weight W in descending is

$$U_{12} = \int_1^2 W\, dy = Wh = \text{potential energy of the body at level 1 with respect to level 2}$$

The work of weight W in rising from level 2 to level 1 is

$$U_{21} = \int_2^1 -W\, dy = -Wh = \text{potential energy of the body at level 2 with respect to level 1}$$

Elastic Potential Energy

The potential energy of elastic members is another common form of potential energy in engineering mechanics. For a linearly deforming helical spring, the axial force F and displacement x are related by the spring constant k,

$$F = kx \qquad (\text{similarly, } M = k\theta \text{ for a torsion spring})$$

The work U of a force F on an initially undeformed spring is

$$U = \frac{1}{2} kx^2 \tag{1.2.33}$$

In the general case, deforming the spring from position x_1 to x_2,

$$U = \frac{1}{2}k\left(x_2^2 - x_1^2\right)$$

Notation for Potential Energy

The change in the potential energy V of a system is

$$U = -\Delta V$$

Note that negative work is done by a system while its own potential energy is increased by the action of an external force or moment. The external agent does positive work at the same time since it acts in the same direction as the resulting displacement.

Potential Energy at Equilibrium

For equilibrium of a system,

$$\frac{dV}{dq} = 0$$

where q = an independent coordinate along which there is possibility of displacement.

For a system with n degrees of freedom,

$$\frac{\partial V}{\partial q_i} = 0, \qquad i = 1, 2, \ldots, n$$

Equilibrium is stable if $(d^2V/dq^2) > 0$.
Equilibrium is unstable if $(d^2V/dq^2) < 0$.

Equilibrium is neutral only if all derivatives of V are zero. In cases of complex configurations, evaluate derivatives of higher order as well.

Moments of Inertia

The topics of inertia are related to the methods of first moments. They are traditionally presented in statics in preparation for application in dynamics or mechanics of materials.

Moments of Inertia of a Mass

The moment of inertia dI_x of an elemental mass dM about the x axis (Figure 1.2.28) is defined as

$$dI_x = r^2\ dM = \left(y^2 + z^2\right) dM$$

where r is the nearest distance from dM to the x axis. The moments of inertia of a body about the three coordinate axes are

$$I_x = \int r^2\, dM = \int \left(y^2 + z^2\right) dM$$

$$I_y = \int \left(x^2 + z^2\right) dM \qquad (1.2.34)$$

$$I_z = \int \left(x^2 + y^2\right) dM$$

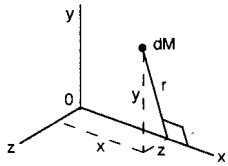


FIGURE 1.2.28 Mass element *dM* in *xyz* coordinates.

Radius of Gyration. The radius of gyration r_g is defined by $r_g = \sqrt{I_x / M}$, and similarly for any other axis. It is based on the concept of the body of mass *M* being replaced by a point mass *M* (same mass) at a distance r_g from a given axis. A thin strip or shell with all mass essentially at a constant distance r_g from the axis of reference is equivalent to a point mass for some analyses.

Moment of Inertia of an Area

The moment of inertia of an elemental area *dA* about the *x* axis (Figure 1.2.29) is defined as

$$dI_x = y^2\, dA$$

where *y* is the nearest distance from *dA* to the *x* axis. The moments of inertia (second moments) of the area *A* about the *x* and *y* axes (because *A* is in the *xy* plane) are

$$I_x = \int y^2\, dA \qquad I_y = \int x^2\, dA \qquad (1.2.35)$$

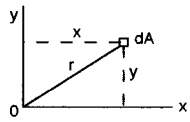


FIGURE 1.2.29 Area *A* in the *xy* plane.

The radius of gyration of an area is defined the same way as it is for a mass: $r_g = \sqrt{I_x / A}$, etc.

Polar Moment of Inertia of an Area

The polar moment of inertia is defined with respect to an axis perpendicular to the area considered. In Figure 1.2.29 this may be the *z* axis. The polar moment of inertia in this case is

$$J_O = \int r^2\, dA = \int \left(x^2 + y^2\right) dA = I_x + I_y \qquad (1.2.36)$$

Parallel-Axis Transformations of Moments of Inertia

It is often convenient to first calculate the moment of inertia about a centroidal axis and then transform this with respect to a parallel axis. The formulas for the transformations are

$$
\begin{aligned}
I &= I_C + Md^2 && \text{for a mass } M \\
I &= I_C + Ad^2 && \text{for an area } A \\
J_O &= J_C + Ad^2 && \text{for an area } A
\end{aligned}
\tag{1.2.37}
$$

where I or J_O = moment of inertia of M or A about any line ℓ
I_C or J_C = moment of inertia of M or A about a line through the mass center or centroid and parallel to ℓ
d = nearest distance between the parallel lines
Note that one of the two axes in each equation must be a centroidal axis.

Products of Inertia

The products of inertia for areas and masses and the corresponding parallel-axis formulas are defined in similar patterns. Using notations in accordance with the preceding formulas, products of inertia are

$$
\begin{aligned}
I_{xy} &= \int xy\, dA \quad \text{for area,} && \text{or} \quad \int xy\, dM \quad \text{for mass} \\
I_{yz} &= \int yz\, dA && \text{or} \quad \int yz\, dM \\
I_{xz} &= \int xz\, dA && \text{or} \quad \int xz\, dM
\end{aligned}
\tag{1.2.38}
$$

Parallel-axis formulas are

$$
\begin{aligned}
I_{xy} &= I_{x'y'} + A\, d_x d_y \quad \text{for area,} && \text{or} \quad I_{x'y'} + M\, d_x d_y \quad \text{for mass} \\
I_{yz} &= I_{y'z'} + A\, d_y d_z && \text{or} \quad I_{y'z'} + M\, d_y d_z \\
I_{xz} &= I_{x'z'} + A\, d_x d_z && \text{or} \quad I_{x'z'} + M\, d_x d_z
\end{aligned}
\tag{1.2.39}
$$

Notes: The moment of inertia is always positive. The product of inertia may be positive, negative, or zero; it is zero if x or y (or both) is an axis of symmetry of the area. Transformations of known moments and product of inertia to axes that are inclined to the original set of axes are possible but not covered here. These transformations are useful for determining the principal (maximum and minimum) moments of inertia and the principal axes when the area or body has no symmetry. The principal moments of inertia for objects of simple shape are available in many texts.

1.3 Dynamics

Stephen M. Birn and Bela I. Sandor

There are two major categories in dynamics, kinematics and kinetics. **Kinematics** involves the time- and geometry-dependent motion of a particle, rigid body, deformable body, or a fluid without considering the forces that cause the motion. It relates position, velocity, acceleration, and time. **Kinetics** combines the concepts of kinematics and the forces that cause the motion.

Kinematics of Particles

Scalar Method

The scalar method of particle kinematics is adequate for one-dimensional analysis. A particle is a body whose dimensions can be neglected (in some analyses, very large bodies are considered particles). The equations described here are easily adapted and applied to two and three dimensions.

Average and Instantaneous Velocity

The average velocity of a particle is the change in distance divided by the change in time. The instantaneous velocity is the particle's velocity at a particular instant.

$$v_{ave} = \frac{\Delta x}{\Delta t} \qquad v_{inst} = \lim_{\Delta t \to 0} \frac{\Delta x}{\Delta t} = \frac{dx}{dt} = \dot{x} \tag{1.3.1}$$

Average and Instantaneous Acceleration

The average acceleration is the change in velocity divided by the change in time. The instantaneous acceleration is the particle's acceleration at a particular instant.

$$a_{ave} = \frac{\Delta v}{\Delta t} \qquad a_{inst} = \lim_{\Delta t \to 0} \frac{\Delta v}{\Delta t} = \frac{dv}{dt} = \dot{v} = \ddot{x} \tag{1.3.2}$$

Displacement, velocity, acceleration, and time are related to one another. For example, if velocity is given as a function of time, the displacement and acceleration can be determined through integration and differentiation, respectively. The following example illustrates this concept.

Example 8

A particle moves with a velocity $v(t) = 3t^2 - 8t$. Determine $x(t)$ and $a(t)$, if $x(0) = 5$.

Solution.

1. Determine $x(t)$ by integration

$$v = \frac{dx}{dt}$$

$$v\,dt = dx$$

$$\int 3t^2 - 8t\,dt = \int dx$$

$$t^3 - 4t^2 + C = x$$

$$\text{from } x(0) = 5 \qquad C = 5$$

$$x(t) = t^3 - 4t^2 + 5$$

2. Determine $a(t)$ by differentiation

$$a = \frac{dv}{dt} = \frac{d}{dt}\left(3t^2 - 8t\right)$$

$$a(t) = 6t - 8$$

There are four key points to be seen from these graphs (Figure 1.3.1).

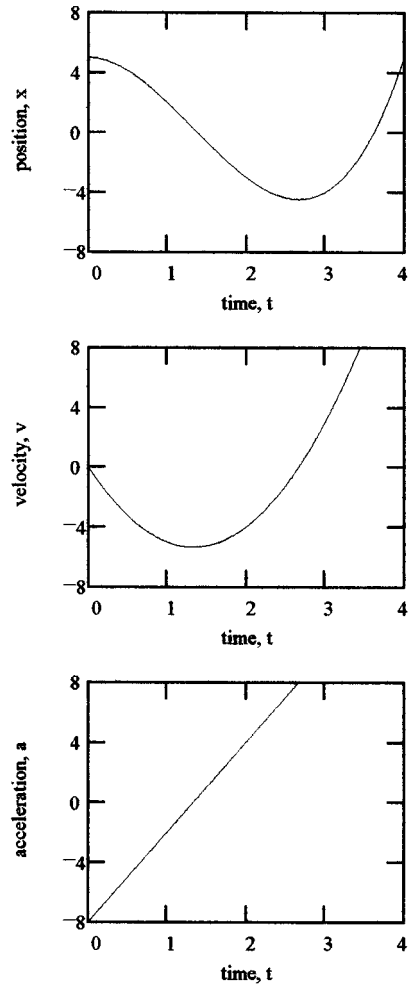


FIGURE 1.3.1 Plots of a particle's kinematics.

1. $v = 0$ at the local maximum or minimum of the x-t curve.
2. $a = 0$ at the local maximum or minimum of the v-t curve.
3. The area under the v-t curve in a specific time interval is equal to the net displacement change in that interval.
4. The area under the a-t curve in a specific time interval is equal to the net velocity change in that interval.

Useful Expressions Based on Acceleration

Equations for nonconstant acceleration:

$$a = \frac{dv}{dt} \Rightarrow \int_{v_0}^{v} dv = \int_{0}^{t} a\, dt \tag{1.3.3}$$

$$v\, dv = a\, dx \Rightarrow \int_{v_0}^{v} v\, dv = \int_{x_0}^{x} a\, dx \tag{1.3.4}$$

Equations for constant acceleration (projectile motion; free fall):

$$\begin{aligned} v &= at + v_0 \\ v^2 &= 2a\left(x - x_0\right) + v_0^2 \\ x &= \frac{1}{2}at^2 + v_0 t + x_0 \end{aligned} \tag{1.3.5}$$

These equations are only to be used when the acceleration is known to be a constant. There are other expressions available depending on how a variable acceleration is given as a function of time, velocity, or displacement.

Scalar Relative Motion Equations

The concept of relative motion can be used to determine the displacement, velocity, and acceleration between two particles that travel along the same line. Equation 1.3.6 provides the mathematical basis for this method. These equations can also be used when analyzing two points on the same body that are not attached rigidly to each other (Figure 1.3.2).

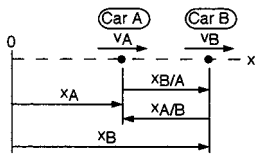


FIGURE 1.3.2 Relative motion of two particles along a straight line.

$$\begin{aligned} x_{B/A} &= x_B - x_A \\ v_{B/A} &= v_B - v_A \\ a_{B/A} &= a_B - a_A \end{aligned} \tag{1.3.6}$$

The notation B/A represents the displacement, velocity, or acceleration of particle B as seen from particle A. Relative motion can be used to analyze many different degrees-of-freedom systems. A degree of freedom of a mechanical system is the number of independent coordinate systems needed to define the position of a particle.

Vector Method

The vector method facilitates the analysis of two- and three-dimensional problems. In general, curvilinear motion occurs and is analyzed using a convenient coordinate system.

Vector Notation in Rectangular (Cartesian) Coordinates

Figure 1.3.3 illustrates the vector method.

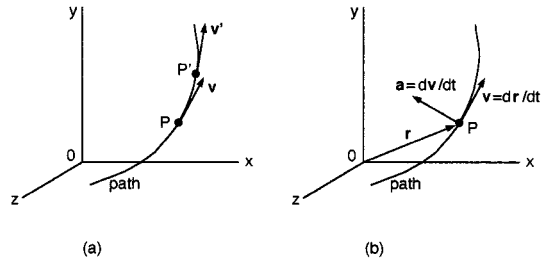


FIGURE 1.3.3 Vector method for a particle.

The mathematical method is based on determining **v** and **a** as functions of the position vector **r**. Note that the time derivatives of unit vectors are zero when the xyz coordinate system is fixed. The scalar components $(\dot{x}, \dot{y}, \ddot{x}, \ldots)$ can be determined from the appropriate scalar equations previously presented that only include the quantities relevant to the coordinate direction considered.

$$\mathbf{r} = x\mathbf{i} + y\mathbf{j} + z\mathbf{k}$$

$$\mathbf{v} = \frac{d\mathbf{r}}{dt} = \frac{dx}{dt}\mathbf{i} + \frac{dy}{dt}\mathbf{j} + \frac{dz}{dt}\mathbf{k} = \dot{x}\mathbf{i} + \dot{y}\mathbf{j} + \dot{z}\mathbf{k} \tag{1.3.7}$$

$$\mathbf{a} = \frac{d\mathbf{v}}{dt} = \frac{d^2x}{dt^2}\mathbf{i} + \frac{d^2y}{dt^2}\mathbf{j} + \frac{d^2z}{dt^2}\mathbf{k} = \ddot{x}\mathbf{i} + \ddot{y}\mathbf{j} + \ddot{z}\mathbf{k}$$

There are a few key points to remember when considering curvilinear motion. First, the instantaneous velocity vector is *always* tangent to the path of the particle. Second, the speed of the particle is the magnitude of the velocity vector. Third, the acceleration vector is *not* tangent to the path of the particle and not collinear with **v** in curvilinear motion.

Tangential and Normal Components

Tangential and normal components are useful in analyzing velocity and acceleration. Figure 1.3.4 illustrates the method and Equation 1.3.8 is the governing equations for it.

$$\mathbf{v} = v\mathbf{n}_t$$

$$\mathbf{a} = a_t\mathbf{n}_t + a_n\mathbf{n}_n$$

$$a_t = \frac{dv}{dt} \qquad a_n = \frac{v^2}{\rho} \tag{1.3.8}$$

$$\rho = \frac{\left[1 + (dy/dx)^2\right]^{3/2}}{d^2y/dx^2}$$

$\rho = r$ = constant for a circular path

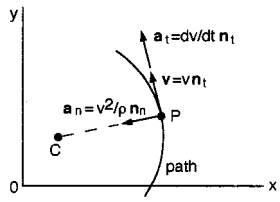


FIGURE 1.3.4 Tangential and normal components. *C* is the center of curvature.

The *osculating plane* contains the unit vectors $\mathbf{n}_t$ and $\mathbf{n}_n$, thus defining a plane. When using normal and tangential components, it is common to forget to include the component of normal acceleration, especially if the particle travels at a constant speed along a curved path.

For a particle that moves in circular motion,

$$v = r\dot{\theta} = r\omega$$
$$a_t = \frac{dv}{dt} = r\ddot{\theta} = r\alpha \tag{1.3.9}$$
$$a_n = \frac{v^2}{r} = r\dot{\theta}^2 = r\omega^2$$

Motion of a Particle in Polar Coordinates

Sometimes it may be best to analyze particle motion by using polar coordinates as follows (Figure 1.3.5):

$$\mathbf{v} = \dot{r}\mathbf{n}_r + r\dot{\theta}\mathbf{n}_\theta \qquad (\text{always tangent to the path})$$
$$\frac{d\theta}{dt} = \dot{\theta} = \omega, \ \text{rad/s} \tag{1.3.10}$$
$$\mathbf{a} = \left(\ddot{r} - r\dot{\theta}^2\right)\mathbf{n}_r + \left(r\ddot{\theta} + 2\dot{r}\dot{\theta}\right)\mathbf{n}_\theta$$

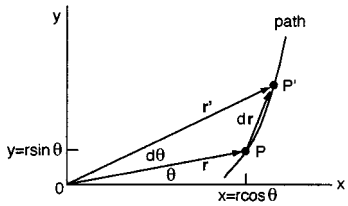


FIGURE 1.3.5 Motion of a particle in polar coordinates.

For a particle that moves in circular motion the equations simplify to

$$\frac{d\dot{\theta}}{dt} = \ddot{\theta} = \dot{\omega} = \alpha, \ \text{rad/s}^2$$
$$\mathbf{v} = r\dot{\theta}\mathbf{n}_\theta \tag{1.3.11}$$
$$\mathbf{a} = -r\dot{\theta}^2\mathbf{n}_r + r\ddot{\theta}\mathbf{n}_\theta$$

Motion of a Particle in Cylindrical Coordinates

Cylindrical coordinates provide a means of describing three-dimensional motion as illustrated in Figure 1.3.6.

$$\mathbf{v} = \dot{r}\mathbf{n}_r + r\dot{\theta}\mathbf{n}_\theta + \dot{z}\mathbf{k}$$
$$\mathbf{a} = \left(\ddot{r} - r\dot{\theta}^2\right)\mathbf{n}_r + \left(r\ddot{\theta} + 2\dot{r}\dot{\theta}\right)\mathbf{n}_\theta + \ddot{z}\mathbf{k} \tag{1.3.12}$$

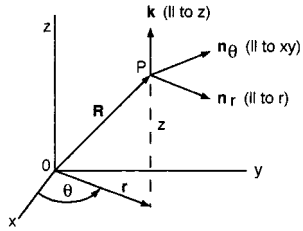


FIGURE 1.3.6 Motion of a particle in cylindrical coordinates.

Motion of a Particle in Spherical Coordinates

Spherical coordinates are useful in a few special cases but are difficult to apply to practical problems. The governing equations for them are available in many texts.

Relative Motion of Particles in Two and Three Dimensions

Figure 1.3.7 shows relative motion in two and three dimensions. This can be used in analyzing the translation of coordinate axes. Note that the unit vectors of the coordinate systems are the same. Subscripts are arbitrary but must be used consistently since $\mathbf{r}_{B/A} = -\mathbf{r}_{A/B}$ etc.

$$\begin{aligned} \mathbf{r}_B &= \mathbf{r}_A + \mathbf{r}_{B/A} \\ \mathbf{v}_B &= \mathbf{v}_A + \mathbf{v}_{B/A} \\ \mathbf{a}_B &= \mathbf{a}_A + \mathbf{a}_{B/A} \end{aligned} \tag{1.3.13}$$

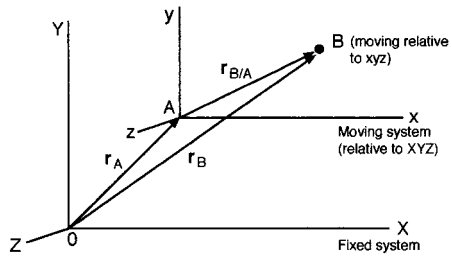


FIGURE 1.3.7 Relative motion using translating coordinates.

Kinetics of Particles

Kinetics combines the methods of kinematics and the forces that cause the motion. There are several useful methods of analysis based on Newton's second law.

Newton's Second Law

The magnitude of the acceleration of a particle is directly proportional to the magnitude of the resultant force acting on it, and inversely proportional to its mass. The direction of the acceleration is the same as the direction of the resultant force.

$$\mathbf{F} = m\mathbf{a} \tag{1.3.14}$$

where m is the particle's mass. There are three key points to remember when applying this equation.

1. **F** is the resultant force.
2. **a** is the acceleration of a single particle (use $\mathbf{a}_C$ for the center of mass for a system of particles).
3. The motion is in a nonaccelerating reference frame.

Equations of Motion

The **equations of motion** for vector and scalar notations in rectangular coordinates are

$$\sum \mathbf{F} = m\mathbf{a}$$
$$\sum F_x = ma_x \qquad \sum F_y = ma_y \qquad \sum F_z = ma_z \tag{1.3.15}$$

The equations of motion for tangential and normal components are

$$\sum F_n = ma_n = m\frac{v^2}{\rho}$$
$$\sum F_t = ma_t = m\dot{v} = mv\frac{dv}{ds} \tag{1.3.16}$$

The equations of motion in a polar coordinate system (radial and transverse components) are

$$\sum F_r = ma_r = m\left(\ddot{r} - r\dot{\theta}^2\right)$$
$$\sum F_\theta = ma_\theta = m\left(r\ddot{\theta} - 2\dot{r}\dot{\theta}\right) \tag{1.3.17}$$

Procedure for Solving Problems

1. Draw a free-body diagram of the particle showing all forces. (The free-body diagram will look unbalanced since the particle is not in static equilibrium.)
2. Choose a convenient nonaccelerating reference frame.
3. Apply the appropriate equations of motion for the reference frame chosen to calculate the forces or accelerations applied to the particle.
4. Use kinematics equations to determine velocities and/or displacements if needed.

Work and Energy Methods

Newton's second law is not always the most convenient method for solving a problem. Work and energy methods are useful in problems involving changes in displacement and velocity, if there is no need to calculate accelerations.

Work of a Force

The total work of a force **F** in displacing a particle *P* from position 1 to position 2 along any path is

$$U_{12} = \int_1^2 \mathbf{F} \cdot d\mathbf{r} = \int_1^2 \left(F_x\, dx + F_y\, dy + F_z\, dz\right) \tag{1.3.18}$$

Potential and Kinetic Energies

Gravitational potential energy: $U_{12} = \int_1^2 W\, dy = Wh = V_g$, where W = weight and h = vertical elevation difference.

Elastic potential energy: $U = \int_{x_1}^{x_2} kx\, dx = \frac{1}{2}k(x_2^2 - x_1^2) = V_e$, where k = spring constant.

Kinetic energy of a particle: $T = 1/2mv^2$, where m = mass and v = magnitude of velocity.

Kinetic energy can be related to work by the *principle of work and energy*,

$$U_{12} = T_2 - T_1 \tag{1.3.19}$$

where U_{12} is the work of a force on the particle moving it from position 1 to position 2, T_1 is the kinetic energy of the particle at position 1 (initial kinetic energy), and T_2 is the kinetic energy of the particle at position 2 (final kinetic energy).

Power

Power is defined as work done in a given time.

$$\text{power} = \frac{dU}{dt} = \frac{\mathbf{F} \cdot d\mathbf{r}}{dt} = \mathbf{F} \cdot \mathbf{v} \tag{1.3.20}$$

where $\mathbf{v}$ is velocity.

Important units and conversions of power are

$$1 \text{ W} = 1 \text{ J/s} = 1 \text{ N} \cdot \text{m/s}$$

$$1 \text{ hp} = 550 \text{ ft} \cdot \text{lb/s} = 33{,}000 \text{ ft} \cdot \text{lb/min} = 746 \text{ W}$$

$$1 \text{ ft} \cdot \text{lb/s} = 1.356 \text{ J/s} = 1.356 \text{ W}$$

Advantages and Disadvantages of the Energy Method

There are four advantages to using the energy method in engineering problems:

1. Accelerations do not need to be determined.
2. Modifications of problems are easy to make in the analysis.
3. Scalar quantities are summed, even if the path of motion is complex.
4. Forces that do not do work are ignored.

The main disadvantage of the energy method is that quantities of work or energy cannot be used to determine accelerations or forces that do no work. In these instances, Newton's second law has to be used.

Conservative Systems and Potential Functions

Sometimes it is useful to assume a conservative system where friction does not oppose the motion of the particle. The work in a conservative system is independent of the path of the particle, and potential energy is defined as

$$\underbrace{U_{12}}_{\substack{\text{work of } \mathbf{F} \\ \text{from 1 to 2}}} = \underbrace{-\Delta V}_{\substack{\text{difference of potential} \\ \text{energies at 1 and 2}}}$$

A special case is where the particle moves in a closed path. One trip around the path is called a *cycle*.

$$U = \oint dU = \oint \mathbf{F} \cdot d\mathbf{r} = \oint \left(F_x \, dx + F_y \, dy + F_z \, dz \right) = 0 \tag{1.3.21}$$

In advanced analysis differential changes in the potential energy function (V) are calculated by the use of partial derivatives,

$$\mathbf{F} = F_x \mathbf{i} + F_y \mathbf{j} + F_z \mathbf{k} = -\left(\frac{\partial V}{\partial x} \mathbf{i} + \frac{\partial V}{\partial y} \mathbf{j} + \frac{\partial V}{\partial z} \mathbf{k} \right)$$

Conservation of Mechanical Energy

Conservation of mechanical energy is assumed if kinetic energy (T) and potential energy (V) change back and forth in a conservative system (the dissipation of energy is considered negligible). Equation 1.3.22 formalizes such a situation, where position 1 is the initial state and position 2 is the final state. The reference (datum) should be chosen to reduce the number of terms in the equation.

$$T_1 + V_1 = T_2 + V_2 \tag{1.3.22}$$

Linear and Angular Momentum Methods

The concept of linear momentum is useful in engineering when the accelerations of particles are not known but the velocities are. The linear momentum is derived from Newton's second law,

$$\mathbf{G} = m\mathbf{v} \tag{1.3.23}$$

The time rate of change of linear momentum is equal to force. When $m\mathbf{v}$ is constant, the conservation of momentum equation results,

$$\sum \mathbf{F} = \dot{\mathbf{G}} = \frac{d}{dt}(m\mathbf{v})$$
$$\sum \mathbf{F} = 0 \qquad m\mathbf{v} = \text{constant} \qquad (\text{conservation of momentum}) \tag{1.3.24}$$

The method of angular momentum is based on the momentum of a particle about a fixed point, using the vector product in the general case (Figure 1.3.8).

$$\mathbf{H}_O = \mathbf{r} \times m\mathbf{v} \tag{1.3.25}$$

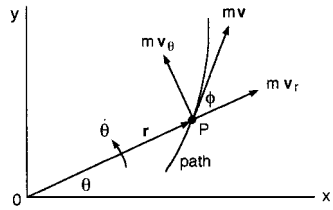


FIGURE 1.3.8 Definition of angular momentum for a particle.

The angular momentum equation can be solved using a scalar method if the motion of the particle remains in a plane,

$$\mathbf{H}_O = mrv\sin\phi = mrv_\theta = mr^2\dot{\theta}$$

If the particle does not remain in a plane, then the general space motion equations apply. They are derived from the cross-product $\mathbf{r} \times m\mathbf{v}$,

$$\mathbf{H}_O = H_x\mathbf{i} + H_y\mathbf{j} + H_z\mathbf{k}$$

$$H_x = m\left(yv_z - zv_y\right)$$

$$H_y = m\left(zv_x - xv_z\right)$$

$$H_z = m\left(xv_y - yv_x\right) \qquad (1.3.25a)$$

Time Rate of Change of Angular Momentum

In general, a force acting on a particle changes its angular momentum: *the time rate of change of angular momentum of a particle is equal to the sum of the moments of the forces acting on the particle.*

Vectors: $$\dot{\mathbf{H}}_O = \frac{d}{dt}(\mathbf{r} \times m\mathbf{v}) = \mathbf{r} \times \sum \mathbf{F} = \sum \mathbf{H}_O \qquad (1.3.26)$$

Scalars: $$\sum M_x = \dot{H}_x \qquad \sum M_y = \dot{H}_y \qquad \sum M_z = \dot{H}_z$$

$$\sum \mathbf{M}_O = 0 \qquad \mathbf{H}_O = \mathbf{r} \times m\mathbf{v} = \text{constant}$$

$$(\text{conservation of angular momentum}) \qquad (1.3.27)$$

A special case is when the sum of the moments about point *O* is zero. This is the conservation of angular momentum. In this case (motion under a central force), if the distance *r* increases, the velocity must decrease, and vice versa.

Impulse and Momentum

Impulse and momentum are important in considering the motion of particles in impact. The linear impulse and momentum equation is

$$\underbrace{\int_{t_1}^{t_2} \mathbf{F}\, dt}_{\text{impulse}} = \underbrace{m\mathbf{v}_2}_{\substack{\text{final}\\ \text{momentum}}} - \underbrace{m\mathbf{v}_1}_{\substack{\text{initial}\\ \text{momentum}}} \qquad (1.3.28)$$

Conservation of Total Momentum of Particles

Conservation of total momentum occurs when *the initial momentum of n particles is equal to the final momentum of those same n particles,*

$$\underbrace{\sum_i^n \left(m_i\mathbf{v}_i\right)_1}_{\substack{\text{total initial}\\ \text{momentum at time } t_1}} = \underbrace{\sum_i^n \left(m_i\mathbf{v}_i\right)_2}_{\substack{\text{total final}\\ \text{momentum at time } t_2}} \qquad (1.3.29)$$

When considering the response of two deformable bodies to direct central impact, the coefficient of restitution is used. This coefficient *e* relates the initial velocities of the particles to the final velocities,

$$e = \frac{v_{Bf} - v_{Af}}{v_A - v_B} = \frac{|\text{relative velocity of separation}|}{|\text{relative velocity of approach}|} \qquad (1.3.30)$$

For real materials, $0 < e < 1$. If both bodies are *perfectly elastic*, $e = 1$, and if either body is *perfectly plastic*, $e = 0$.

Kinetics of Systems of Particles

There are three distinct types of systems of particles: discrete particles, continuous particles in fluids, and continuous particles in rigid or deformable bodies. This section considers methods for discrete particles that have relevance to the mechanics of solids. Methods involving particles in rigid bodies will be discussed in later sections.

Newton's Second Law Applied to a System of Particles

Newton's second law can be extended to systems of particles,

$$\sum_{i=1}^{n} \mathbf{F}_i = \sum_{i=1}^{n} m_i \mathbf{a}_i \tag{1.3.31}$$

Motion of the Center of Mass

The center of mass of a system of particles moves under the action of internal and external forces as if the total mass of the system and all the external forces were at the center of mass. Equation 1.3.32 defines the position, velocity, and acceleration of the center of mass of a system of particles.

$$m\mathbf{r}_C = \sum_{i=1}^{n} m_i \mathbf{r}_i \qquad m\mathbf{v}_C = \sum_{i=1}^{n} m_i \mathbf{v}_i \qquad m\mathbf{a}_C = \sum_{i=1}^{n} m_i \mathbf{a}_i \qquad \sum \mathbf{F} = m\mathbf{a}_C \tag{1.3.32}$$

Work and Energy Methods for a System of Particles

Gravitational Potential Energy. The gravitational potential energy of a system of particles is the sum of the potential energies of the individual particles of the system.

$$V_g = g\sum_{i=1}^{n} m_i y_i = \sum_{i=1}^{n} W_i y_i = mgy_C = Wy_C \tag{1.3.33}$$

where g = acceleration of gravity
y_C = vertical position of center of mass with respect to a reference level

Kinetic Energy. The kinetic energy of a system of particles is the sum of the kinetic energies of the individual particles of the system with respect to a fixed reference frame,

$$T = \frac{1}{2}\sum_{i=1}^{n} m_i v_i^2 \tag{1.3.34}$$

A translating reference frame located at the mass center C of a system of particles can be used advantageously, with

$$T = \underbrace{\frac{1}{2} m v_C^2}_{\substack{\text{motion of total} \\ \text{mass imagined to} \\ \text{be concentrated at C}}} + \underbrace{\frac{1}{2}\sum_{i=1}^{n} m_i v_i'^2}_{\substack{\text{motion of all} \\ \text{particles relative to } C}} \qquad (v' \text{ are with respect to a translating frame}) \tag{1.3.35}$$

Work and Energy

The work and energy equation for a system of particles is similar to the equation stated for a single particle.

$$\sum_{i=1}^{n} U_i' = \sum_{i=1}^{n} V_i + \sum_{i=1}^{n} T_i$$
$$U' = \Delta V + \Delta T \tag{1.3.36}$$

Momentum Methods for a System of Particles

Moments of Forces on a System of Particles. The moments of external forces on a system of particles about a point O are given by

$$\sum_{i=1}^{n} \left(\mathbf{r}_i \times \mathbf{F}_i\right) = \sum_{i=1}^{n} \mathbf{M}_{i_O} + \sum_{i=1}^{n} \left(\mathbf{r}_i \times m_i \mathbf{a}_i\right) \tag{1.3.37}$$

Linear and Angular Momenta of a System of Particles. The resultant of the external forces on a system of particles equals the time rate of change of linear momentum of that system.

$$\mathbf{G} = \sum_{i=1}^{n} m_i \mathbf{v}_i \qquad \sum \mathbf{F} = \dot{\mathbf{G}} \tag{1.3.38}$$

The angular momentum equation for a system of particles about a fixed point O is

$$\mathbf{H}_O = \sum_{i=1}^{n} \left(\mathbf{r}_i \times m_i \mathbf{a}_i\right)$$
$$\sum \mathbf{M}_O = \dot{\mathbf{H}}_O = \sum_{i=1}^{n} \left(\mathbf{r}_i \times m_i \mathbf{a}_i\right) \tag{1.3.39}$$

The last equation means that *the resultant of the moments of the external forces on a system of particles equals the time rate of change of angular momentum of that system.*

Angular Momentum about the Center of Mass

The above equations work well for reference frames that are stationary, but sometimes a special approach may be useful, noting that *the angular momentum of a system of particles about its center of mass C is the same whether it is observed from a fixed frame at point O or from the centroidal frame which may be translating but not rotating*. In this case

$$\mathbf{H}_O = \mathbf{H}_C + \mathbf{r}_C \times m\mathbf{v}_C$$
$$\sum \mathbf{M}_O = \dot{\mathbf{H}}_C + \mathbf{r}_C \times m\mathbf{a}_C \tag{1.3.40}$$

Conservation of Momentum

The conservation of momentum equations for a system of particles is analogous to that for a single particle.

$$\left.\begin{aligned}\mathbf{G} &= \text{constant}\\ \mathbf{H}_O &= \text{constant}\\ \mathbf{H}_C &= \text{constant}\end{aligned}\right\} \quad \text{not the same constants in general}$$

Impulse and Momentum of a System of Particles

The linear impulse momentum for a system of particles is

$$\sum_{i=1}^{n}\int_{t_1}^{t_2}\mathbf{F}_i\,dt = \mathbf{G}_2 - \mathbf{G}_1 = m\mathbf{v}_{C_2} - m\mathbf{v}_{C_1} \tag{1.3.41}$$

The angular impulse momentum for a system of particles is

$$\sum_{i=1}^{n}\int_{t_1}^{t_2}\mathbf{M}_{i_O}\,dt = \mathbf{H}_{O_2} - \mathbf{H}_{O_1} \tag{1.3.42}$$

Kinematics of Rigid Bodies

Rigid body kinematics is used when the methods of particle kinematics are inadequate to solve a problem. A rigid body is defined as one in which the particles are rigidly connected. This assumption allows for some similarities to particle kinematics. There are two kinds of rigid body motion, translation and rotation. These motions may occur separately or in combination.

Translation

Figure 1.3.9 models the translational motion of a rigid body.

$$\begin{aligned}\mathbf{r}_B &= \mathbf{r}_A + \mathbf{r}_{B/A} \qquad \left(\mathbf{r}_{B/A} = \text{constant}\right)\\ \mathbf{v}_B &= \dot{\mathbf{r}}_B = \dot{\mathbf{r}}_A = \mathbf{v}_A\\ \mathbf{a}_B &= \dot{\mathbf{v}}_B = \dot{\mathbf{v}}_A = \mathbf{a}_A\end{aligned} \tag{1.3.43}$$

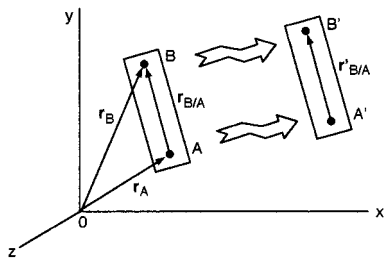


FIGURE 1.3.9 Translational motion of a rigid body.

These equations represent an important fact: *when a rigid body is in translation, the motion of a single point completely specifies the motion of the whole body.*

Rotation about a Fixed Axis

Figure 1.3.10 models a point P in a rigid body rotating about a fixed axis with an angular velocity ω.

The velocity $\mathbf{v}$ of point P is determined assuming that the magnitude of $\mathbf{r}$ is constant,

$$\mathbf{v} = \omega \times \mathbf{r} \tag{1.3.44}$$

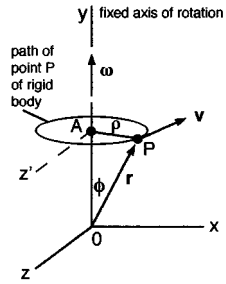


FIGURE 1.3.10 Rigid body rotating about a fixed axis.

The acceleration **a** of point *P* is determined conveniently by using normal and tangential components,

$$\mathbf{a}_P = \underbrace{\alpha \times \mathbf{r}}_{\mathbf{a}_t} + \underbrace{\omega \times (\omega \times \mathbf{r})}_{\mathbf{a}_n} \qquad a_t = \rho\alpha \qquad a_n = \rho\omega^2 \tag{1.3.45}$$

Note that *the angular acceleration α and angular velocity ω are valid for any line perpendicular to the axis of rotation of the rigid body at a given instant.*

Kinematics Equations for Rigid Bodies Rotating in a Plane

For rotational motion with or without a fixed axis, if displacement is measured by an angle θ,

Angular speed: $$\omega = \frac{d\theta}{dt}$$

Angular acceleration: $$\alpha = \frac{d\omega}{dt} = \omega\frac{d\omega}{d\theta}$$

For a constant angular speed ω,

Angular displacement: $\theta = \theta_o + \omega t \quad (\theta = \theta_o \text{ at } t = 0)$

For a constant angular acceleration α,

$$\omega = \omega_o + \alpha t \qquad (\omega = \omega_o \text{ at } t = 0)$$

$$\theta = \theta_o + \omega_o t + \frac{1}{2}\alpha t^2$$

$$\omega^2 = \omega_o^2 + 2\alpha(\theta - \theta_o)$$

Velocities in General Plane Motion

General plane motion of a rigid body is defined by simultaneous translation and rotation in a plane. Figure 1.3.11 illustrates how the velocity of a point A can be determined using Equation 1.3.46, which is based on relative motion of particles.

$$\mathbf{v}_A = \underbrace{\mathbf{v}_B}_{\text{translation}} + \underbrace{\omega \times \mathbf{r}_{A/B}}_{\text{rotation}} \tag{1.3.46}$$

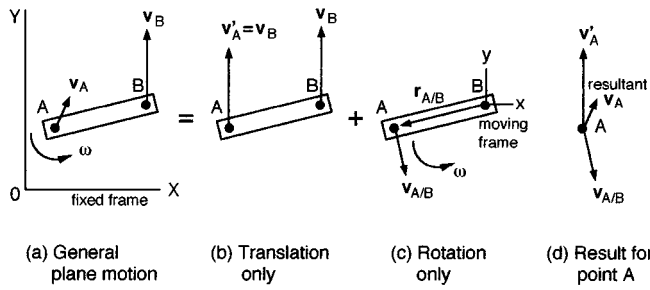


FIGURE 1.3.11 Analysis of velocities in general plane motion.

There are five important points to remember when solving general plane motion problems, including those of interconnected rigid bodies.

1. The angular velocity of a rigid body in plane motion is independent of the reference point.
2. The common point of two or more pin-jointed members must have the same absolute velocity even though the individual members may have different angular velocities.
3. The points of contact in members that are in temporary contact may or may not have the same absolute velocity. If there is sliding between the members, the points in contact have different absolute velocities. The absolute velocities of the contacting particles are always the same if there is no sliding.
4. If the angular velocity of a member is not known, but some points of the member move along defined paths (i.e., the end points of a piston rod), these paths define the directions of the velocity vectors and are useful in the solution.
5. The geometric center of a wheel rolling on a flat surface moves in rectilinear motion. If there is no slipping at the point of contact, the linear distance the center point travels is equal to that portion of the rim circumference that has rolled along the flat surface.

Instantaneous Center of Rotation

The method of *instantaneous center of rotation* is a geometric method of determining the angular velocity when two velocity vectors are known for a given rigid body. Figure 1.3.12 illustrates the method. This procedure can also be used to determine velocities that are parallel to one of the given velocities, by similar triangles.

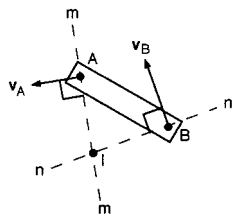


FIGURE 1.3.12 Schematic for instantaneous center of rotation.

Velocities $\mathbf{v}_A$ and $\mathbf{v}_B$ are given; thus the body is rotating about point I at that instant. Point I has zero velocity at that instant, but generally has an acceleration. This method does *not* work for the determination of angular accelerations.

Acceleration in General Plane Motion

Figure 1.3.13 illustrates a method of determining accelerations of points of a rigid body. This is similar to (but more difficult than) the procedure of determining velocities.

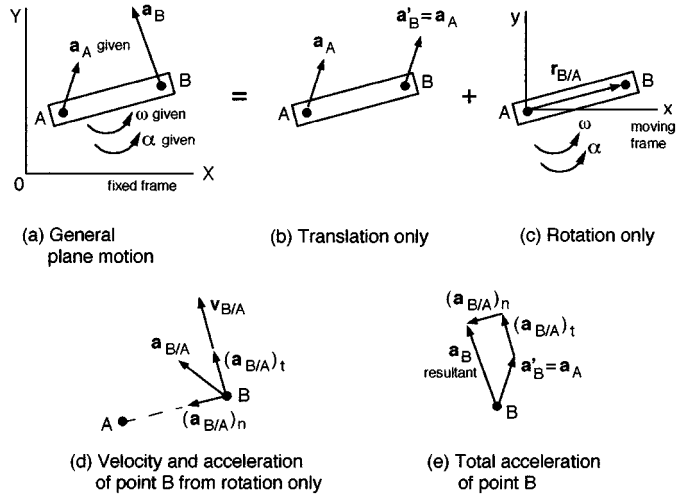


FIGURE 1.3.13 Accelerations in general plane motion.

$$\mathbf{a}_B = \mathbf{a}_A + \alpha \times \mathbf{r}_{B/A} + \omega \times \left(\omega \times \mathbf{r}_{B/A}\right)$$

$$\mathbf{a}_B = \underbrace{\mathbf{a}_A}_{\text{translation}} + \underbrace{\left(\mathbf{a}_{B/A}\right)_t + \left(\mathbf{a}_{B/A}\right)_n}_{\text{rotation}} \tag{1.3.47}$$

There are six key points to consider when solving this kind of a problem.

1. The angular velocity and acceleration of a rigid body in plane motion are independent of the reference point.
2. The common points of pin-jointed members must have the same absolute acceleration even though the individual members may have different angular velocities and angular accelerations.
3. The points of contact in members that are in temporary contact may or may not have the same absolute acceleration. Even when there is no sliding between the members, only the tangential accelerations of the points in contact are the same, while the normal accelerations are frequently different in magnitude and direction.
4. The instantaneous center of zero velocity in general has an acceleration and should *not* be used as a reference point for accelerations unless its acceleration is known and included in the analysis.
5. If the angular acceleration of a member is not known, but some points of the member move along defined paths, the geometric constraints of motion define the directions of normal and tangential acceleration vectors and are useful in the solution.
6. The geometric center of a wheel rolling on a flat surface moves in rectilinear motion. If there is no slipping at the point of contact, the linear acceleration of the center point is parallel to the flat surface and equal to $r\alpha$ for a wheel of radius r and angular acceleration α.

General Motion of a Rigid Body

Figure 1.3.14 illustrates the complex general motion (three-dimensional) of a rigid body. It is important to note that here the angular velocity and angular acceleration vectors are not necessarily in the same direction as they are in general plane motion.

Equations 1.3.48 give the velocity and acceleration of a point on the rigid body. These equations are the same as those presented for plane motion.

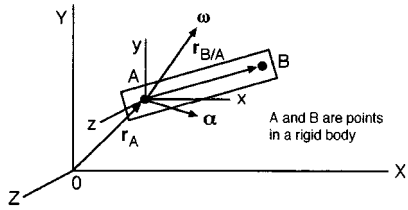


FIGURE 1.3.14 General motion of a rigid body.

$$\mathbf{v}_B = \mathbf{v}_A + \omega \times \mathbf{r}_{B/A}$$

$$\mathbf{a}_B = \mathbf{a}_A + \alpha \times \mathbf{r}_{B/A} + \omega \times \left(\omega \times \mathbf{r}_{B/A}\right) \tag{1.3.48}$$

$$\mathbf{a}_B = \mathbf{a}_A + \left(\mathbf{a}_{B/A}\right)_t + \left(\mathbf{a}_{B/A}\right)_n$$

The most difficult part of solving a general motion problem is determining the angular acceleration vector. There are three cases for the determination of the angular acceleration.

1. The direction of ω is constant. This is plane motion and $\alpha = \dot{\omega}$ can be used in scalar solutions of problems.
2. The magnitude of ω is constant but its direction changes. An example of this is a wheel which travels at a constant speed on a curved path.
3. Both the magnitude and direction of ω change. This is *space motion* since all or some points of the rigid body have three-dimensional paths. An example of this is a wheel which accelerates on a curved path.

A useful expression can be obtained from item 2 and Figure 1.3.15. The rigid body is fixed at point *O* and ω has a constant magnitude. Let ω rotate about the Y axis with angular velocity Ω. The angular acceleration is determined from Equation 1.3.49.

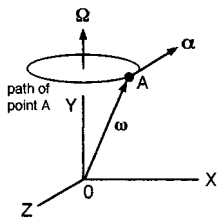


FIGURE 1.3.15 Rigid body fixed at point *O*.

$$\alpha = \frac{d\omega}{dt} = \Omega \times \omega \tag{1.3.49}$$

For *space motion* it is essential to combine the results of items 1 and 2, which provide components of α for the change in magnitude and the change in direction. The following example illustrates the procedure.

Example 9

The rotor shaft of an alternator in a car is in the horizontal plane. It rotates at a constant angular speed of 1500 rpm while the car travels at v = 60 ft/sec on a horizontal road of 400 ft radius (Figure 1.3.16). Determine the angular acceleration of the rotor shaft if v increases at the rate of 8 ft/sec^2.

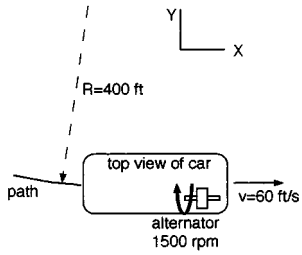


FIGURE 1.3.16 Schematic of shaft's motion.

Solution. There are two components of α. One is the change in the direction of the rotor shaft's ω_x, and the other is the change in magnitude from the acceleration of the car.

1. Component from the change in direction. Determine ω_c of the car.

$$v = r\omega_c$$

$$\omega_c = 0.15 \text{ rad/sec } \mathbf{k}$$

Use Equation 1.3.49:

$$\alpha = \omega_c \times \omega = \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ 0 & 0 & 0.15 \\ 157.1 & 0 & 0 \end{vmatrix} = 23.6\mathbf{j} \text{ rad/sec}^2$$

2. Component from the acceleration of the car. Use Equation 1.3.9:

$$\alpha_C r = a_t$$

$$\alpha_C = 0.02\mathbf{k} \text{ rad/sec}^2$$

The angular acceleration of the rotor shaft is

$$\alpha = (23.6\mathbf{j} + 0.02\mathbf{k}) \text{ rad/sec}^2$$

This problem could also be solved using the method in the next section.

Time Derivative of a Vector Using a Rotating Frame

The basis of determining time derivatives of a vector using a rotating frame is illustrated in Figure 1.3.17.

$$(\mathbf{Q})_{XYZ} = \left(\dot{\mathbf{Q}}\right)_{xyz} + \Omega \times \mathbf{Q}$$

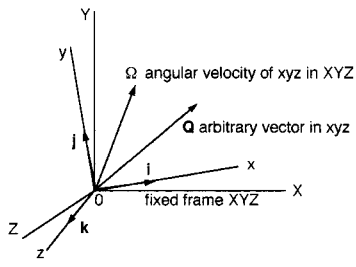


FIGURE 1.3.17 Time derivative of a vector using a rotating reference frame.

Analysis of Velocities and Accelerations Using Rotating and Translating Frames

With the concept of general motion understood, an advantageous method of determining velocities and accelerations is available by the method of rotating reference frames. There are two cases in which this method can be used.

For a common origin of *XYZ* and *xyz*, with **r** being a position vector to a point *P*,

$$\begin{aligned}\mathbf{v}_P &= \mathbf{v}_{xyz} + \Omega \times \mathbf{r} \\ \mathbf{a}_P &= \mathbf{a}_{xyz} + \dot{\Omega} \times \mathbf{r} + \Omega \times (\Omega \times \mathbf{r}) + 2\Omega \times \mathbf{v}_{xyz}\end{aligned} \tag{1.3.50}$$

For the origin *A* of *xyz* translating with respect *XYZ*:

$$\begin{aligned}\mathbf{v}_P &= \mathbf{v}_A + \left(\dot{\mathbf{r}}_{P/A}\right)_{xyz} + \Omega \times \mathbf{r}_{P/A} \\ \mathbf{a}_P &= \mathbf{a}_A + \mathbf{a}_{xyz} + \dot{\Omega} \times \mathbf{r}_{P/A} + \Omega \times \left(\Omega \times \mathbf{r}_{P/A}\right) + 2\Omega \times \mathbf{v}_{xyz}\end{aligned} \tag{1.3.51}$$

where Ω is the angular velocity of the *xyz* frame with respect to *XYZ*. $2\Omega \times \mathbf{v}_{xyz}$ is the Coriolis acceleration.

Kinetics of Rigid Bodies in Plane Motion

Equation of Translational Motion

The fundamental equation for rigid body translation is based on Newton's second law. In Equation 1.3.52, **a** is the acceleration of the center of mass of the rigid body, no matter where the resultant force acts on the body. *The sum of the external forces is equal to the mass of the rigid body times the acceleration of the mass center of the rigid body*, independent of any rotation of the body.

$$\sum \mathbf{F} = m\mathbf{a}_C \tag{1.3.52}$$

Equation of Rotational Motion

Equation 1.3.53 states that *the sum of the external moments on the rigid body is equal to the moment of inertia about an axis times the angular acceleration of the body about that axis*. The angular acceleration α is for the rigid body rotating about an axis. This equation is independent of rigid body translation.

$$\sum \mathbf{M}_C = I_C \alpha \tag{1.3.53}$$

where $\Sigma \mathbf{M}_C = \dot{\mathbf{H}}_C$, $\mathbf{H}_C = I_C \omega$. An application is illustrated in Color Plate 2.

Applications of Equations of Motion

It is important to use the equations of motion properly. For plane motion, three scalar equations are used to define the motion in a plane.

$$\sum F_x = ma_{C_x} \qquad \sum F_y = ma_{C_y} \qquad \sum M_C = I_C \alpha \tag{1.3.54}$$

If a rigid body undergoes only translation,

$$\sum F_x = ma_{C_x} \qquad \sum F_y = ma_{C_y} \qquad \sum M_C = 0 \tag{1.3.55}$$

If the rigid body undergoes pure rotation about the center of mass,

$$\sum F_x = 0 \qquad \sum F_y = 0 \qquad \sum M_C = I_C\alpha \tag{1.3.56}$$

Rigid body motions are categorized according to the constraints of the motion:

1. *Unconstrained Motion*: Equations 1.3.54 are directly applied with all three equations independent of one another.
2. *Constrained Motion*: Equations 1.3.54 are not independent of one another. Generally, a kinematics analysis has to be made to determine how the motion is constrained in the plane. There are two special cases:
   a. Point constraint: the body has a fixed axis.
   b. Line constraint: the body moves along a fixed line or plane.

When considering systems of rigid bodies, it is important to remember that at most only three equations of motion are available from each free-body diagram for plane motion to solve for three unknowns. The motion of interconnected bodies must be analyzed using related free-body diagrams.

Rotation about a Fixed Axis Not Through the Center of Mass

The methods presented above are essential in analyzing rigid bodies that rotate about a fixed axis, which is common in machines (shafts, wheels, gears, linkages). The mass of the rotating body may be nonuniformly distributed as modeled in Figure 1.3.18.

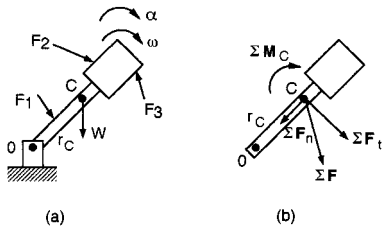


FIGURE 1.3.18 Rotation of a rigid body about a fixed axis.

Note that r_C is the nearest distance between the fixed axis O and the mass center C. The figure also defines the normal and tangential coordinate system used in Equations 1.3.57, which are the scalar equations of motion using normal and tangential components. The sum of the forces must include all reaction forces on the rigid body at the axis of rotation.

$$\sum F_n = mr_C\omega^2 \qquad \sum F_t = mr_C\alpha \qquad \sum M_O = I_O\alpha \tag{1.3.57}$$

General Plane Motion

A body that is translating and rotating is in general plane motion. The scalar equations of motion are given by Equation 1.3.54. If an arbitrary axis A is used to find the resultant moment,

$$\sum \mathbf{M}_A = I_A\alpha + \mathbf{r} \times m\mathbf{a}_C \tag{1.3.58}$$

where *C* is the center of mass. It is a common error to forget to include the cross-product term in the analysis.

There are two special cases in general plane motion, *rolling* and *sliding*.

Figure 1.3.19 shows pure rolling of a wheel without slipping with the center of mass *C* at the geometric center of the wheel. This is called pure rolling of a balanced wheel.

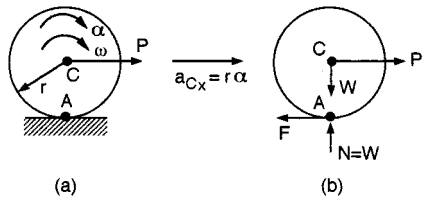
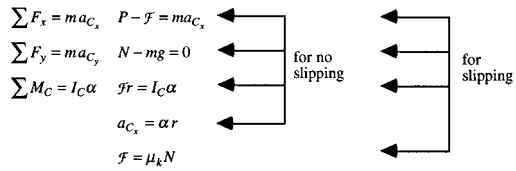


FIGURE 1.3.19 Pure rolling of a wheel.

From this figure the scalar equation of motion results,

$$a_{C_x} = r\alpha \qquad \sum \mathbf{M}_A = I_A \alpha \tag{1.3.59}$$

For balanced wheels either sliding or not sliding, the following schematic is helpful.



If slipping is not certain, assume there is no slipping and check whether $\mathcal{F} \le \mu_s N$. If $\mathcal{F} > \mu_s N$ (not possible; there is sliding), start the solution over using $\mathcal{F} = \mu_k N$ but not using $a_{C_x} = r\alpha$, which is not valid here.

For the problem involving unbalanced wheels (the mass center and geometric center do not coincide), Equations 1.3.60 result.

$$a_{C_x} \neq r\alpha \qquad a_G = r\alpha$$
$$\mathbf{a}_C = \mathbf{a}_G + \mathbf{a}_{C/G} = \mathbf{a}_G + \left(\mathbf{a}_{C/G}\right)_n + \left(\mathbf{a}_{C/G}\right)_t \tag{1.3.60}$$

Energy and Momentum Methods for Rigid Bodies in Plane Motion

Newton's second law in determining kinetics relationships is not always the most efficient, although it always works. As for particles, energy and momentum methods are often useful to analyze rigid bodies in plane motion.

Work of a Force on a Rigid Body

The work of a force acting on a rigid body moving from position 1 to 2 is

$$U_{12} = \int_1^2 \mathbf{F} \cdot d\mathbf{r} = \int_1^2 \mathbf{F} \cdot \mathbf{v}\, dt \tag{1.3.61}$$

Work of a Moment

The work of a moment has a similar form, for angular positions θ,

$$U_{12} = \int_{\theta_1}^{\theta_2} \mathbf{M} \cdot d\theta \tag{1.3.62}$$

In the common case where the moment vector **M** is perpendicular to the plane of motion, $\mathbf{M} \cdot d\theta = M\, d\theta$.

It is important to note those forces that do no work:

1. Forces that act at fixed points on the body do not do work. For example, the reaction at a fixed, frictionless pin does no work on the body that rotates about that pin.
2. A force which is always perpendicular to the direction of the motion does no work.
3. The weight of a body does no work when the body's center of gravity moves in a horizontal plane.
4. The friction force $\mathcal{F}$ at a point of contact on a body that rolls without slipping does no work. This is because the point of contact is the instantaneous center of zero velocity.

Kinetic Energy of a Rigid Body

The kinetic energy of a particle only consists of the energy associated with its translational motion. The kinetic energy of a rigid body also includes a term for the rotational energy of the body,

$$T = T_{trans} + T_{rot} = \frac{1}{2} m v_C^2 + \frac{1}{2} I_C \omega^2 \tag{1.3.63}$$

where C is the center of mass of the rigid body.

The kinetic energy of a rigid body rotating about an arbitrary axis at point O is

$$T = \frac{1}{2} I_O \omega^2$$

Principle of Work and Energy

The principle of work and energy for a rigid body is the same as used for particles with the addition of the rotational energy terms.

$$T_2 = T_1 + U_{12} \tag{1.3.64}$$

where T_1 = initial kinetic energy of the body
T_2 = final kinetic energy of the body
U_{12} = work of all external forces and moments acting on the body moving from position 1 to 2

This method is advantageous when displacements and velocities are the desired quantities.

Conservation of Energy

The conservation of energy in a conservative rigid body system is

$$T_1 + V_1 = T_2 + V_2 \tag{1.3.65}$$

where T = kinetic energy
V = total potential energy (gravitational and elastic)

Power

The net power supplied to or required of the system is

$$\text{power} = \dot{T}_{trans} + \dot{T}_{rot} + \dot{V}_g + \dot{V}_e \tag{1.3.66}$$

This can be calculated by taking time derivatives of the kinetic and potential energy terms. Each term is considered positive when it represents the power supplied to the system and negative when power is taken from the system.

Impulse and Momentum of a Rigid Body

Impulse and momentum methods are particularly useful when time and velocities are of interest. Figure 1.3.20 shows how rigid bodies are to be considered for this kind of analysis. Notice that rotational motion of the rigid body must be included in the modeling.

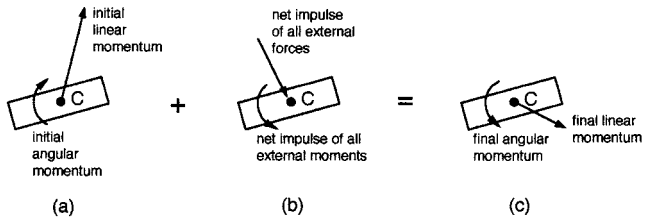


FIGURE 1.3.20 Impulse and momentum for rigid bodies.

The impulse of the external forces in the given interval is

$$\int_{t_1}^{t_2} \sum \mathbf{F}\, dt = m_{C_2}\left(\mathbf{v}_{C_2} - \mathbf{v}_{C_1}\right) \tag{1.3.67}$$

where t is time, C is the center of mass, and $\sum\mathbf{F}$ includes all external forces.

The impulse of the external moments in the given interval is

$$\int_{t_1}^{t_2} \sum \mathbf{M}_C\, dt = \mathbf{H}_{C_2} - \mathbf{H}_{C_1} \tag{1.3.68}$$

For plane motion, if $\sum\mathbf{M}$ is parallel to ω, the scalar expressions are

$$\begin{aligned} &\int_{t_1}^{t_2} \sum M_C\, dt = I_C\left(\omega_2 - \omega_1\right) \\ &\int_{t_1}^{t_2} \sum M_O\, dt = I_O\left(\omega_2 - \omega_1\right) \quad \text{for rotation about a fixed point } O \end{aligned} \tag{1.3.69}$$

Impulse and Momentum of a System of Rigid Bodies

A system of rigid bodies can be analyzed using one of the two following procedures, illustrated in Figure 1.3.21.

1. Apply the principle of impulse and momentum to each rigid member separately. The mutual forces acting between members must be included in the formulation of the solution.
2. Apply the principle of impulse and momentum to the entire system of bodies, ignoring the mutual forces between members.

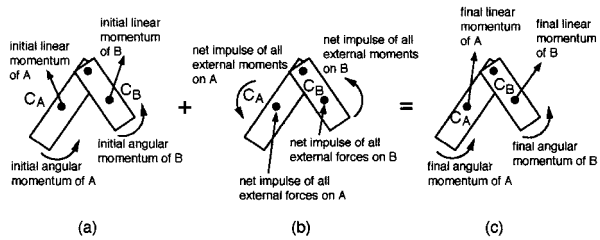


FIGURE 1.3.21 System of rigid bodies.

Conservation of Momentum

The principle of conservation of linear and angular momentum of particles can be extended to rigid bodies that have no external forces or moments acting on them. The conservation of linear momentum means that the center of mass C moves at a constant speed in a constant direction,

$$\sum \mathbf{F} = 0 \Rightarrow \Delta \mathbf{G} = 0$$
$$\mathbf{v}_{C_1} = \mathbf{v}_{C_2} \tag{1.3.70}$$

Likewise, for conservation of angular momentum of rigid bodies,

$$\sum \mathbf{M} = 0 \Rightarrow \Delta \mathbf{H}_C = 0$$
$$I_C \omega_1 = I_C \omega_2 \tag{1.3.71}$$

For a system of rigid bodies, use the same fixed reference point O for all parts of the system. Thus, for plane motion,

$$\Delta \mathbf{H}_O = 0 \qquad I_O \omega_1 = I_O \omega_2 \tag{1.3.72}$$

There are two important points to remember when using these equations. First, $\Delta \mathbf{H}_C = 0$ does not imply that $\Delta \mathbf{H}_O = 0$, or vice versa. Second, conservation of momentum does not require the simultaneous conservation of both angular and linear momenta (for example, there may be an angular impulse while linear momentum is conserved).

Kinetics of Rigid Bodies in Three Dimensions

The concepts of plane rigid body motion can be extended to the more complicated problems in three dimensions, such as of gyroscopes and jet engines. This section briefly covers some fundamental topics. There are many additional topics and useful methods that are included in the technical literature.

Angular Momentum in Three Dimensions

For analyzing three-dimensional angular momentum, three special definitions are used. These can be visualized by considering a spinning top (Figure 1.3.22).

Precession — rotation of the angular velocity vector about the y axis.
Space Cone — locus of the absolute positions of the instantaneous axis of rotation.
Body Cone — locus of the positions of the instantaneous axis relative to the body. The body cone appears to roll on the space cone (not shown here).

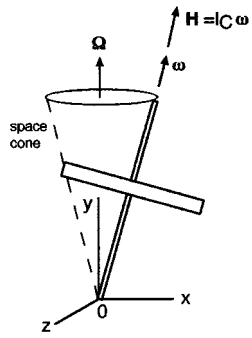


FIGURE 1.3.22 Motion of an inclined, spinning top.

Equations 1.3.73 provide the scalar components of the total angular momentum.

$$\begin{aligned} H_x &= I_x\omega_x - I_{xy}\omega_y - I_{xz}\omega_z \\ H_y &= -I_{xy}\omega_x + I_y\omega_y - I_{yz}\omega_z \\ H_z &= -I_{zx}\omega_x - I_{zy}\omega_y + I_z\omega_z \end{aligned} \tag{1.3.73}$$

Impulse and Momentum of a Rigid Body in Three-Dimensional Motion

The extension of the planar motion equations of impulse and momentum to three dimensions is straightforward.

$$\text{System momenta} = \begin{cases} \text{linear momentum of mass center } (\mathbf{G}) \\ \text{angular momentum about mass center } (\mathbf{H}_C) \end{cases} \tag{1.3.74}$$

where **G** and **H** have different units. The principle of impulse and momentum is applied for the period of time t_1 to t_2,

$$\begin{aligned} \mathbf{G}_2 &= \mathbf{G}_1 + (\text{external linear impulses})\Big|_1^2 \\ \mathbf{H}_{C_2} &= \mathbf{H}_{C_1} + (\text{external angular impulses})\Big|_1^2 \end{aligned} \tag{1.3.75}$$

Kinetic Energy of a Rigid Body in Three-Dimensional Motion

The total kinetic energy of a rigid body in three dimensions is

$$T = \underbrace{\frac{1}{2}mv_C^2}_{\substack{\text{translation} \\ \text{of mass center}}} + \underbrace{\frac{1}{2}\omega \cdot \mathbf{H}_C}_{\substack{\text{rotation about} \\ \text{mass center}}} \tag{1.3.76}$$

For a rigid body that has a fixed point *O*,

$$T = \frac{1}{2}\omega \cdot \mathbf{H}_O \tag{1.3.77}$$

Equations of Motion in Three Dimensions

The equations of motion for a rigid body in three dimensions are extensions of the equations previously stated.

$$\sum \mathbf{F} = m\mathbf{a}_C$$
$$\sum \mathbf{M}_C = \dot{\mathbf{H}}_C = \left(\dot{\mathbf{H}}_C\right)_{xyz} + \Omega \times \mathbf{H}_C \tag{1.3.78}$$

where $\mathbf{a}_C$ = acceleration of mass center
$\mathbf{H}_C$ = angular momentum of the body about its mass center
xyz = frame fixed in the body with origin at the mass center
Ω = angular velocity of the xyz frame with respect to a fixed XYZ frame

Note that an arbitrary fixed point O may be used for reference if done consistently.

Euler's Equations of Motion

Euler's equations of motion result from the simplification of allowing the xyz axes to coincide with the principal axes of inertia of the body.

$$\sum M_x = I_x\dot{\omega}_x - \left(I_y - I_z\right)\omega_y\omega_z$$
$$\sum M_y = I_y\dot{\omega}_y - \left(I_z - I_x\right)\omega_z\omega_x \tag{1.3.79}$$
$$\sum M_z = I_z\dot{\omega}_z - \left(I_x - I_y\right)\omega_x\omega_y$$

where all quantities must be evaluated with respect to the appropriate principal axes.

Solution of Problems in Three-Dimensional Motion

In order to solve a three-dimensional problem it is necessary to apply the six independent scalar equations.

$$\sum F_x = ma_{C_x} \qquad \sum F_y = ma_{C_y} \qquad \sum F_z = ma_{C_z}$$

$$\sum M_x = \dot{H}_x + \omega_y H_z - \omega_z H_y$$
$$\sum M_y = \dot{H}_y + \omega_z H_x - \omega_x H_z \tag{1.3.80}$$
$$\sum M_z = \dot{H}_z + \omega_x H_y - \omega_y H_x$$

These equations are valid in general. Some common cases are briefly stated.

Unconstrained motion. The six governing equations should be used with xyz axes attached at the center of mass of the body.

Motion of a body about a fixed point. The governing equations are valid for a body rotating about a noncentroidal fixed point O. The reference axes xyz must pass through the fixed point to allow using a set of moment equations that do not involve the unknown reactions at O.

Motion of a body about a fixed axis. This is the generalized form of plane motion of an arbitrary rigid body. The analysis of unbalanced wheels and shafts and corresponding bearing reactions falls in this category.

1.4 Vibrations

Bela I. Sandor with assistance by Stephen M. Birn

Vibrations in machines and structures should be analyzed and controlled if they have undesirable effects such as noise, unpleasant motions, or fatigue damage with potentially catastrophic consequences. Conversely, vibrations are sometimes employed to useful purposes, such as for compacting materials.

Undamped Free and Forced Vibrations

The simplest vibrating system has motion of one degree of freedom (DOF) described by the coordinate x in Figure 1.4.1. (An analogous approach is used for torsional vibrations, with similar results.)

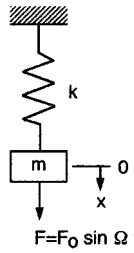


FIGURE 1.4.1 Model of a simple vibrating system.

Assuming that the spring has no mass and that there is no damping in the system, the equation of motion for **free vibration** (motion under internal forces only; $F = 0$) is

$$m\ddot{x} + kx = 0 \qquad \text{or} \qquad \ddot{x} + \omega^2 x = 0 \tag{1.4.1}$$

where $\omega = \sqrt{k/m}$ = natural circular frequency in rad/sec.

The displacement x as a function of time t is

$$x = C_1 \sin\omega t + C_2 \cos\omega t \tag{1.4.2}$$

where C_1 and C_2 are constants depending on the initial conditions of the motion. Alternatively,

$$x = A\sin(\omega t + \phi)$$

where $C_1 = A\cos\phi$, $C_2 = A\sin\phi$, and ϕ is the phase angle, another constant. A complete cycle of the motion occurs in time τ, the *period of simple harmonic motion*,

$$\tau = \frac{2\pi}{\omega} = 2\pi\sqrt{\frac{m}{k}} \ (\text{seconds per cycle})$$

The *frequency* in units of cycles per second (cps) or hertz (Hz) is $f = 1/\tau$.

The simplest case of **forced vibration** is modeled in Figure 1.4.1, with the force F included. Using typical simplifying assumptions as above, the equation of motion for a harmonic force of forcing frequency Ω,

$$m\ddot{x} + kx = F_o \sin\Omega t \tag{1.4.3}$$

The vibrations of a mass m may also be induced by the displacement $d = d_o \sin \Omega t$ of a foundation or another mass M to which m is attached by a spring k. Using the same reference point and axis for both x and d, the equation of motion for m is

$$\begin{aligned} m\ddot{x} + k\left(x - d_o \sin \Omega t\right) &= 0 \\ m\ddot{x} + kx &= kd_o \sin \Omega t \end{aligned} \tag{1.4.4}$$

where d_o is the amplitude of vibration of the moving support M, and Ω is its frequency of motion.

The general solution of the forced vibration in the *steady state* (after the initial, transient behavior) is

$$\begin{aligned} x &= A \sin \Omega t \\ A &= \frac{F_o}{k - m\Omega^2} = \frac{F_o/k}{1 - \left(\Omega/\omega\right)^2} \end{aligned} \tag{1.4.5}$$

where Ω is the forcing frequency and ω is the natural circular frequency of the system of m and k.

Resonance. The amplitude of the oscillations in forced vibrations depends on the frequency ratio Ω/ω. Without damping or physical constraints, the amplitude would become infinite at $\Omega = \omega$, the condition of **resonance**. Dangerously large amplitudes may occur at resonance and at other frequency ratios near the resonant frequency. A *magnification factor* is defined as

$$MF = \frac{F}{F_o/k} = \frac{A}{d_o} = \frac{1}{1 - \left(\Omega/\omega\right)^2} \tag{1.4.6}$$

Several special cases of this are noted:

1. Static loading: $\Omega = 0$, or $\Omega \ll \omega$; MF $\simeq$ 1.
2. Resonance: $\Omega = \omega$; $MF = \infty$.
3. High-frequency excitation: $\Omega \gg \omega$; MF $\simeq$ 0.
4. Phase relationships: The vibration is *in phase* for $\Omega < \omega$, and it is 180° *out of phase* for $\Omega > \omega$.

Damped Free and Forced Vibrations

A vibrating system of one degree of freedom and damping is modeled in Figure 1.4.2. The equation of motion for *damped free vibrations* ($F = 0$) is

$$m\ddot{x} + c\dot{x} + kx = 0 \tag{1.4.7}$$

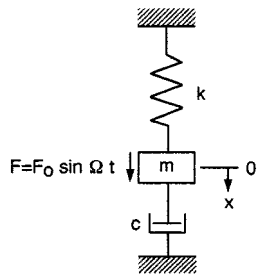


FIGURE 1.4.2 Model of a damped vibrating system.

The displacement x as a function of time t is

$$x = e^{\lambda t} \tag{1.4.8}$$

$$\lambda_{1,2} = \frac{-c}{2m} \pm \sqrt{\left(\frac{c}{2m}\right)^2 - \frac{k}{m}}$$

The value of the coefficient of viscous damping c that makes the radical zero is the *critical damping coefficient* $c_c = 2m\sqrt{k/m} = 2m\omega$. Three special cases of damped free vibrations are noted:

1. Overdamped system: $c > c_c$; the motion is *nonvibratory* or *aperiodic.*
2. Critically damped system: $c = c_c$; this motion is also nonvibratory; x decreases at the fastest rate possible without oscillation of the mass.
3. Underdamped system: $c < c_c$; the roots $\lambda_{1,2}$ are complex numbers; the displacement is

$$x = Ae^{-(c/2m)t}\sin\left(\omega_d t + \phi\right)$$

where A and ϕ are constants depending on the initial conditions, and the *damped natural frequency* is

$$\omega_d = \omega\sqrt{1 - \left(\frac{c}{c_c}\right)^2}$$

The ratio c/c_c is the *damping factor* ζ. The damping in a system is determined by measuring the rate of decay of free oscillations. This is expressed by the *logarithmic decrement* δ, involving any two successive amplitudes x_i and x_{i+1},

$$\delta = \ln\frac{x_i}{x_{i+1}} = \frac{2\pi\zeta}{\sqrt{1-\zeta^2}} \simeq 2\pi\zeta$$

The simplifying approximation for δ is valid for up to about 20% damping ($\zeta \simeq 0.2$).

The *period of the damped vibration* is $\tau_d = 2\pi/\omega_d$. It is a constant, but always larger than the period of the same system without damping. In many real systems the damping is relatively small ($\zeta < 0.2$), where $\tau_d \simeq \tau$ and $\omega_d \simeq \omega$ can be used.

The equation of motion for *damped forced vibrations* (Figure 1.4.2; $F \neq 0$) is

$$m\ddot{x} + c\dot{x} + kx = F_o \sin\Omega t \tag{1.4.9}$$

The solution for steady-state vibration of the system is

$$x = A\sin\left(\Omega t - \phi\right) \tag{1.4.10}$$

where the amplitude and phase angle are from

$$A = \frac{F_o}{\sqrt{\left(c\Omega\right)^2 + \left(k - m\Omega^2\right)^2}}$$

$$\tan\phi = \frac{c\Omega}{k - m\Omega^2}$$

The *magnification factor* for the amplitude of the oscillations is

$$MF = \frac{A}{F_o/k} = \frac{A}{d_o} = \frac{1}{\sqrt{\left[2\zeta(\Omega/\omega)\right]^2 + \left[1-(\Omega/\omega)^2\right]^2}} \tag{1.4.11}$$

This quantity is sketched as a function of the frequency ratio Ω/ω for several damping factors in Figure 1.4.3. Note that the amplitude of vibration is reduced at all values of Ω/ω if the coefficient of damping c is increased in a particular system.

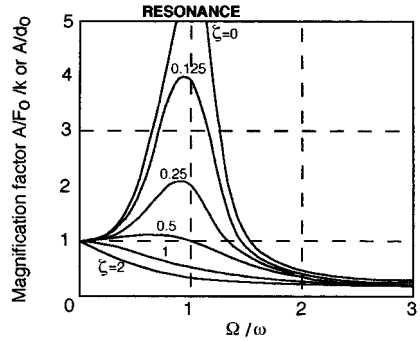


FIGURE 1.4.3 Magnification factor in damped forced vibration.

Vibration Control

Vibration Isolation

It is often desirable to reduce the forces transmitted, or the noise and motions inside or in the neighborhood of vibrating machines and structures. This can be done to some extent within the constraints of space and additional weight and cost by the use of isolators, such as rubber engine mounts and wheel suspension systems in cars. Many kinds of isolating materials and systems are available commercially.

The effectiveness of vibration isolation is expressed by the *transmissibility TR*, the ratio of the force transmitted F_T to the disturbing force F_o. A simple isolation system is modeled as a spring and a dashpot in parallel, for which the transmissibility is given by Equation 1.4.12 and sketched in Figure 1.4.4.

$$TR = \frac{F_T}{F_o} = \frac{\sqrt{1+4\zeta^2(\Omega/\omega)^2}}{\sqrt{\left[1-(\Omega/\omega)^2\right]^2 + 4\zeta^2(\Omega/\omega)^2}} \tag{1.4.12}$$

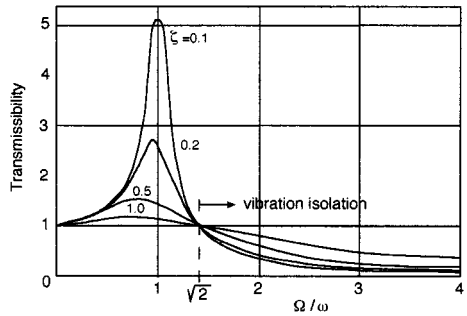


FIGURE 1.4.4 Transmissibility patterns of a vibration isolator.

When damping is negligible,

$$TR \simeq \frac{1}{(\Omega/\omega)^2 - 1}$$

Note from the figure that

1. Vibration isolation occurs at $\Omega/\omega > \sqrt{2}$.
2. Isolation efficiency increases with decreasing stiffness of the isolation mounts.
3. Damping reduces isolation efficiency. However, some damping is normally required if resonance may occur in a system even for short periods.
4. The response curves are essentially independent of damping when Ω/ω is large $(\gtrsim 3)$ and damping is low $(\zeta \lesssim 0.2)$. Here $TR \simeq 1/[(\Omega/\omega)^2 - 1]$.
5. For a system with more than one excitation frequency, the lowest excitation frequency is of primary importance.

The efficiency of an isolating system is defined by the reduction R in transmissibility,

$$R = 1 - TR$$

If a certain reduction R in transmissibility is desired, the appropriate stiffness k of an isolation system is obtained from $\omega = \sqrt{k/m}$ and

$$\frac{\Omega}{\omega} = \sqrt{\frac{2 - R}{1 - R}}$$

A small magnitude of stiffness k makes the reduction R in transmissibility large. It is difficult to achieve isolation for very low excitation frequencies because of the required large static deflections. To obtain highly efficient isolation at low excitation frequencies, a large supporting mass M may be utilized, with the value of $\omega = \sqrt{k/(m + M)}$.

Vibration Absorption

In some cases a vibratory force is purposely generated in a system by a secondary spring-mass system to oppose a primary disturbing force and thereby reduce or eliminate the undesirable net effect. An interesting example of this is the "tuned-mass damper" in a few skyscrapers, designed to counter the oscillatory motions caused by wind. The secondary spring-mass system has disadvantages of its own, such as extra weight, complexity, and effectiveness limited to a single frequency.

Balancing of Rotating Components

The conditions of static or dynamic unbalance of rotating bodies have long been recognized. These can be analyzed by the methods of elementary mechanics, simple tests can be performed in many cases, and adequate corrections can be made routinely to achieve balance, such as for the wheels of automotive vehicles. Three categories of increasing complexity are distinguished.

1. *Static unbalance*. The distributed or lumped masses causing unbalance are in a single axial plane and all on the same side of the axis of rotation (Figure 1.4.5). Thin disks are also in this category. Static unbalance is detected in a static test since the center of gravity of the body is not on the axis, and correction is made by adding or removing mass at a convenient radial distance from the axis.
2. *Static balance with dynamic unbalance*. This may be the case when the masses causing unbalance are in a single axial plane but on opposite sides of the axis of rotation (Figure 1.4.6a). Static balance is achieved if the center of gravity of the body is on the axis, but dynamic unbalance

results from the couple of the unbalance forces ($m\omega^2 r$) during rotation, causing a shaking of the axle.

3. *Static and dynamic unbalance*. This is the general case of unbalance, which can be visualized by letting m_1 and m_2 and the axis of rotation not all lie in the same plane (Figure 1.4.6b).

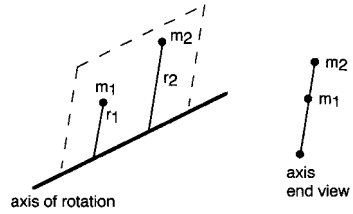


FIGURE 1.4.5 Schematic of static unbalance.

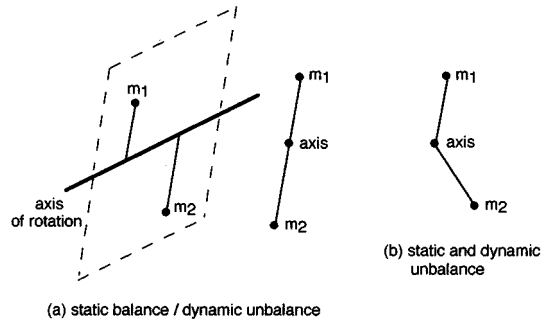


FIGURE 1.4.6 Schematic of two cases of dynamic unbalance.

The magnitude and angular position of a body's unbalance can be determined using a dynamic balancing machine. Here the shaking forces are measured by electronically sensing the small oscillations of the bearings that can be correlated with the position of the body.

Critical Speed of Rotating Shafts

A rotating shaft may become dangerously unstable and whirl with large lateral amplitudes of displacement at a critical speed of rotation. The critical speed, in revolutions per second, corresponds with the natural frequency of lateral vibration of the system. Thus, it can be analytically predicted fairly well and can be safely measured in a real but nonrotating machine with high precision.

If unavoidable, as at startup, the critical speed should be passed over rapidly. Other ways of minimizing the problems of whirling shafts include the proper balancing of rotors and the replacing of bent shafts and worn bearings.

Random Vibrations. Shock Excitation

Many structures are subjected to nonharmonic excitations and respond with transient vibrations rather than steady-state motions. Random vibration is often caused by *shock* excitation, which implies that the loading occurs suddenly, in a short time with respect to the natural period of vibration of the system. Such a loading, typically caused by impact conditions, may be highly irregular in terms of amplitude, waveform, and repetition (Figure 1.4.7), but normally it is possible to extract practically uniform critical events from the loading history for purposes of future design and life prediction.

For most practical purposes, this plot represents aperiodic motion, where the important quantities are the maximum and average large amplitudes and the projected total repetitions (in this case, at the rate of about 1000 per day) over the design life of the structure. The small-amplitude transient vibrations

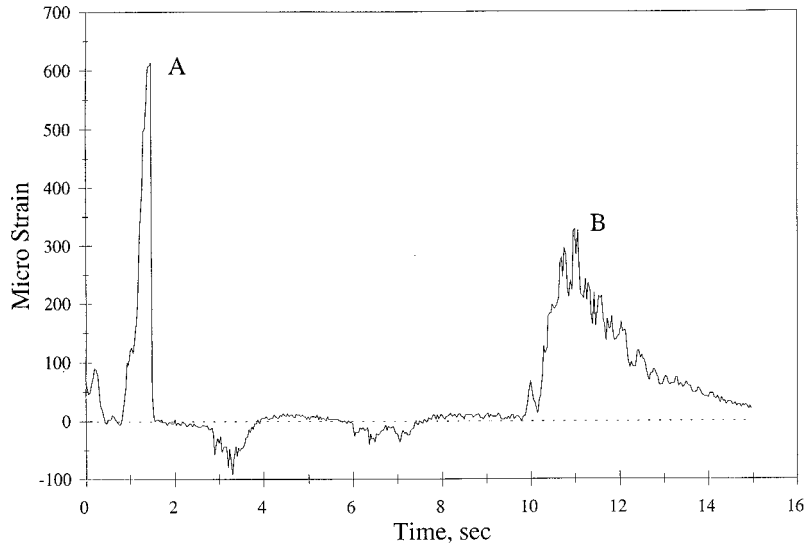


FIGURE 1.4.7 Strain-time history at one strain-gage location on a steel bridge caused by two trucks moving in opposite directions. (A) Garbage truck in the near lane; (B) tractor trailer in the far lane. Weights unknown. (Data courtesy Mark J. Fleming, University of Wisconsin-Madison.)

associated with the large events are likely to be negligible here in terms of both dynamic behavior and fatigue damage, although the relatively large number of small oscillations may cause one to be concerned in some cases.

Random vibrations are difficult to deal with analytically. Numerical methods involving computers are advantageous to obtain response (or shock) spectrums of a system, assuming key parameters and simple models of nonharmonic excitations such as impulsive forces and force step functions. Since the maximum transient response is relatively insensitive to damping, an undamped system is useful in modeling response spectrums. Experimental techniques are needed to verify the analytical predictions, especially when the behavior of a multiple-degree-of-freedom system is determined from the response spectrum of a single-degree-of-freedom system.

Multiple-Degree-of-Freedom Systems. Modal Analysis

The analysis of a system with more than one degree of freedom requires an independent coordinate for each degree of freedom to describe the configurations. Thus, an *n*-degree-of-freedom system has *n* natural frequencies and *n* normal *modes* of vibration. Complex systems can be classified as (1) discrete and lumped-parameter systems with finite numbers of degrees of freedom or (2) continuous elastic bodies of distributed mass with infinite number of degrees of freedom (in theory). A common example of the latter is a vibrating beam, with the first two modes of vibration shown in Figure 1.4.8. Each *nodal point* is a point of zero deflection. Usually the *fundamental natural frequency* (the lowest) is the most important, and only the lowest few frequencies are considered in practice.

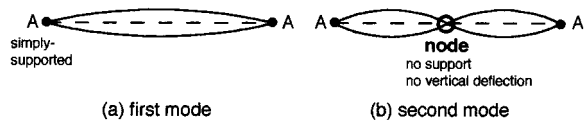


FIGURE 1.4.8 Simply supported beam in two modes of vibration.

A system's harmonic vibrations are its *principal modes*. There are also many ways in which the system can vibrate nonharmonically. Periodic motion of complex wave form can be analyzed as a combination of principal-mode vibrations.

The classical method of mathematical solution and the experimental techniques become increasingly cumbersome and sometimes inaccurate for a system of more than a few degrees of freedom. The recent emergence of sophisticated numerical (finite element; Figure 1.4.9) and experimental (electro-optics) techniques has resulted in significant progress in this area. The synergistic aspects of several new methods are especially remarkable. For example, damage caused by vibrations can significantly affect a system's own modal behavior and, consequently, the rate of damage evolution. Such nonlinear changes of a system can now be investigated and eventually predicted by the hybrid applications of computerized numerical methods, fatigue and fracture mechanics (Section 1.6), and high-speed, noncontacting, full-field vibration and stress imaging (Sections 1.4, "Vibration-Measuring Instruments," and 1.5, "Experimental Stress Analysis and Mechanical Testing"). These enhance the already powerful modern methods of *modal analysis* for accurately describing the response of multiple-degree-of-freedom systems.

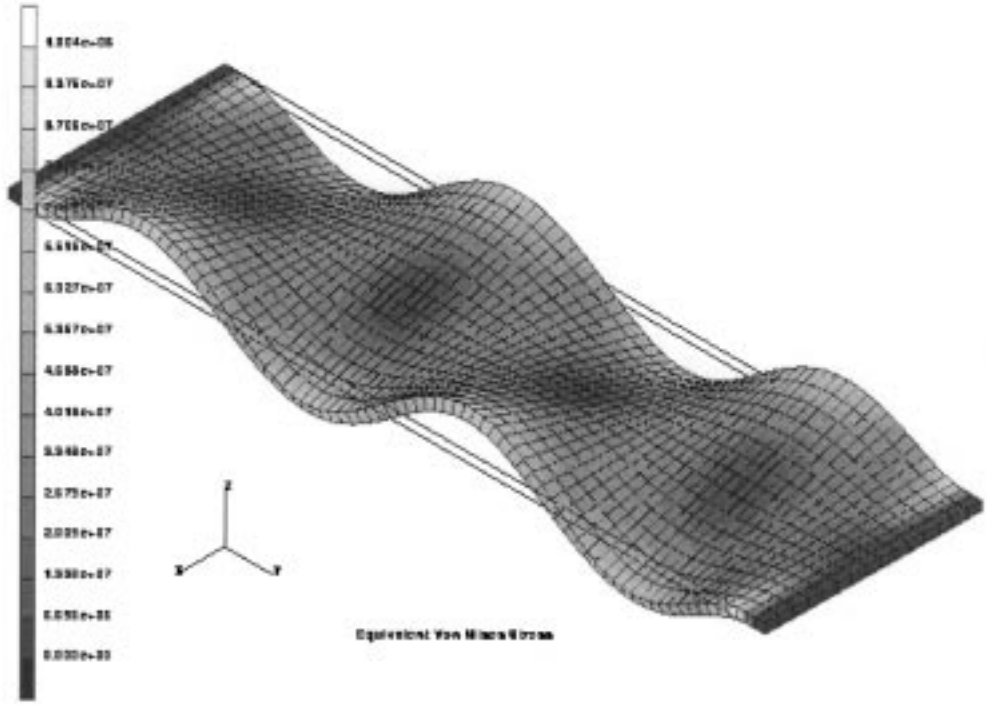


FIGURE 1.4.9 Modal analysis of a vibrating plate. (Photo courtesy David T. Corr, University of Wisconsin-Madison.)

Vibration-Measuring Instruments

There are many kinds of instruments for the experimental investigation of vibrating systems. They range from simple, inexpensive devices to sophisticated electro-optics with lasers or infrared detectors, with the list still expanding in many areas.

The basic quantities of interest regarding a vibrating system are the displacement, velocity, acceleration, and frequency. A typical sensor (or pickup or transducer) for determining these is the piezoelectric accelerometer, which is attached to the vibrating machine or structure to be analyzed. The complete setup normally includes amplifiers, frequency analyzer, oscilloscope, and recorders. An instrumented

impact hammer may be used to provide well-defined impulse excitation to determine the natural frequencies of structures. The frequency analyzer can display the accelerometer output in either the time or the frequency domain.

Other kinds of devices used for vibration sensing include seismic spring-mass systems, electrical-resistance strain gages, and electromagnetic transducers.

Care must be exercised in matching a transducer to the task at hand, since reliable data can be obtained only if the transducer has a "flat-response" frequency region for the measurements of interest. For example, electromagnetic vibrometers (or seismometers) are low-frequency transducers that have low natural frequency compared to the frequency of the motion to be measured. At the other extreme, piezoelectric accelerometers are designed to have higher natural frequency than the frequency to be measured.

It is also important to use transducers of negligible mass compared to the mass of the vibrating system being measured. Very small, light-weight accelerometers are available to satisfy this condition in many cases. There are situations, however, where only noncontacting means of motion measurement provide satisfactory results. Optical techniques are prominent in this area, offering several advantages besides the noncontacting measurement capability. They can be full-field techniques, which means that data may be obtained rapidly from many points on a body using one instrument. They have excellent resolution and precision, and some of them are easy to use. Three kinds of optical instruments are distinguished here for vibratory system analysis, depending on the primary quantity measured:

1. *Displacement measurement*. Holography and speckle pattern imaging have excellent resolution, but they are adversely affected by unstable measuring conditions. They are most useful in laboratory applications.
2. *Velocity measurement*. Laser Doppler systems provide time-resolved, accelerometer-like measurements. They are relatively unaffected by measuring conditions, and are simple and rugged enough to use either in the laboratory or in the field. Several important capabilities of such a vibration pattern imaging system are worth mentioning (Color Plates 3 to 7):
   - Noncontacting; the structure's response is not affected by the instrumentation; applicable in some hazardous environments (hot structures etc.), and short or long range (over 200 m) on natural surfaces
   - Single-point or full-field data acquisition at high resolution from areas of 0.5×0.5 mm to 8×8 m; up to 500 individual points can be programmed
   - Wide frequency range; 0 to 100 kHz (for example, Figure 1.4.10)
   - Sensitivity to a wide range of vibration velocities; 0.005 to 1000 mm/sec
   - Large depth of focus; ±3 m at 10-m working distance
   - Node spacing down to a few millimeters can be resolved
   - Resolution of small displacements, down to the wavelength of the laser source (typically, ≈1 Å)
   - Safe, class II laser system; <1 mW output
   - Conventional signal processing is used to give multipoint modal parameters in familiar format for analytical comparisons
3. *Dynamic stress measurement*. Differential thermography via dynamic thermoelasticity (Figure 1.4.11) has recently become a powerful technique for measuring the modal response of vibrating structures and, uniquely, for directly assessing the structural integrity and durability aspects of the situation. This approach uses high-speed infrared electro-optics and has predictive capability because it can be quantitatively combined with modern fatigue and fracture mechanics methods. For example, it can effectively relate vibration modes to complex fracture modes and damage evolution rates of a real component even under arbitrary and unknown loading with unknown boundary conditions. See Section 1.5, "Experimental Stress Analysis and Mechanical Testing," for more on the dynamic thermoelasticity technique.

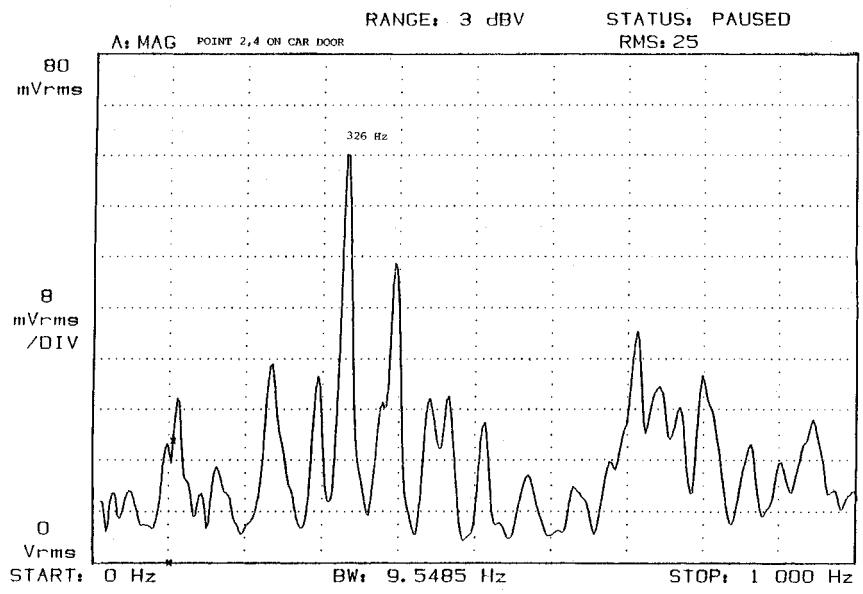


FIGURE 1.4.10 Laser-based, noncontacting vibration analysis of a point on a car door. (Data courtesy of Ometron Inc., Sterling, VA.)

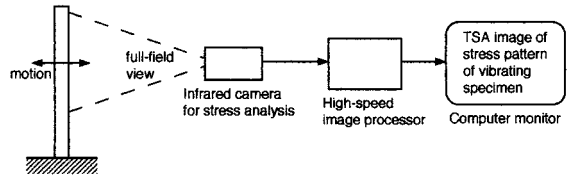


FIGURE 1.4.11 Schematic of modal analysis of a jet engine turbine blade by thermal imaging of the stress field caused by complex vibration. For sample data, see Color Plate 8.

1.5 Mechanics of Materials

Bela I. Sandor

Mechanics of materials, also called strength of materials, provides quantitative methods to determine stresses (the intensity of forces) and strains (the severity of deformations), or overall deformations or load-carrying abilities of components and structures. The stress-strain behavior of materials under a wide range of service conditions must be considered in many designs. It is also crucial to base the analysis on correct modeling of component geometries and external loads. This can be difficult in the case of multiaxial loading, and even more so if time- or temperature-dependent material behaviors must be considered.

Proper modeling involves free-body diagrams and equations of equilibrium. However, it is important to remember that *the equilibrium equations of statics are valid only for forces or for moments of forces*, and not for stresses.

Stress

The intensity of a force is called stress and is defined as the force acting on an infinitesimal area. A normal stress σ is defined as

$$\sigma = \lim_{dA \to 0} \frac{dF}{dA} \tag{1.5.1}$$

where dF is a differential normal force acting on a differential area dA. It is often useful to calculate the average normal stress $\sigma = P/A$, where P is the resultant force on an area A. A shear stress τ caused by a shearing force V is defined likewise,

$$\tau = \lim_{dA \to 0} \frac{dV}{dA} \tag{1.5.2}$$

An average shear stress is obtained from V/A.

It is helpful to consider the general cases of stresses using rectangular elements in two and three dimensions, while ignoring the deformations caused by the stresses.

Plane Stress

There are relatively simple cases where all stress vectors lie in the same plane. This is represented by a two-dimensional element in Figure 1.5.1, where σ_x and/or σ_y may be either tensile (pulling on the element as shown) or compressive (pushing on the element; not shown). Normal stresses are easy to visualize and set up correctly.

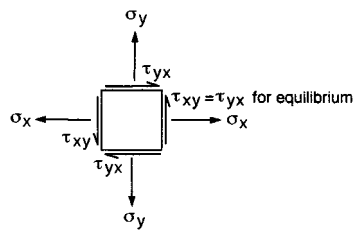


FIGURE 1.5.1 Generalized plane stress.

Shear stresses need to be discussed here in a little detail. The notation means that τ_{xy}, for example, is a shear stress acting in the y direction, on a face that is perpendicular to the x axis. It follows that τ_{yx} is acting in the x direction, on a face that is perpendicular to the y axis. The four shear stress vectors are

pointed as they are because of the requirement that the element be in equilibrium: the net forces and moments of forces on it must be zero. Thus, reversing the direction of all four τ's in Figure 1.5.1 is possible, but reversing less than four is not realistic.

Three-Dimensional State of Stress

The concept of plane stress can be generalized for a three-dimensional element as shown in Figure 1.5.2, working with the three primary faces of the cube and not showing stresses on the hidden faces, for clarity.

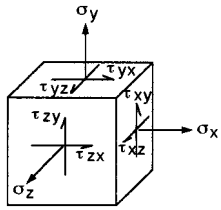


FIGURE 1.5.2 Three-dimensional general state of stress.

Again, the normal stresses are easy to set up, while the shear stresses may require considerable attention. The complex cases of stresses result from multiaxial loading, such as combined axial, bending, and torsional loading. Note that even in complex situations simplifications are possible. For example, if the right face in Figure 1.5.2 is a free surface, $\sigma_x = \tau_{xz} = \tau_{xy} = 0$. This leaves a plane stress state with σ_y, σ_z, and τ_{yz}, at most.

Stress Transformation

A free-body element with known stresses on it allows the calculation of stresses in directions other than the given *xyz* coordinates. This is useful when potentially critical welded or glued joints, or fibers of a composite, are along other axes. The stress transformations are simplest in the case of plane stress and can be done in several ways. In any case, at a given point in a material there is only one state of stress at a particular instant. At the same time, the components of the stresses depend on the orientation of the chosen coordinate system.

The stress transformation equations depend on the chosen coordinate system and the sign convention adopted. A common arrangement is shown in Figure 1.5.3, where (a) is the known set of stresses and (b) is the unknown set, denoted by primes.

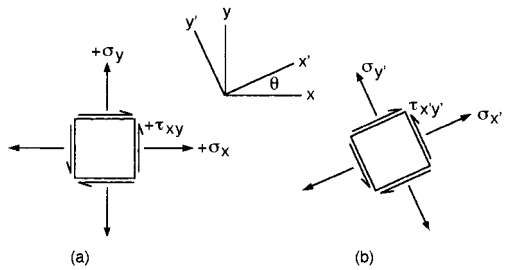


FIGURE 1.5.3 Elements for stress transformation.

In the present sign convention an outward normal stress is positive, and an upward shear stress on the right-hand face of the element is positive. The transformation equations are

$$\sigma_{x'} = \frac{\sigma_x + \sigma_y}{2} + \frac{\sigma_x - \sigma_y}{2}\cos 2\theta + \tau_{xy}\sin 2\theta$$

$$\sigma_{y'} = \frac{\sigma_x + \sigma_y}{2} - \frac{\sigma_x - \sigma_y}{2}\cos 2\theta - \tau_{xy}\sin 2\theta \qquad (1.5.3)$$

$$\tau_{x'y'} = -\frac{\sigma_x - \sigma_y}{2}\sin 2\theta + \tau_{xy}\cos 2\theta$$

If a result is negative, it means that the actual direction of the stress is opposite to the assumed direction.

Principal Stresses

It is often important to determine the maximum and minimum values of the stress at a point and the orientations of the planes of these stresses. For plane stress, the maximum and minimum normal stresses, called **principal stresses**, are obtained from

$$\sigma_{1,2} = \frac{\sigma_x + \sigma_y}{2} \pm \sqrt{\left(\frac{\sigma_x - \sigma_y}{2}\right)^2 + \tau_{xy}^2} \qquad (1.5.4)$$

There is no shear stress acting on the principal planes on which the principal stresses are acting. However, there are shear stresses on other planes. The maximum shear stress is calculated from

$$\tau_{\max} = \sqrt{\left(\frac{\sigma_x - \sigma_y}{2}\right)^2 + \tau_{xy}^2} \qquad (1.5.5)$$

This stress acts on planes oriented 45° from the planes of principal stress. There is a normal stress on these planes of τ_{max}, the average of σ_x and σ_y,

$$\sigma_{ave} = \frac{\sigma_x + \sigma_y}{2} \qquad (1.5.6)$$

Mohr's Circle for Plane Stress

The equations for plane stress transformation have a graphical solution, called Mohr's circle, which is convenient to use in engineering practice, including "back-of-the-envelope" calculations. Mohr's circle is plotted on a $\sigma - \tau$ coordinate system as in Figure 1.5.4, with the center C of the circle always on the σ axis at $\sigma_{ave} = (\sigma_x + \sigma_y)/2$ and its radius $R = \sqrt{[(\sigma_x - \sigma_y)/2]^2 + \tau_{xy}^2}$. The positive τ axis is downward for convenience, to make θ on the element and the corresponding 2θ on the circle agree in sense (both counterclockwise here).

The following aspects of Mohr's circle should be noted:

1. The center C of the circle is always on the σ axis, but it may move left and right in a dynamic loading situation. This should be considered in failure prevention.
2. The radius R of the circle is τ_{max}, and it may change, even pulsate, in dynamic loading. This is also relevant in failure prevention.
3. Working back and forth between the rectangular element and the circle should be done carefully and consistently. An angle θ on the element should be represented by 2θ in the corresponding circle. If τ is positive downward for the circle, the sense of rotation is identical in the element and the circle.

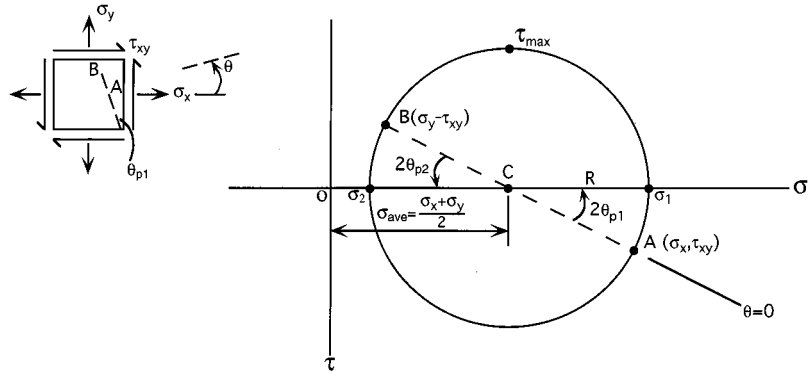


FIGURE 1.5.4 Mohr's circle.

4. The principal stresses σ_1 and σ_2 are on the σ axis ($\tau = 0$).
5. The planes on which σ_1 and σ_2 act are oriented at $2\theta_p$ from the planes of σ_x and σ_y (respectively) in the circle and at θ_p in the element.
6. The stresses on an arbitrary plane can be determined by their σ and τ coordinates from the circle. These coordinates give magnitudes and signs of the stresses. The physical meaning of $+\tau$ vs. $-\tau$ regarding material response is normally not as distinct as $+\sigma$ vs. $-\sigma$ (tension vs. compression).
7. To plot the circle, either use the calculated center *C* coordinate and the radius *R*, or directly plot the stress coordinates for two mutually perpendicular planes and draw the circle through the two points (*A* and *B* in Figure 1.5.4) which must be diametrically opposite on the circle.

Special Cases of Mohr's Circles for Plane Stress

See Figures 1.5.5 to 1.5.9

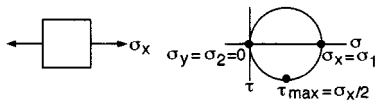


FIGURE 1.5.5 Uniaxial tension.

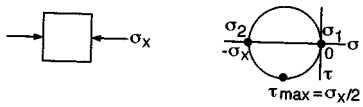


FIGURE 1.5.6 Uniaxial compression.

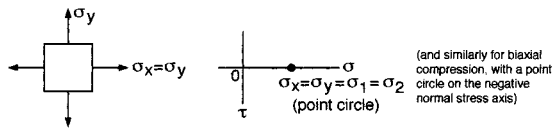


FIGURE 1.5.7 Biaxial tension: $\sigma_x = \sigma_y$ (and similarly for biaxial compression: $-\sigma_x = -\sigma_y$).

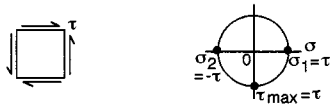


FIGURE 1.5.8 Pure shear.

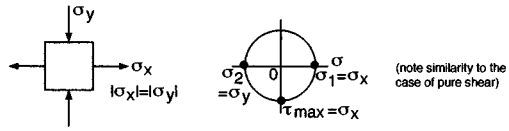


FIGURE 1.5.9 Biaxial tension-compression: $|\sigma_x| = |\sigma_y|$ (similar to the case of pure shear).

Absolute Maximum Shear Stress

In the case of a general three-dimensional state of stress, the transformations to arbitrary planes are complex and beyond the scope of this book. It is useful to note, however, that in general there are three principal stresses at any point in a material. (Plane stress is a special case with one of these stresses being zero.) If the three principal stresses are known, it is easy to determine the absolute maximum shear stress, which is valuable in assessing a material's performance in service. The idea is to view the element as three separate two-dimensional elements, each time from a different principal direction, and plot the Mohr's circles for them in the same diagram. This is illustrated schematically in Figure 1.5.10 for an element with three tensile principal stresses, of a maximum, a minimum, and an intermediate value.

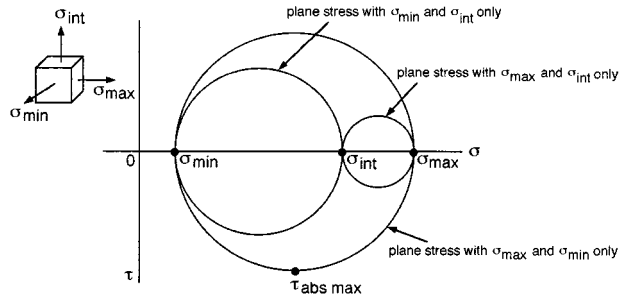


FIGURE 1.5.10 Principal stresses of three-dimensional element.

The Mohr's circles are interrelated since the three views of the element have common principal stresses associated with them. The absolute maximum shear stress is

$$\tau_{\text{abs max}} = \frac{\sigma_{\max} - \sigma_{\min}}{2} \tag{1.5.7}$$

Note that in calculating the absolute maximum shear stress for a state of plane stress, the actual third principal stress of $\sigma_3 = 0$ may be significant if that is the minimum stress, and should be used in Equation 1.5.7 instead of a larger intermediate stress. For example, assume $\sigma_x = 200$ ksi and $\sigma_y = 100$ ksi in a case of plane stress. Using these as $\sigma_{\max}$ and $\sigma_{\min}$, $\tau_{\max} = (200 - 100)/2 = 50$ ksi. However, the fact that $\sigma_z = 0$ is important here. Thus, correctly, $\tau_{\text{abs max}} = (200 - 0)/2 = 100$ ksi. There is an important lesson here: apparently negligible quantities cannot always be ignored in mechanics of materials.

Strain

Solid materials deform when forces are acting on them. Large deformations are possible in some materials. Extremely small deformations are difficult to measure, but they still may be significant in critical geometry change or gradual damage evolution. Deformations are normally nonuniform even in apparently uniform components of machines and structures. The severity of deformation is called strain, which is separately defined for volumetric change and angular distortion of a body.

Normal Strain

The elongation or shortening of a line segment of unit length is called normal strain or axial strain. To define this quantitatively, consider a uniform bar of length L_o, and call this original length the gage length. Assume the bar elongates by an amount e to a new length L_1 under the action of a force F (Figure 1.5.11) or by thermal expansion. The normal strain ε is defined, with the gage length approaching zero in the limit, as

$$\varepsilon = \frac{e}{L_o} \tag{1.5.8}$$

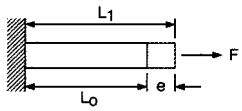


FIGURE 1.5.11 Model for calculating axial or normal strain.

The strain calculated this way is called engineering strain, which is useful and fairly accurate for small deformations. Elongation is considered positive.

Normal strain is a dimensionless quantity, but it is customary to label it in a way that indicates strain, such as in./in., or m/m, or %, or μ in./in., or $\mu\varepsilon$ (microstrain), depending on the system of units and the numerical representation.

True Strain

A difficulty of proper definition arises if the deformation e is not infinitesimal, because in a sense the gage length itself is increasing. The correct definition in such a case is based on the instantaneous length L and infinitesimal changes dL in that length. Thus, the total true strain in a member axially deforming from length L_o to a final length L_f by an amount e is

$$\varepsilon = \int_{L_o}^{L_f} \frac{dL}{L} = \ell n \frac{L_f}{L_o} = \ell n(1+e) \tag{1.5.9}$$

True strain is practically identical to engineering strain up to a few percent of engineering strain. The approximate final length L_f of an axially deformed, short line segment of original length L_o is sometimes expressed as

$$L_f \approx (1+\varepsilon)L_o \tag{1.5.10}$$

Shear Strain

Angular distortions are called shear strains. More precisely, shear strain γ is the change in angle of two originally perpendicular ($\theta = \pi/2$) line segments. For consistency, assume that a decreasing angle represents positive shear strain, and an increasing angle is from negative shear strain. The angle γ is measured in radians. A useful way to show shear strain is given in Figure 1.5.12.

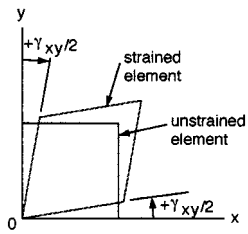


FIGURE 1.5.12 Shear strain in two dimensions.

Strain Transformation

The method of transforming strain at a point is similar to that for stress. In general, there are three components of normal strain, ε_x, ε_y, and ε_z, and three components of shear strain, γ_{xy}, γ_{xz}, and γ_{yz}. Transformations of plane strain components are the simplest.

For a consistent approach, assume that strain transformation is desired from an *xy* coordinate system to an *x′y′* set of axes, where the latter is rotated counterclockwise (+θ) from the *xy* system. The transformation equations for plane strain are

$$\varepsilon_{x'} = \frac{\varepsilon_x + \varepsilon_y}{2} + \frac{\varepsilon_x - \varepsilon_y}{2}\cos 2\theta + \frac{\gamma_{xy}}{2}\sin 2\theta$$

$$\varepsilon_{y'} = \frac{\varepsilon_x + \varepsilon_y}{2} - \frac{\varepsilon_x - \varepsilon_y}{2}\cos 2\theta - \frac{\gamma_{xy}}{2}\sin 2\theta \qquad (1.5.11)$$

$$\frac{\gamma_{x'y'}}{2} = -\frac{\varepsilon_x - \varepsilon_y}{2}\sin 2\theta + \frac{\gamma_{xy}}{2}\cos 2\theta$$

Note the similarity between the strain and stress transformation equations, as well as the differences.

Principal Strains

For isotropic materials only, principal strains (with no shear strain) occur along the principal axes for stress. In plane strain the principal strains ε_1 and ε_2 are expressed as

$$\varepsilon_{1,2} = \frac{\varepsilon_x + \varepsilon_y}{2} +/- \sqrt{\left(\frac{\varepsilon_x - \varepsilon_y}{2}\right)^2 - \left(\frac{\gamma_{xy}}{2}\right)^2} \qquad (1.5.12)$$

The angular position θ_p of the principal axes (measured positive counterclockwise) with respect to the given *xy* system is determined from

$$\tan 2\theta_p = \frac{\gamma_{xy}}{\varepsilon_x - \varepsilon_y} \qquad (1.5.13)$$

Like in the case of stress, the maximum in-plane shear strain is

$$\frac{\gamma_{x'y'\max}}{2} = \sqrt{\left(\frac{\varepsilon_x - \varepsilon_y}{2}\right)^2 + \left(\frac{\gamma_{xy}}{2}\right)^2} \qquad (1.5.14)$$

which occurs along axes at 45° from the principal axes, determined from

$$\tan 2\theta = -\frac{\varepsilon_x - \varepsilon_y}{\gamma_{xy}} \qquad (1.5.15)$$

The corresponding average normal strain is

$$\varepsilon_{ave} = \frac{\varepsilon_x + \varepsilon_y}{2} \qquad (1.5.16)$$

Mohr's Circle for Plane Strain

As in the case of stress, there is a graphical overview by Mohr's circle of the directional dependence of the normal and shear strain components at a point in a material. This circle has a center C at $\varepsilon_{ave} = (\varepsilon_x + \varepsilon_y)/2$ which is always on the ε axis, but is shifting left and right in a dynamic loading situation. The radius R of the circle is

$$R = \sqrt{\left(\frac{\varepsilon_x - \varepsilon_y}{2}\right)^2 + \left(\frac{\gamma_{xy}}{2}\right)^2} \tag{1.5.17}$$

Note the proper labeling (ε vs. $\gamma/2$) and preferred orientation of the strain axes as shown in Figure 1.5.13. This sets up a favorable uniformity of angular displacement between the element ($+\theta$ counterclockwise) and the circle ($+2\theta$ counterclockwise).

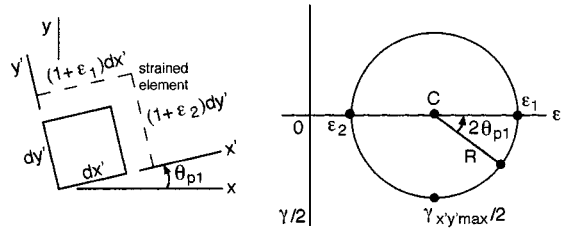


FIGURE 1.5.13 Mohr's circle for plane strain.

Mechanical Behaviors and Properties of Materials

The stress-strain response of a material depends on its chemical composition, microstructure, geometry, the magnitude and rate of change of stress or strain applied, and environmental factors. Numerous quantitative mechanical properties are used in engineering. Some of the basic properties and common variations of them are described here because they are essential in mechanics of materials analyses.

Stress-Strain Diagrams

There are several distinctive shapes of uniaxial tension or compression stress-strain plots, depending on the material, test conditions, and the quantities plotted. The chosen representative schematic diagram here is a true stress vs. true strain curve for a ductile, nonferrous metal tested in tension (Figure 1.5.14). The important mechanical properties listed in Table 1.5.1 are obtained from such a test or a similar one in pure shear (not all are shown in Figure 1.5.14).

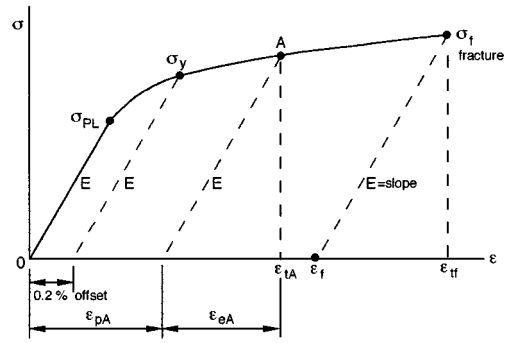


FIGURE 1.5.14 True stress vs. true strain for a ductile, nonferrous metal.

Table 1.5.1 Basic Mechanical Properties

Symbol	Definition	Remarks
E	Modulus of elasticity; Young's modulus; $E = \sigma/\varepsilon_e$	Hooke's law; T and ε_p effects small
G	Shear modulus of elasticity; $G = \dfrac{\tau}{\gamma_e} = E/2(1+\nu)$	T and ε_p effects small
ν	Poisson's ratio; $\nu = \dfrac{\varepsilon_{lateral}}{\varepsilon_{longit.}}$	T and ε_p effects small
σ_{PL}	Proportional limit; at onset of noticeable yielding (or at onset of nonlinear elastic behavior)	Flow property; inaccurate; T and ε_p effects large
σ_y	0.2% offset yield strength (but yielding can occur at $\sigma < \sigma_y$ if $\sigma_{PL} < \sigma_y$)	Flow property; accurate; T and ε_p effects large
σ_f	True fracture strength; $\sigma_f = \dfrac{P_f}{A_f}$	Fracture property; T and ε_p effects medium
ε_f	True fracture ductility; $\varepsilon_f = \ln\dfrac{A_O}{A_f} = \ln\dfrac{100}{100 - \%\text{RA}}$	Max. ε_p; fracture property; T and ε_p effects medium
% RA	Percent reduction of area; $\%\text{RA} = \dfrac{A_O - A_f}{A_O} \times 100$	Fracture property; T and ε_p effects medium
n	Strain hardening exponent; $\sigma = K\varepsilon_p^n$	Flow property; T and ε_p effects small to large
Toughness	Area under σ vs. ε_p curve	True toughness or intrinsic toughness; T and ε_p effects large
σ_u	Ultimate strength; $\dfrac{P_{\max}}{A_O}$	Fracture property; T and ε_p effects medium
M_r	Modulus of resilience; $M_r = \dfrac{\sigma_{PL}^2}{2E}$	Area under original elastic portion of $\sigma - \varepsilon$ curve

Notes: T is temperature; ε_p refers to prior plastic strain, especially cyclic plastic strain (fatigue) (these are qualitative indicators here; exceptions are possible)

$$\underset{\text{total}}{\varepsilon_t} = \underset{\text{elastic}}{\varepsilon_e} + \underset{\text{plastic}}{\varepsilon_p} = \frac{\sigma}{E} + \left(\frac{\sigma}{K}\right)^{1/n} = \frac{\sigma}{E} + \varepsilon_f\left(\frac{\sigma}{\sigma_f}\right)^{1/n}$$

Another useful mechanical property (not measured from the $\sigma - \varepsilon$ plot) is hardness. This is a flow property, with some qualitative correlations to the other properties.

It is important to appreciate that the mechanical properties of a material depend on its chemical composition and its history of thermal treatment and plastic deformations (cold work; cyclic plasticity). For example, consider the wide ranges of monotonic and cyclic stress-strain curves for 1045 steel (a given chemical composition) at room temperature, as functions of its hardness resulting from thermal treatment (Figure 1.5.15). See Section 1.6, "Fatigue," for more on cycle-dependent material behaviors.

Generalized Stress-Strain Expressions. Hooke's Law

An important special case of stress-strain responses is when the material acts entirely elastically ($\varepsilon_p = 0$, $\varepsilon_t = \varepsilon_e$). In this case, for uniaxial loading, the basic **Hooke's law** $\sigma = E\varepsilon$ can be used, and similarly for unidirectional shear, $\tau = G\gamma$. For multiaxial loading (Color Plate 9), the generalized Hooke's law is applicable,

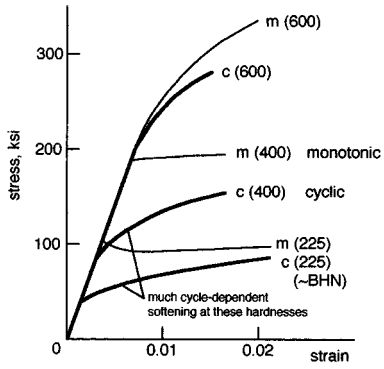


FIGURE 1.5.15 Influence of hardness and deformation history on the stress-strain response of SAE 1045 steel.

$$\varepsilon_x = \frac{1}{E}\left[\sigma_x - \nu\left(\sigma_y + \sigma_z\right)\right]$$

$$\varepsilon_y = \frac{1}{E}\left[\sigma_y - \nu\left(\sigma_x + \sigma_z\right)\right] \tag{1.5.18}$$

$$\varepsilon_z = \frac{1}{E}\left[\sigma_z - \nu\left(\sigma_x + \sigma_y\right)\right]$$

Other useful expressions for ideally elastic behavior are as follows. Relating the axial and shear moduli,

$$G = \frac{E}{2(1+\nu)} \tag{1.5.19}$$

The change in volume per unit volume is the volumetric strain or dilatation,

$$e = \frac{1-2\nu}{E}\left(\sigma_x + \sigma_y + \sigma_z\right) \tag{1.5.20}$$

The bulk modulus k is the ratio of a uniform stress (hydrostatic) to the dilatation,

$$k = \frac{\sigma}{e} = \frac{E}{3(1-2\nu)} \tag{1.5.21}$$

For most metals, $\nu \approx 1/3$ and $k \approx E$.

Uniaxial Elastic Deformations

The total elastic deformation δ of axially loaded bars, columns, and wires is calculated with the aid of basic expressions. Using $\sigma = E\varepsilon$ and $\sigma = P(x)/A(x)$,where $P(x)$ and $A(x)$ are, respectively, the internal force and cross-sectional area of a bar at a distance x from one end,

$$\delta = \int_0^L \frac{P(x)}{A(x)E}\,dx \tag{1.5.22}$$

where L is the total length considered.

In most cases, $A(x)$ is a constant; $P(x)$ may also be a constant, except where several different axial forces are applied, and occasionally for vertical bars and columns, where the member's own weight may cause $P(x)$ to vary significantly along the length. If $A(x)$, $P(x)$, and E are constants,

$$\delta = \frac{PL}{AE} \tag{1.5.23}$$

Thermally Induced Deformations

Thermal expansion or contraction is a linearly dependent, recoverable deformation like purely elastic deformations are. For a homogeneous and isotropic material, the thermally induced deformation from the original length L is calculated from

$$\delta_T = \alpha \Delta T L \tag{1.5.24}$$

where α is the linear coefficient of thermal expansion (strain per degree of temperature, a material property), and ΔT is the change in temperature.

The thermal strain can be prevented or reduced by constraining a member. In that case the stresses and strains can be calculated using the methods pertaining to statically indeterminate members.

Stresses in Beams

To calculate stresses in beams, one must first model the beam correctly in terms of its supports and loading (such as simply supported, with distributed loading), determine the appropriate unknown external reactions, and establish the corresponding shear and moment diagrams using a consistent sign convention. Both normal and shear stresses may have to be calculated, but typically the normal stresses are the most significant.

Flexure Formula

The normal stresses at a particular cross section in a beam are caused by the bending moment that acts at that cross section, and are distributed by magnitude and sign (both tension and compression) so that the beam is in equilibrium. The basic concept for calculating the stresses is that there is a neutral axis n-n of $\varepsilon = \sigma = 0$ in the beam, and that the longitudinal normal strain varies linearly with distance y from the neutral axis.

If the beam is behaving entirely elastically, the stress distribution is also linear, as in Figure 1.5.16. In this case, the stress at a distance y from the neutral axis is calculated from $M = \int \sigma(y) y \, dA$ and results in

$$\sigma(y) = \frac{My}{I} \tag{1.5.25}$$

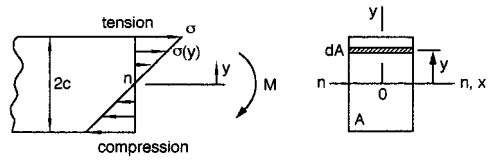


FIGURE 1.5.16 Internal normal stresses in a beam caused by bending.

where I = moment of inertia of the cross-sectional area about the neutral axis.

The maximum stress, with the appropriate sign, is

$$\sigma = \frac{Mc}{I} \tag{1.5.26}$$

There are several special cases of bending that require additional considerations and analysis as outlined below.

Inelastic Bending

A beam may plastically deform under an increasing moment, yielding first in its outer layers and ultimately throughout its depth. Such a beam is analyzed by assuming that the normal strains are still linearly varying from zero at the neutral axis to maximum values at the outer layers. Thus, the stress distributions depend on the stress-strain curve of the material. With the stress distribution established, the neutral axis can be determined from $\int\sigma(y)\,dA = 0$, and the resultant moment from $M = \int y\sigma(y)\,dA$.

A fully plastic beam of rectangular cross section and flat-top yielding supports 50% more bending moment than its maximum elastic moment.

Neutral Axis of Semisymmetric Area

If the cross-sectional area is semisymmetric, such as a T-shape, and the loading is in a centroidal plane of symmetry, the neutral axis for elastic deformations is at the centroid *C* of the area as shown in Figure 1.5.17, and Equation 1.5.25 can be used. Note that the magnitudes of the maximum tensile and compressive stresses are not the same in this case.

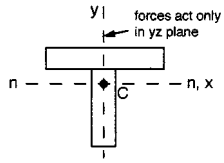


FIGURE 1.5.17 Neutral axis of a semisymmetric area.

Unsymmetric Bending

In the general case, the cross-sectional area has an arbitrary shape and the loading is arbitrarily applied. The problem of an arbitrary area is handled by choosing the centroidal *xy* coordinate system such that the axes are principal axes of inertia for the area. The principal axes can be determined by using inertia transformation equations or Mohr's circle of inertia. Having an axis of symmetry is a simple special case because the principal axes are the axis of symmetry and the axis perpendicular to it.

The **flexure formula** can be applied directly if the principal axes of inertia are known, and the bending moment is applied about one of these centroidal principal axes. A more complex case is if the moment is not about a principal axis as shown in Figure 1.5.18.

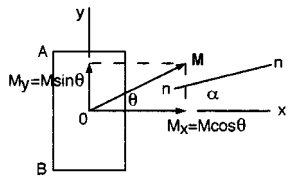


FIGURE 1.5.18 Schematic of arbitrary bending moment.

Different texts may present different formulas for calculating the bending stresses in such situations, depending on the choice of a coordinate system and the sign convention adopted. It is better not to rely on a cookbook formula, but to break down the problem into simple, easily visualized parts, and then reason out an algebraic superposition of the stress components. To illustrate this approach schematically, consider the stresses at points *A* and *B* in Figure 1.5.18. Instead of working directly with the applied moment **M**, resolve **M** into its components M_x and M_y. M_x causes a tensile stress $\sigma_{z_{A_1}}$ at *A* and a compressive stress $-\sigma_{z_{B_1}}$ at *B*. M_y causes tensile stresses at both *A* and *B*, $\sigma_{z_{A_2}}$ and $\sigma_{z_{B_2}}$. The magnitudes of these stress components are readily calculated from the flexure formula with the appropriate dimensions and inertias for each. The resultant stresses are

$$\sigma_A = \sigma_{z_{A_1}} + \sigma_{z_{A_2}}$$

$$\sigma_B = -\sigma_{z_{B_1}} + \sigma_{z_{B_2}}$$

The neutral axis at angle α in the general case is not coincident with the direction of **M** (Figure 1.5.18). In the present case α is defined by

$$\tan\alpha = \frac{I_x}{I_y}\tan\theta \tag{1.5.27}$$

Composite Beams

Nonhomogeneous beams are often designed to take advantage of the properties of two different materials. The approach for analyzing these is to imagine a transformation of the beam's cross section to an equivalent cross section of a different shape but of a single material, so that the flexure formula is usable. This is illustrated for a beam *A* with reinforcing plates *B*, as in Figure 1.5.19.

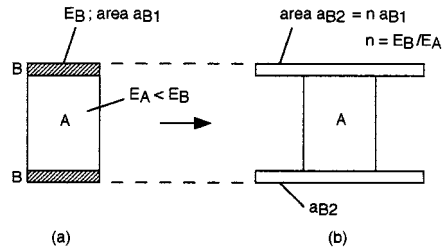


FIGURE 1.5.19 Equivalent area method for a symmetric composite beam.

The transformation factor *n* is obtained from

$$n = \frac{E_B}{E_A} \tag{1.5.28}$$

Note that in a composite beam the strains vary linearly with distance from the neutral axis, but the stresses do not, because of the different elastic moduli of the components. The actual stress σ in the transformed area is determined by first calculating the "pretend" stress σ' for the uniform transformed area and then multiplying it by *n*,

$$\sigma = n\sigma' \tag{1.5.29}$$

Nonsymmetric composite beams (such as having only one reinforcing plate *B* in Figure 1.5.19) are analyzed similarly, but first require the location of the neutral axis.

Reinforced concrete beams are important special cases of composite beams. The stress analysis of these is influenced by the fact that concrete is much weaker in tension than in compression. Empirical approaches are particularly useful in this area.

Curved Beams

The stress analysis of curved beams requires some additional considerations. For example, the flexure formula is about 7% in error (the calculated stresses are too low) when the beam's radius of curvature is five times its depth (hooks, chain links). The curved-beam formula provides realistic values in such cases.

Shear Stresses in Beams

Transverse loads on beams are common, and they cause transverse and complementary longitudinal shear stresses in the beams. Schematically, the transverse shear stresses are distributed on a rectangular cross section as shown in Figure 1.5.20. The shear stress is zero at free surfaces by definition.

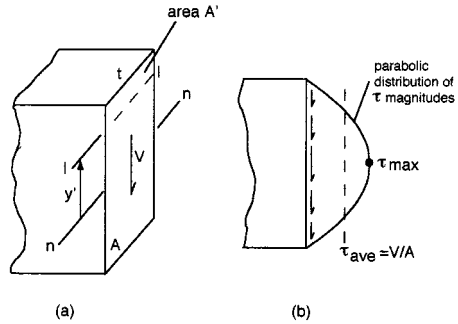


FIGURE 1.5.20 Transverse shear stress distribution.

The internal shear stress is calculated according to Figure 1.5.20 from

$$\tau = \frac{VQ}{It} \tag{1.5.30}$$

where τ = shear stress value at any point on the line $\ell - \ell$ at a distance y' from the neutral axis
V = total shear force on cross-sectional area A
Q = $\bar{y}'A'$; A' = area above line $\ell - \ell$; $\bar{y}'$ = distance from neutral axis to centroid of A'
I = moment of inertia of entire area A about neutral axis
t = width of cross section where τ is to be determined

This shear formula gives $\tau_{max} = 1.5\ V/A$ if t is constant for the whole section (rectangle).

Note that the magnitude of the shear stress distribution changes sharply where there is an abrupt change in width t, such as in an I-beam, Figure 1.5.21.

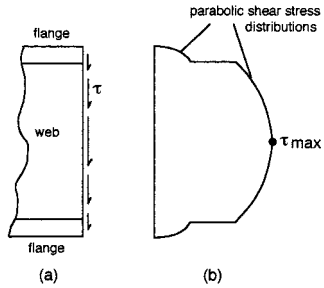


FIGURE 1.5.21 Shear stress distribution for I-beam.

Shear Flow

In the analysis of built-up members, such as welded, bolted, nailed, or glued box beams and channels, a useful quantity is the shear flow q measured in force per unit length along the beam,

$$q = \frac{VQ}{I} \tag{1.5.31}$$

where all quantities are defined as for Equation 1.5.30. Care must be taken to use the appropriate value for Q. For example, consider a channel section of three flat pieces glued together as in Figure 1.5.22. There are two critical joint regions B here, and the area A' is between them. The shear flow is carried by the two joints together, so the actual force per unit length on one joint is $q/2$ here.

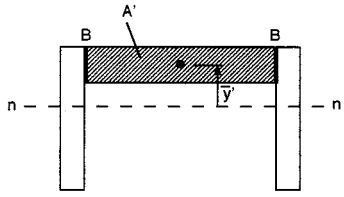


FIGURE 1.5.22 Critical joint regions of a built-up beam.

Shear Flow in Thin-Walled Beams

The shear-flow distribution over the cross section of a thin-walled member is governed by equilibrium requirements. Schematic examples of this are given in Figure 1.5.23. Note the special case of unsymmetrical loading in part (c), which causes a bending and a twisting of the beam. The twisting is prevented if the vertical force V is applied at the shear center C, defined by the quantity e,

$$e = \frac{dH}{V} \tag{1.5.32}$$

where d is the centroidal distance between the two horizontal flanges and H is the shear force in the flanges (q_{ave} times width of flange).

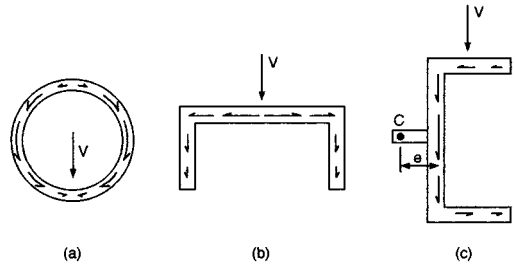


FIGURE 1.5.23 Shear flow distributions.

Deflections of Beams

Small deflections of beams can be determined relatively easily. The first step is to assess a beam's loading and support conditions and sketch an exaggerated elastic deflection curve as in Figure 1.5.24.

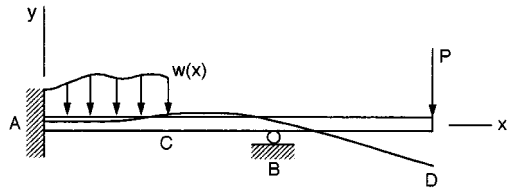


FIGURE 1.5.24 Exaggerated elastic curve of a beam in bending.

The boundary conditions at the supports are useful in the solution for the whole beam. Here at the fixed end *A* there is no vertical displacement and no rotation, while at the roller support *B* there is no vertical displacement but rotation of the continuous beam occurs. The boundary and continuity conditions can be determined by inspection in simple cases.

Moment vs. Curvature

For a homogeneous and elastic beam,

$$\frac{1}{\rho} = \frac{M}{EI} \tag{1.5.33}$$

where ρ = radius of curvature at a specific point on the elastic curve; 1/ρ is the curvature. The product *EI* is called the flexural rigidity; it is often a constant along the whole beam.

Integration Method for Slope and Displacement

For small displacements, $1/\rho = d^2y/dx^2$. In the general case, a distributed external loading $w(x)$ should be included in the modeling of the problem. A set of expressions is available to solve for the deflections in rectangular coordinates:

$$-w(x) = \frac{dV}{dx} = EI\frac{d^4y}{dx^4}$$

$$V(x) = \frac{dM}{dx} = EI\frac{d^3y}{dx^3} \tag{1.5.34}$$

$$M(x) = EI\frac{d^2y}{dx^2}$$

The deflection *y* of the elastic curve is obtained by successive integrations, using appropriate constants of integration to satisfy the boundary and continuity conditions. In general, several functions must be written for the moment $M(x)$, one for each distinct region of the beam, between loading discontinuities. For example, these regions in Figure 1.5.24 are *AC*, *CB*, and *BD*. Considerable care is required to set up a solution with a consistent sign convention and selection of coordinates for simple and efficient forms of $M(x)$.

In practice, even relatively complex problems of beam deflections are solved using the principle of superposition and handbook values of slopes and deflections for subsets of basic loadings and supports. The literature contains a large variety of such subsets. A sampling of these is given in Table 1.5.2.

Deflection Caused by Shear

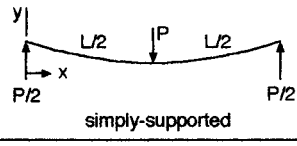
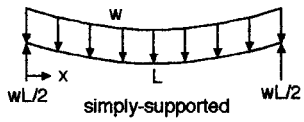
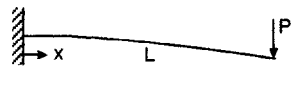
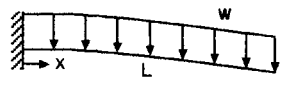
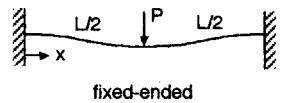
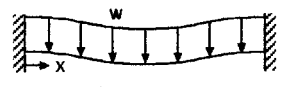
The transverse shear acting on a beam causes a displacement that tends to be significant compared to bending deflections only in very short beams. The shear deflection over a length *L* is approximated by

$$y = \frac{\tau_{ave} L}{G} \tag{1.5.35}$$

Torsion

The simplest torsion members have circular cross sections. The main assumptions in their analysis are that cross-sectional circles remain plane circles during twisting of a shaft and that radial lines on any cross section remain straight and rotate through the same angle. The length and diameter of the shaft are unchanged in small angular displacements. It is useful in the analysis that shear strain γ varies linearly

TABLE 1.5.2

Beam	Slope: ldy/dxl	Max. deflection
 simply-supported	$PL^2 / 16EI$ at x=0, L	$PL^3 / 48EI$ at x=L/2
 simply-supported	$wL^3 / 24EI$ at x=0, L	$5wL^4 / 384EI$ at x=L/2
 cantilever	$PL^2 / 2EI$ at x=L	$PL^3 / 3EI$ at x=L
 cantilever	$wL^3 / 6EI$ at x=L	$wL^4 / 8EI$ at x=L
 fixed-ended	0 at x=0, L/2, L	$PL^3 / 192EI$ at x=L/2
 fixed-ended	0 at x=0, L/2, L	$wL^4 / 384EI$ at x=L/2

along any radial line, from zero at the centerline of a solid or tubular shaft to a maximum at the outer surface,

$$\gamma = \frac{\rho}{r}\gamma_{max} \tag{1.5.36}$$

where ρ = radial distance to any element in the shaft
r = radius of the shaft

Using $\tau = G\gamma$ for an elastically deforming material (Figure 1.5.25),

$$\tau = \frac{\rho}{r}\tau_{max} \tag{1.5.37}$$

The torsion formula relating shear stress to the applied torque T is from $T = 2\pi\int\tau\rho^2\, d\rho$,

$$\tau_{max} = \frac{Tr}{J} \quad \text{or} \quad \tau = \frac{T\rho}{J} \tag{1.5.38}$$

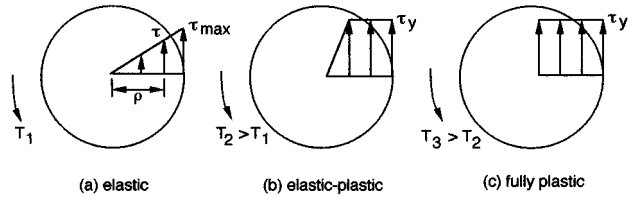


FIGURE 1.5.25 Shear stress distributions in a shaft.

where J = the polar moment of inertia of the cross-sectional area; for a solid circle, $J = \pi r^4/2$; for a tube, $J = (\pi/2)(r_0^4 - r_i^4)$.

Power Transmission

The power P transmitted by a shaft under torque T and rotating at angular velocity ω is

$$P = T\omega \tag{1.5.39}$$

where $\omega = 2\pi f$; f = frequency of rotation or number of revolutions per second.

Angle of Twist

For a homogeneous shaft of constant area and G over a length L, under a torque T, the angular displacement of one end relative to the other is

$$\phi = \frac{TL}{JG} \tag{1.5.40}$$

For a shaft consisting of segments with various material and/or geometric properties, under several different torques in each, the net angular displacement is calculated from the vector sum of the individual twists,

$$\phi = \sum \frac{T_i L_i}{J_i G_i} \tag{1.5.41}$$

The right-hand rule is used for a sign convention for torques and angles: both T and ϕ are positive, with the thumb pointing outward from a shaft and the fingers curling in the direction of torque and/or rotation, as in Figure 1.5.26. Note that regardless of the number of torques applied to a shaft at various places along its length, there is only one torque at a given cross section, and this torque is a constant in that segment of the shaft (until another external torque is encountered, requiring a different free-body diagram).

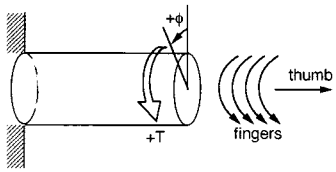


FIGURE 1.5.26 Right-hand rule for positive torque and angle.

Inelastic Torsion

A shaft may plastically deform under an increasing torque, yielding first in its outer layers and ultimately throughout the cross section. Such a shaft is analyzed by assuming that the shear strains are still linearly

varying from zero at the centerline to a maximum at the outer layers. Thus, the shear stress distribution depends on the shear stress-strain curve of the material. For example, an elastic, elastic-plastic, and fully plastic solid circular shaft is modeled in Figure 1.5.25, assuming flat-top yielding at τ_y. The torque T in any case is obtained by integrating the shear stresses over the whole area,

$$T = 2\pi \int_A \tau \rho^2 \, d\rho \tag{1.5.42}$$

The fully plastic torque in this case is 33% greater than the maximum elastic torque.

Noncircular Shafts

The analysis of solid noncircular members, such as rectangles and triangles, is beyond the scope of this book. The reason for the difficulty is that plane sections do not remain plane, but warp. It can be noted here, however, that a circular shaft utilizes material the most efficiently since it has a smaller maximum shear stress and a smaller angle of twist than a noncircular shaft of the same weight per unit length under the same torque.

Noncircular tubes with thin walls can be analyzed using the concept of shear flow that must be continuous and constant on the closed path of the cross-sectional area. The shear stress under a torque T is essentially constant over a uniformly thin wall (from inside to outside), and is given by

$$\tau = \frac{T}{2tA_m} \tag{1.5.43}$$

where t = thickness of the tube
A_m = mean area within the centerline of the wall thickness

The angle of twist for an elastically deforming thin-walled tube of length L and constant thickness t is

$$\phi = \frac{TL}{4A_m^2 Gt} \oint ds \tag{1.5.44}$$

where the line integral represents the total length of the wall's centerline boundary in the cross section (for a circular tube, this becomes $\approx 2\pi r$). For a tube with variable thickness t, the integrand becomes $\oint ds/t$.

Statically Indeterminate Members

Members that have more supports or constraints than the minimum required for static equilibrium are called statically indeterminate. They can be analyzed if a sufficient number of additional relationships are available. These are fundamentally similar to one another in terms of compatibility for displacements, and are described separately for special cases.

Statically Indeterminate Axially Loaded Members

Several subsets of these are common; three are shown schematically in Figure 1.5.27.

1. From a free-body diagram of part (a), assuming upward forces F_A and F_B at ends A and B, respectively, the force equilibrium equation is

$$F_A + F_B - P = 0$$

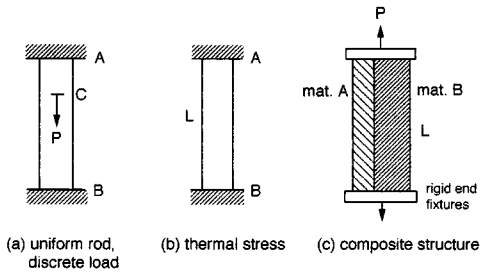


FIGURE 1.5.27 Statically indeterminate axially loaded members.

The displacement compatibility condition is that both ends are fixed, so

$$\delta_{A/B} = 0$$

Then

$$\frac{F_A L_{AC}}{AE} - \frac{F_B L_{BC}}{AE} = 0, \quad F_A = P\frac{L_{BC}}{L_{AB}}, \quad F_B = P\frac{L_{AC}}{L_{AB}}$$

Alternatively, first assume that $F_B = 0$, and calculate the total downward displacement (tensile) of the free end *B*. Then calculate the required force F_B to compressively deform the rod upward so that after the superposition there is no net displacement of end *B*. The results are the same as above for elastically deforming members.

2. Constrained thermal expansion or contraction of part (b) is handled as above, using the expression for thermally induced deformation,

$$\delta_T = \alpha\, \Delta T L \tag{1.5.45}$$

where α = linear coefficient of thermal expansion
ΔT = change in temperature

3. The force equilibrium equation of part (c) is

$$P - F_A - F_B = 0$$

Here the two different component materials are deforming together by the same amount, so

$$\delta_A = \delta_B$$

$$\frac{F_A L}{A_A E_A} = \frac{F_B L}{A_B E_B}$$

providing two equations with two unknowns, F_A and F_B. Note that rigid supports are not necessarily realistic to assume in all cases.

Statically Indeterminate Beams

As for axially loaded members, the redundant reactions of beams are determined from the given conditions of geometry (the displacement compatibility conditions). There are various approaches for solving problems of statically indeterminate beams, using the methods of integration, moment-areas, or

superposition. Handbook formulas for the slopes and deflections of beams are especially useful, noting that the boundary conditions must be well defined in any case. The method of superposition is illustrated in Figure 1.5.28.

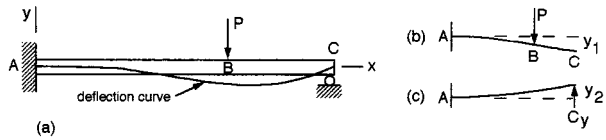


FIGURE 1.5.28 A statically indeterminate beam.

Choosing the reaction at C as the redundant support reaction (otherwise, the moment at A could be taken as redundant), and first removing the unknown reaction C_y, the statically determinate and stable primary beam is obtained in Figure 1.5.28b. Here the slope and deflection at A are both zero. The slopes at B and C are the same because segment BC is straight. Next the external load P is removed, and a cantilever beam fixed at A and with load C_y is considered in Figure 1.5.28c. From the original boundary conditions at C, $-y_1 + y_2 = 0$, and the problem can be solved easily using any appropriate method.

Statically Indeterminate Torsion Members

Torsion members with redundant supports are analyzed essentially the same way as other kinds of statically indeterminate members. The unknown torques, for example, are determined by setting up a solution to satisfy the requirements of equilibrium ($\sum T = 0$), angular displacement compatibility, and torque-displacement (angle = TL/JG) relationships. Again, the boundary conditions must be reasonably well defined.

Buckling

The elastic buckling of relatively long and slender members under axial compressive loading could result in sudden and catastrophic large displacements. The critical buckling load is the smallest for a given ideal column when it is pin-supported at both ends; the critical load is larger than this for other kinds of supports. An ideal column is made of homogeneous material, is perfectly straight prior to loading, and is loaded only axially through the centroid of its cross-sectional area.

Critical Load. Euler's Equation

The buckling equation (Euler's equation) for a pin-supported column gives the critical or maximum axial load P_{cr} as

$$P_{cr} = \frac{\pi^2 EI}{L^2} \tag{1.5.46}$$

where E = modulus of elasticity
I = smallest moment of inertia of the cross-sectional area
L = unsupported length of the pinned column

A useful form of this equation gives the critical average stress prior to any yielding, for arbitrary end conditions,

$$\sigma_{cr} = \frac{\pi^2 E}{(kL/r)^2} \tag{1.5.47}$$

where $r = \sqrt{I/A}$ = radius of gyration of cross-sectional area A
L/r = slenderness ratio
k = effective-length factor; constant, dependent on the end constraints
kL/r = effective-slenderness ratio

The slenderness ratio indicates, for a given material, the tendency for elastic buckling or failure by yielding (where the Euler formula is not applicable). For example, buckling is expected in mild steel if L/r is approximately 90 or larger, and in an aluminum alloy if $L/r > 60$. Yielding would occur first at smaller values of L/r. Ratios of 200 or higher indicate very slender members that cannot support large compressive loads.

Several common end conditions of slender columns are shown schematically in Figure 1.5.29.

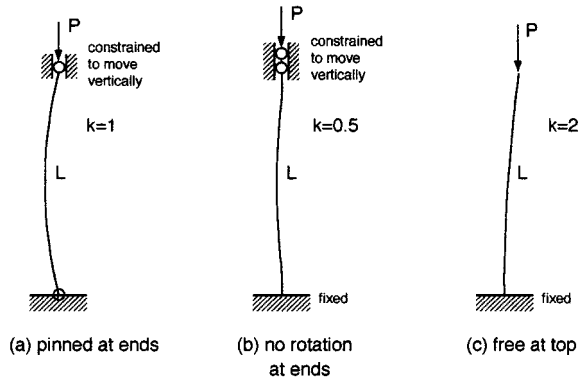


FIGURE 1.5.29 Common end conditions of slender columns.

Secant Formula

Real columns are not perfectly straight and homogeneous and are likely to be loaded eccentrically. Such columns first bend and deflect laterally, rather than buckle suddenly. The maximum elastic compressive stress in this case is caused by the axial and bending loads and is calculated for small deflections from the secant formula,

$$\sigma_{\max} = \frac{P}{A}\left[1 + \frac{ec}{r^2}\sec\left(\frac{L}{2r}\sqrt{\frac{P}{EA}}\right)\right] \tag{1.5.48}$$

where e is the eccentricity of the load P (distance from the neutral axis of area A) and c is measured from the neutral axis to the outer layer of the column where $\sigma_{\max}$ occurs.

The load and stress are nonlinearly related; if there are several loads on a column, the loads should be properly combined first before using the secant formula, rather than linearly superposing several individually determined stresses. Similarly, factors of safety should be applied to the resultant load.

Inelastic Buckling

For columns that may yield before buckling elastically, the generalized Euler equation, also called the Engesser equation, is appropriate. This involves substituting the tangent modulus E_T (tangent to the stress-strain curve) for the elastic modulus E in the Euler equation,

$$\sigma_{cr} = \frac{\pi^2 E_T}{(kL/r)^2} \tag{1.5.49}$$

Note that E_T must be valid for the stress σ_{cr}, but E_T is dependent on stress when the deformations are not entirely elastic. Thermal or plastic-strain events may even alter the stress-strain curve of the material, thereby further changing E_T. Thus, Equation 1.5.49 should be used with caution in a trial-and-error procedure.

Impact Loading

A mass impacting another object causes deformations that depend on the relative velocity between them. The simplest model for such an event is a mass falling on a spring. The maximum dynamic deformation d of a linearly responding spring is related to the static deformation d_{st} (the deformation caused by a weight W applied slowly) by a factor that depends on h, the height of free fall from a static position.

$$d = d_{st}\left(1 + \sqrt{1 + \frac{2h}{d_{st}}}\right) \tag{1.5.50}$$

The dynamic and static stresses are related in a similar way,

$$\sigma = \sigma_{st}\left(1 + \sqrt{1 + \frac{2h}{d_{st}}}\right) \tag{1.5.51}$$

The quantity in parentheses is called the impact factor, which shows the magnification of deflection or stress in impacts involving free fall. Note that the real impact factor is somewhat smaller than what is indicated here, because some energy is always dissipated by friction during the fall and deceleration of the body. This includes internal friction during plastic flow at the points of contact between the bodies. Other small errors may result from neglecting the mass and possible inelasticity of the spring.

A special value of the impact factor is worth remembering. When the load is applied suddenly without a prior free fall, $h = 0$, and

$$d = 2d_{st} \quad \text{and} \quad \sigma = 2\sigma_{st}$$

This means that the minimum impact factor is about two, and it is likely to be larger than two, causing perhaps a “bottoming out” of the spring, or permanent damage somewhere in the structure or the payload supported by the spring.

Combined Stresses

Combinations of different kinds of loads on a member are common. The resultant states of stress at various points of interest can be determined by superposition if the material does not yield. The three-dimensional visualization and correct modeling of such a problem is typically the most difficult part of the solution, followed by routine calculations of the stress components and resultants. No new methods of analysis are needed here.

The approach is to sketch an infinitesimal cube at each critical point in the member and determine the individual stresses (magnitudes and signs) acting on that element, generally caused by axial, shear, bending, torsion, and internal pressure loading. This is illustrated in Figure 1.5.30, for a case of medium complexity.

Consider a solid circular rod of radius R, fixed at $z = 0$ (in the xy plane), and loaded by two forces at point B of a rigid arm. Set up the stress analysis at point A ($-R$, 0, 0), assuming there is no stress concentration at the wall fixture of the rod (Figure 1.5.30a).

First the equivalent loading at the origin 0 is determined (Figure 1.5.30b). This can be done most accurately in vector form. The individual stresses at point A are set up in the subdiagram (c). Check that

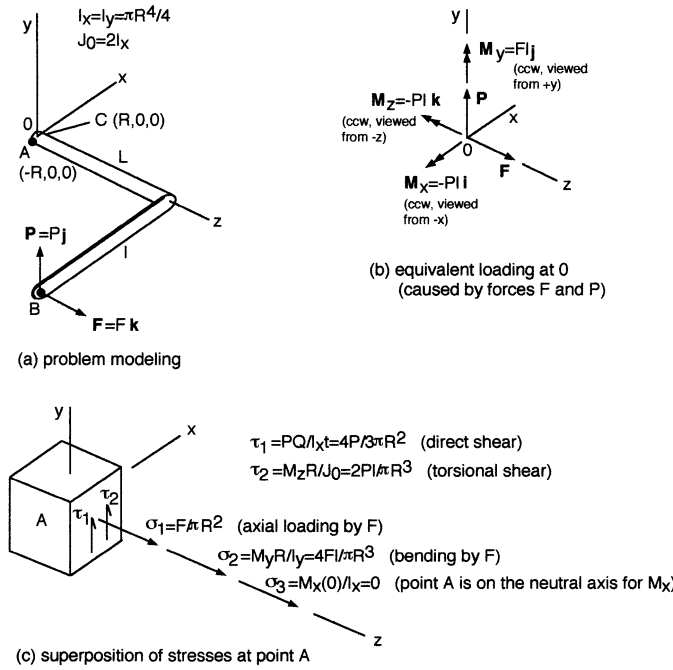


FIGURE 1.5.30 Illustration of stress analysis for combined axial, shear, bending, and torsion loading.

each stress even in symbolic form has the proper units of force per area. The net normal force in this case is $\sigma_1 + \sigma_2$, and the net shear stress is $\tau_1 + \tau_2$.

The state of stress is different at other points in the member. Note that some of the stresses at a point could have different signs, reducing the resultant stress at that location. Such is the case at a point *C* diametrically opposite to point *A* in the present example (*R*, 0, 0), where the axial load **F** and $\mathbf{M}_y$ generate normal stresses of opposite signs. This shows the importance of proper modeling and setting up a problem of combined loads before doing the numerical solution.

Pressure Vessels

Maan H. Jawad and Bela I. Sandor

Pressure vessels are made in different shapes and sizes (Figure 1.5.31 and Color Plate 10) and are used in diverse applications. The applications range from air receivers in gasoline stations to nuclear reactors in submarines to heat exchangers in refineries. The required thicknesses for some commonly encountered pressure vessel components depend on the geometry as follows.

Cylindrical Shells

The force per unit length in the hoop (tangential) direction, N_t, required to contain a given pressure p in a cylindrical shell is obtained by taking a free-body diagram (Figure 1.5.32a) of the cross section. Assuming the thickness t to be much smaller than the radius R and summing forces in the vertical direction gives

$$2N_tL = 2RLp$$

or

$$N_t = pR \tag{1.5.52}$$

(a)

(b)

FIGURE 1.5.31 Various pressure vessels. (Photos courtesy Nooter Corp., St. Louis, MO.)

The corresponding hoop stress is $\sigma_t = pR/t$.

The longitudinal force per unit length, N_x, in the cylinder due to pressure is obtained by summing forces in the axial direction (Figure 1.5.32b),

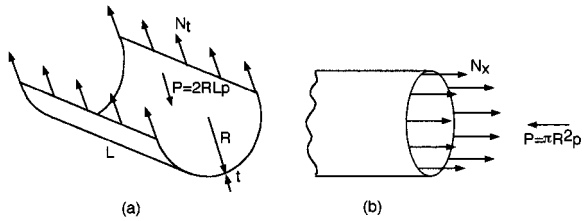


FIGURE 1.5.32 Analysis of cylindrical pressure vessels.

$$2\pi RN_x = \pi R^2 p$$

or

$$N_x = pR/2 \tag{1.5.53}$$

The corresponding axial stress is $\sigma_x = pR/2t$. It is seen that the magnitude of N_t (and σ_t) is twice that of N_x (and σ_x). If S is the allowable stress and t is the required minimum thickness,

$$t = pR/S \tag{1.5.54}$$

Spherical Shells

A free-body diagram of the spherical cross section is shown in Figure 1.5.33. Summation of forces gives

$$t = pR/2S \tag{1.5.55}$$

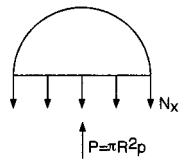


FIGURE 1.5.33 Analysis of spherical pressure vessels.

Example 10

Determine the required thickness of the shell and heads of the air receiver shown in Figure 1.5.34 if p = 100 psi and S = 15,000 psi.

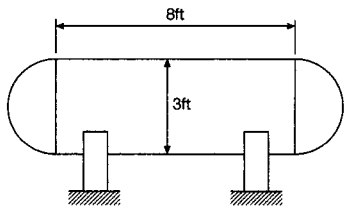


FIGURE 1.5.34 Sketch of a pressure vessel.

Solution. From Equation 1.5.54, the required thickness for the cylindrical shell is

$$t = 100 \times 18/15{,}000 = 0.12 \text{ in.}$$

The required head thickness from Equation 1.5.55 is

$$t = 100 \times 18/2 \times 15{,}000 = 0.06 \text{ in.}$$

Conical Shells

The governing equations for the longitudinal and circumferential forces in a conical shell (Figure 1.5.35a) due to internal pressure are similar to Equations 1.5.52 and 1.5.53 for cylindrical shells, with the radius taken normal to the surface. Thus,

$$N_t = pr/\cos\alpha \tag{1.5.56}$$

$$N_x = pr/2\cos\alpha \tag{1.5.57}$$

where α is half the apex angle of the cone.

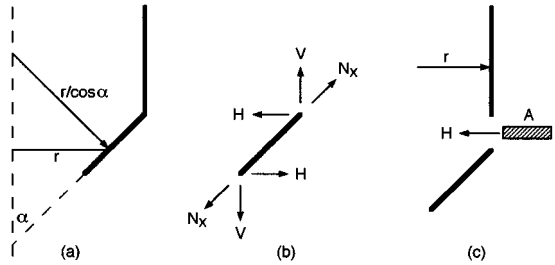


FIGURE 1.5.35 Analysis of conical shells.

The junction between a conical and cylindrical shell, Figure 1.5.35b, is subjected to an additional force, *H*, in the horizontal direction due to internal pressure. The magnitude of this additional force per unit length can be obtained by taking a free-body diagram as shown in Figure 1.5.35b,

$$H = N_x \sin\alpha \tag{1.5.58}$$

A ring is usually provided at the cone-to-cylinder junction to carry the horizontal force H. The required area *A* of the ring is obtained from Figure 1.5.35c as

$$H2r = 2AS$$

or

$$A = Hr/S = \left(N_x \sin\alpha\right)r/S = \left(pr^2 \sin\alpha\right)/(2S\cos\alpha) \tag{1.5.59}$$

The stress in the ring is compressive at the large end of the cone and tensile at the small end of the cone due to internal pressure. This stress may reverse in direction due to other loading conditions such as weight of contents and end loads on the cone due to wind and earthquake loads.

Example 11

Determine the required thickness of the two cylindrical shells and cone shown in Figure 1.5.36a due to an internal pressure of 200 psi. Calculate the area of the rings required at the junctions. Assume the allowable stress to be 20 ksi in tension and 10 ksi in compression.

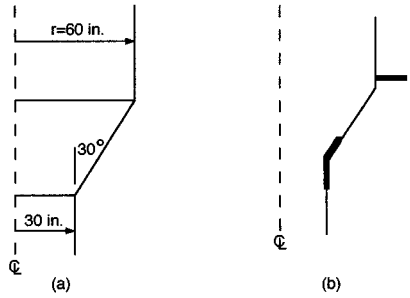


FIGURE 1.5.36 Cylindrical shells with cone connection.

Solution. From Equation 1.5.54, the thickness of the large cylinder is

$$t = 200 \times 60/20{,}000 = 0.60 \text{ in.}$$

The thickness of the small cylinder is

$$t = 200 \times 30/20{,}000 = 0.30 \text{ in.}$$

The thickness of the cone is obtained from Equation 1.5.56 as

$$t = 200 \times 60/(20{,}000 \times \cos 30°) = 0.69 \text{ in.}$$

The required area of the ring at the large end of the cone is obtained from Equation 1.5.59 using the allowable compressive stress of 10 ksi.

$$\text{A} = 200 \times 60^2 \times \sin 30°/(2 \times 10{,}000 \times \cos 30°) = 20.78 \text{ in.}^2$$

The required area of the ring at the small end of the cone is obtained from Equation 1.5.59 using the allowable tensile stress of 20 ksi.

$$\text{A} = 200 \times 30^2 \times \sin 30°/(2 \times 20{,}000 \times \cos 30°) = 2.60 \text{ in.}^2$$

The rings at the junction are incorporated in a number of ways such as those shown in Figure 1.5.36b.

Nozzle Reinforcement

Reinforcements around openings in pressure vessels are needed to minimize the local stress in the area of the opening. The calculation for the needed reinforcement around an opening is based on the concept that pressure in a given area of a vessel is contained by the material in the vessel wall surrounding the pressure. Thus in Figure 1.5.37, if we take an infinitesimal length *dL* along the cylinder, the force caused by the pressure within this length is given by the quantity *pR dL*. The force in the corresponding vessel wall is given by *St dL*. Equating these two quantities results in the expression $t = pR/S$ which is given earlier as Equation 1.5.54. Similarly for the nozzle in Figure 1.5.37, $T = pr/S$. The intersection of the nozzle with the cylinder results in an opening where the pressure in area *ABCD* is not contained by any material. Accordingly, an additional area must be supplied in the vicinity of the opening to prevent overstress of the vessel. The required area *A* is determined from Figure 1.5.37 as

$$A = pRr/S$$

Substituting Equation 1.5.54 into this expression gives

$$A = tr \tag{1.5.60}$$

This equation indicates that the needed additional area is equal to the removed area of the vessel wall.

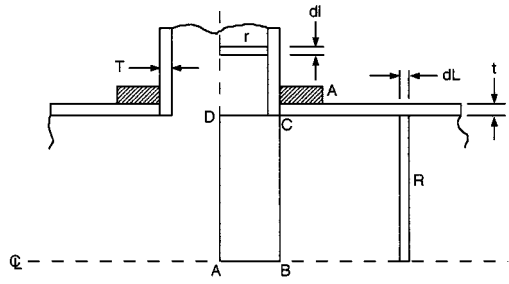


FIGURE 1.5.37 Nozzle reinforcement.

Creep-Fatigue of Boilers and Pressure Vessels

See Figure 1.6.27 in Section 1.6, "Fatigue."

Composite Materials for Pressure Vessels

Ian K. Glasgow

Some pressure vessels can be made of fibrous composite materials with high strength-to-weight ratios. The advantages of using such a material are remarkable in the case of a tubular vessel, where the hoop stress is twice the longitudinal stress, if the fiber quantities and orientations are optimally designed to resist the applied load caused by internal pressure. Simplistically (since a basic element of a composite is strong along the fibers and weak perpendicular to the fibers), this requires twice as many fibers oriented circumferentially as axially. In practice, fibers are commonly laid at ± (a winding angle) at which the hoop and axial stress components are equal, to efficiently create an optimized configuration.

Example 12

Determine the minimum weight of the tube portion of a thin-walled cylindrical pressure vessel of r = 8 in. (20 mm), ℓ = 10 ft (3.05 m), p = 8 ksi (55 MPa); t = ? Assume using a typical graphite/epoxy composite of 60% fibers by volume with allowable tensile stress σ_y = 300 ksi (207 MPa) at 0.058 lb/in.^{3} (1600 kg/m^3). For comparison, consider a steel of σ_y = 200 ksi (138 MPa) at 0.285 lb/in.^{3} (7890 kg/m^3).

Solution.

Composite: $\sigma_y = pr/t$, t = 0.213 in. (5.41 mm) for circumferential plies
$\sigma_y = pr/2t$, t = 0.107 in. (2.72 mm) for axial plies

Total minimum wall thickness: 0.32 in. (8.13 mm)
Total material in tube: 112 lb (50.8 kg)
Steel: $\sigma_y = pr/t$, t = 0.32 in. (8.13 mm)
Total material in tube: 550 lb (249 kg) = 4.9 (composite material)
Note that there are additional considerations in practice, such as cost and potential problems in making adequate connections to the tube.

Experimental Stress Analysis and Mechanical Testing

Michael L. Brown and Bela I. Sandor

Experimental stress analysis is based mostly on the measurement of strains, which may be transformed into stresses. A variety of techniques are available to measure strains. A few of these are described here.

Properties of Strain-Measuring Systems

Strain-measuring systems are based on a variety of sensors, including mechanical, optical, and electrical devices. Each has some special advantages but can usually be adapted for other needs as well. No one system is entirely satisfactory for all practical requirements, so it is necessary to optimize the gage system to each problem according to a set of desirable characteristics. Some of the common characteristics used to evaluate the system's adequacy for a typical application are

1. The calibration constant for the gage should be stable; it should not vary with either time, temperature, or other environmental factors.
2. The gage should be able to measure strains with an accuracy of ±1 με over a strain range of ±10%.
3. The gage size, i.e., the gage length l_0 and width w_0, should be small so that strain at a point is approximated with small error.
4. The response of the gage, largely controlled by its inertia, should be sufficient to permit recording of dynamic strains with frequency components exceeding 100 kHz.
5. The gage system should permit on-location or remote readout.
6. Both the gage and the associated auxiliary equipment should be inexpensive.
7. The gage system should be easy to install and operate.
8. The gage should exhibit a linear response to strain over a wide range.

Three of these basic characteristics deserve further mention here: the gage length l_0, the gage sensitivity, and the range of the strain gage. The gage length is often the most important because in nonlinear strain fields the error will depend on the gage length.

Sensitivity is the smallest value of strain that can be read on the scale associated with the strain gage and should not be mistaken for accuracy or precision. The sensitivity chosen should not be higher than necessary because it needlessly increases the complexity of the measuring method and introduces new problems.

The range of the strain gage refers to the maximum value of strain that can be recorded. Since the range and sensitivity of the gage are interrelated, it is often necessary to compromise between the two for optimal performance of both. Various compromises have resulted in the two main kinds of strain gages, extensometers and electrical strain gages. There are numerous electrical strain gage systems, but only electrical-resistance strain gages will be considered here.

Extensometers

Various extensometers involving mechanical, electrical, magnetic, or optical devices are used in material test systems. A typical extensometer (Figure 1.5.38) is used in the conventional tensile test where the stress-strain diagram is recorded. This kind of extensometer is attached to the specimen by knife edges and spring clips. Electrical-resistance strain gages are attached to the cross-flexural member and provide the strain output. The main advantage of extensometers is that they can be reused and recalibrated after each test. The disadvantages are that they are much larger and more expensive than electrical-resistance strain gages.

Electrical-Resistance Strain Gages

The electrical-resistance strain gage fulfills most of the requirements of an optimum system and is widely used for experimental stress analysis. The electrical-resistance strain gage consists of a metal-foil grid bonded to a polymer backing (Figure 1.5.39). A Wheatstone bridge is often used in this system to enhance the ability to measure changes in resistance. As a specimen is deformed the strain is transmitted

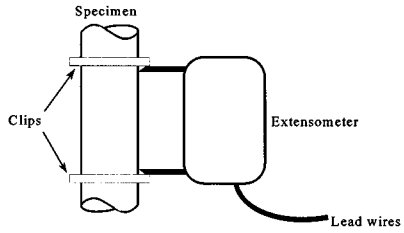


FIGURE 1.5.38 Extensometer attached to a tensile specimen.

to the grid, which has a current applied to it. The change in resistance of the grid is converted to a voltage signal output of the Wheatstone bridge. The basic equation used with this system is

$$\frac{\Delta R}{R} = S_A \varepsilon \tag{1.5.61}$$

where R is the resistance of the gage, ε is the applied strain, and S_A is the sensitivity, or gage factor, of the metallic alloy used in the conductor. The most commonly used alloy is a copper-nickel alloy called Advance, for which the sensitivity is 2.1.

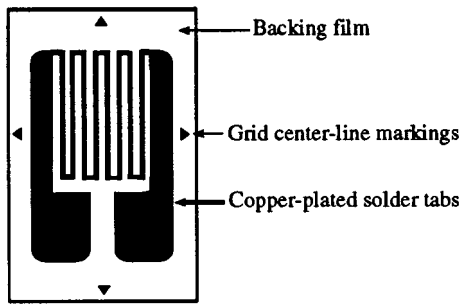


FIGURE 1.5.39 Model of metal-foil strain gages.

Electrical-Resistance Strain Gage Mounting Methods

For precision strain measurements, both the correct adhesive and proper mounting procedures must be employed. The adhesive serves a vital function in the strain-measuring system; it must transmit the strain from the specimen to the sensing element without distortion. Bonding a strain gage to a specimen is one of the most critical steps in the entire process of measuring strain with an electric-resistance strain gage. When mounting a strain gage, it is important to carefully prepare the surface of the component where the gage is to be located. This includes sanding, degreasing, etching, cleaning, and finally neutralizing the surface where the gage is to be mounted. Next, the surface is marked to allow accurate orientation of the strain gage. The gage is then put in place and held with tape while the adhesive is allowed to dry. Several of the adhesive systems commonly used for this are epoxy cements, cyanoacrylate cement, polyester adhesives, and ceramic adhesives. Once the adhesive has been placed, the drying process becomes vitally important, as it can cause residual stresses in the grid work of the gage which could influence the output. After allowing the adhesive to dry, the cure must be tested to ensure complete drying. Failure to do so will affect the stability of the gage and the accuracy of the output. The cure state of the adhesive can be tested by various resistance tests. Also, the bonded surface is inspected to determine if any voids are present between the gage and the specimen due to bubbling of the adhesive.

After the bonding process is complete, the lead wires are attached from the soldering tabs of the gage to an anchor terminal, which is also bonded to the test specimen. This anchoring terminal is used to protect the fragile metal-foil gages. Finally, wires are soldered from this anchor terminal to the instrumentation being used to monitor the resistance changes.

Gage Sensitivities and Gage Factor

The electrical-resistance strain gage has a sensitivity to both axial and transverse strain. The magnitude of the transverse strain transmitted to the grid depends on a number of factors, including the thickness and elastic modulus of the adhesive, the carrier material, the grid material, and the width-to-thickness ratio of the axial segments of the grid. Sometimes it is necessary to calculate the true value of strain that includes all contributions, from

$$\varepsilon_a = \frac{(\Delta R/R)}{S_g}\frac{1-\nu_0 K_t}{1+K_t(\varepsilon_t/\varepsilon_a)} \tag{1.5.62}$$

where ε_a is the normal strain along the axial direction of the gage, ε_t is the normal strain along the transverse direction of the gage, $\nu_0 = 0.285$ is Poisson's ratio for the calibration beam, and K_t is the transverse-sensitivity factor of the gage. The strain gage sensitivity factor, S_g, is a calibration constant provided by the manufacturer. By using Equations 1.5.61 and 1.5.62 the percent error involved in neglecting the transverse sensitivity can be calculated. These errors can be significant for large values of both K_t and $\varepsilon_t/\varepsilon_a$, so it may be necessary to correct for the transverse sensitivity of the gage (Figure 1.5.40).

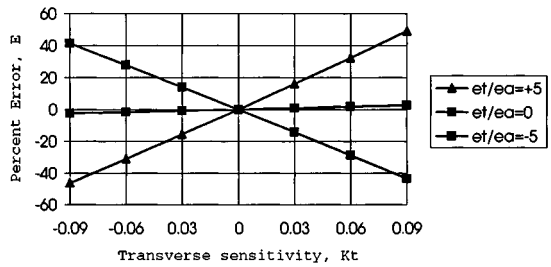


FIGURE 1.5.40 Error as a function of transverse-sensitivity factor with the biaxial strain ratio as a parameter.

Strain Analysis Methods

Electrical-resistance strain gages are normally employed on the free surface of a specimen to establish the stress at a particular point on this surface. In general it is necessary to measure three strains at a point to completely define either the stress or the strain field. For this general case, where nothing is known about the stress field or its directions before experimental analysis, three-element rosettes are required to establish the stress field. This is accomplished by using the three-element gage with orientations at arbitrary angles, as shown in Figure 1.5.41. Using this setup, the strains ε_x, ε_y, and γ_{xy} can be determined. These values can be used to determine the principal strains and principal directions,

$$\varepsilon_1 = \frac{1}{2}(\varepsilon_{xx}+\varepsilon_{yy}) + \frac{1}{2}\sqrt{(\varepsilon_{xx}-\varepsilon_{yy})^2+\gamma_{xy}^2}$$

$$\varepsilon_2 = \frac{1}{2}(\varepsilon_{xx}+\varepsilon_{yy}) - \frac{1}{2}\sqrt{(\varepsilon_{xx}-\varepsilon_{yy})^2+\gamma_{xy}^2} \tag{1.5.63}$$

$$\tan 2\phi = \frac{\gamma_{xy}}{\varepsilon_{xx}-\varepsilon_{yy}}$$

where ϕ is the angle between the principal axis (σ_1) and the *x* axis. The principal stresses can be computed using the principal strains,

$$\sigma_1 = \frac{E}{1-\nu^2}\left(\varepsilon_1 + \nu\varepsilon_2\right)$$
$$\sigma_2 = \frac{E}{1-\nu^2}\left(\varepsilon_2 + \nu\varepsilon_1\right) \tag{1.5.64}$$

These expressions give the complete state of stress since the principal directions are known from Equation 1.5.63.

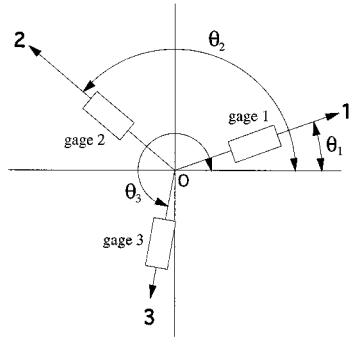


FIGURE 1.5.41 Three gage elements placed at arbitrary angles relative to the *x* and *y* axes.

Optical Methods of Strain Analysis

Moiré Method of Strain Analysis. The moiré technique depends on an optical phenomenon of fringes caused by relative displacement of two sets of arrays of lines. The arrays used to produce the fringes may be a series of straight parallel lines, a series of radial lines emanating from a point, a series of concentric circles, or a pattern of dots. The straight parallel line "grids" are used most often for strain analysis work and consist of equal width lines with opaque spacing of the same width between them. These straight parallel lines are spaced in a "grating" scheme of typically 50 to 1000 lines per inch for moiré work. In the cross-grid system of two perpendicular line arrays, the grid placed on the specimen is referred to as the model grid. The second grid is referred to as the reference grid and is overlaid on top of the model grid. Often a thin layer of oil or some other low-friction substance is placed between the model grid and the reference grid to keep them in contact while attempting to minimize the transmission of strains from the model to the reference grid.

To obtain a moiré fringe pattern the grids are first aligned on the unloaded model so that no pattern is present. The model is loaded and light is transmitted through the two grids. Strain displacement is observed in the model grid while the reference grid remains unchanged. A moiré fringe pattern is formed each time the model grating undergoes a deformation in the primary direction equal to the pitch p of the reference grating. For a unit gage length, $\Delta L = np$, where ΔL is the change in length per unit length, p is the pitch of the reference grating and n is the number of fringes in the unit gage length. In order to calculate ε_x, ε_y, and γ_{xy}, two sets of gratings must be applied in perpendicular directions. Then displacements u and v (displacements in the x and y directions, respectively) can be established and the Cartesian strain components can be calculated from slopes of the displacement surfaces: $\varepsilon_{xx} = \partial u/\partial x$, $\varepsilon_{yy} = \partial v/\partial y$, and $\gamma_{xy} = \partial v/\partial x + \partial u/\partial y$. The displacement gradients in the z direction, $\partial w/\partial x$ and $\partial w/\partial y$, have been neglected here because they are not considered in moiré analysis of in-plane deformation fields.

Photoelasticity. The method of photoelasticity is based on the physical behavior of transparent, noncrystalline, optically isotropic materials that exhibit optically anisotropic characteristics, referred to as temporary double refraction, while they are stressed. To observe and analyze these fringe patterns a device called a polariscope is used. Two kinds of polariscope are common, the plane polariscope and the circular polariscope.

The plane polariscope (Figure 1.5.42) consists of a light source, two polarizing elements, and the model. The axes of the two polarizing elements are oriented at a 90° angle from each other. If the specimen is not stressed, no light passes through the analyzer and a dark field is observed. If the model is stressed, two sets of fringes, isoclinics and isochromatics, will be obtained. Black isoclinic fringe patterns are the loci of points where the principal-stress directions coincide with the axis of the polarizer. These fringe patterns are used to determine the principal stress directions at all points of a photoelastic model. When the principal stress difference is zero ($n = 0$) or sufficient to produce an integral number of wavelengths of retardation ($n = 1, 2, 3, ...$), the intensity of light emerging from the analyzer is zero. This condition for extinction gives a second fringe pattern, called isochromatics, where the fringes are the loci of points exhibiting the same order of extinction ($n = 0, 1, 2, 3, \ldots$).

$$n = N = \frac{h}{f_\sigma}(\sigma_1 - \sigma_2) \tag{1.5.65}$$

where N is the isochromatic fringe order. The order of extinction n depends on the principal stress difference $(\sigma_1 - \sigma_2)$, the thickness h of the model, and the material fringe value f_σ. When monochromatic light is used, the isochromatic fringes appear as dark bands. When white light is used, the isochromatic fringes appear as a series of colored bands. Black fringes appear in this case only where the principal stress difference is zero.

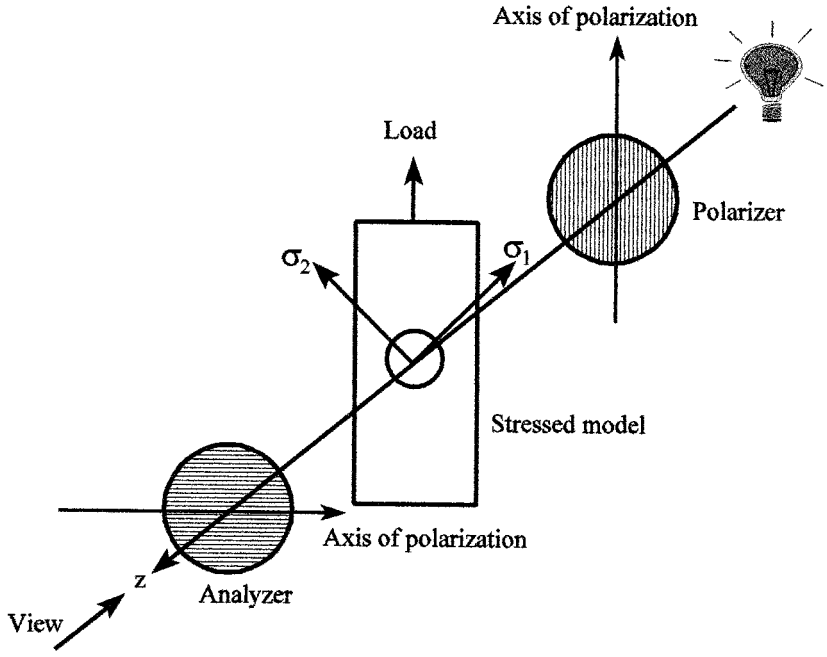


FIGURE 1.5.42 Schematic of a stressed photoelastic model in a plane polariscope.

A circular polariscope is a plane polariscope with two additional polarizing plates, called quarter-wave plates, added between the model and the original polarizing plates (Figure 1.5.43). The two quarter-wave plates are made of a permanently doubly refracting material. The circular polariscope is used to eliminate the isoclinic fringes while maintaining the isochromatic fringes. To accomplish this, monochromatic light must be used since the quarter-wave plates are designed for a specific wavelength of light. For the dark-field arrangement shown, no light is passed through the polariscope when the model is unstressed. A light-field arrangement is achieved by rotating the analyzer 90°. The advantage of using

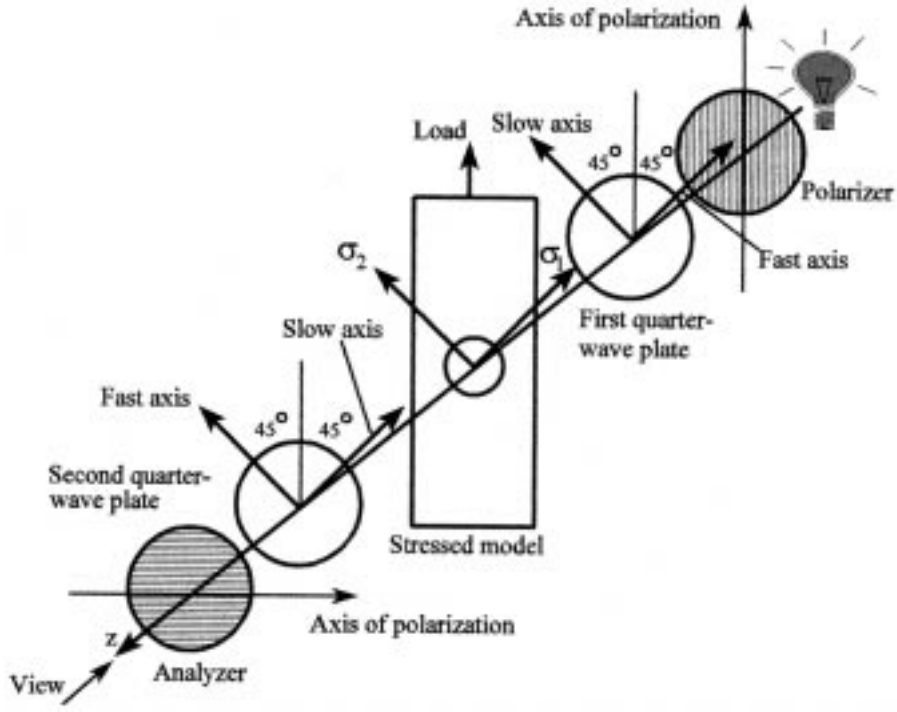


FIGURE 1.5.43 Schematic of a stressed photoelastic model in a circular polariscope.

both light- and dark-field analysis is that twice as much data is obtained for the whole-field determination of $\sigma_1 - \sigma_2$. If a dark-field arrangement is used, n and N still coincide, as in Equation 1.5.65. If a light-field arrangement is used, they are not coincident. In this case Equation 1.5.65 becomes

$$N = \frac{1}{2} + n = \frac{h}{f_\sigma}(\sigma_1 - \sigma_2) \qquad n = 0,\ 1,\ 2,\ 3,\ \ldots \tag{1.5.66}$$

By determining both the isoclinic fringes and the isochromatic fringes, the principal-stress directions and the principal-stress difference can be obtained. In order to obtain the individual principal stresses, a stress separation technique would need to be employed.

The advantages of the photoelastic method are that it allows a full-field stress analysis and it makes it possible to determine both the magnitude and direction of the principal stresses. The disadvantages are that it requires a plastic model of the actual component and it takes a considerable effort to separate the principal stresses.

Thermoelastic Stress Analysis. Modern thermoelastic stress analysis (TSA) employs advanced differential thermography (or AC thermography) methods based on dynamic thermoelasticity and focal-plane-array infrared equipment capable of rapidly measuring small temperature changes (down to 0.001°C) caused by destructive or nondestructive alternating stresses. Stress resolutions comparable to those of strain gages can be achieved in a large variety of materials. The digitally stored data can be processed in near-real time to determine the gradient stress fields and related important quantities (such as combined-mode stress intensity factors) in complex components and structures, with no upper limit in temperature. The efficient, user-friendly methods can be applied in the laboratory and in the field, in vehicles, and structures such as bicycles, automobiles, aircraft, surgical implants, welded bridges, and microelectronics. Optimum

design, rapid prototyping, failure analysis, life prediction, and rationally accelerated testing can be facilitated with the new TSA methods (Color Plates 8 and 11 to 14).

Brittle Coatings. If a coating is applied to a specimen that is thin in comparison with the thickness of the specimen, then the strains developed at the surface of the specimen are transmitted without significant change to the coating. This is the basis of the brittle coating method of stress analysis. The two kinds of coatings available are resin-based and ceramic-based coatings. The ceramic-based coatings are seldom used due to the high application temperatures (950 to 1100°F) required. The coatings are sprayed on the component until a layer approximately 0.003 to 0.010 in. thick has accumulated. It is also necessary to spray calibration bars with the coating at the same time in order to obtain the threshold strain at which the coating will crack. These calibration bars are tested in a cantilever apparatus and the threshold strain is calculated using the flexure formula and Hooke's law. Once the threshold strain is known and the actual specimen has been tested, the principal stress perpendicular to the crack can be determined by using Hooke's law. The procedure is to load the component, apply the coating, and then quickly release the loading in steps to observe any cracks.

The main advantages of this method are that both the magnitude and direction of the principal strains can be quickly obtained and that the coating is applied directly to the component. This also allows a quick analysis of where the maximum stress regions are located so that a better method can be used to obtain more accurate results. The main disadvantage is that the coatings are very sensitive to ambient temperature and might not have sufficiently uniform thickness.

Mechanical Testing

Standards. Many engineering societies have adopted mechanical testing standards; the most widely accepted are the standards published by the American Society for Testing and Materials. Standards for many engineering materials and mechanical tests (tension, compression, fatigue, plane strain fracture toughness, etc.) are available in the *Annual Book of ASTM Standards*.

Open-Loop Testing Machines. In an open-loop mechanical testing system there is no feedback to the control mechanism that would allow for continuous adjustment of the controlled parameter. Instead, the chosen parameter is "controlled" by the preset factory adjustments of the control mechanism. It is not possible for such a machine to continually adjust its operation to achieve a chosen (constant or not constant) displacement rate or loading rate.

A human operator can be added to the control loop in some systems in an attempt to maintain some parameter, such as a loading rate, at a constant level. This is a poor means of obtaining improved equipment response and is prone to error.

Closed-Loop Testing Machines. In a closed-loop, most commonly electrohydraulic, testing system, a servo controller is used to continuously control the chosen parameter. When there is a small difference between the desired value that has been programmed in and the actual value that is being measured, the servo controller adjusts the flow of hydraulic fluid to the actuator to reduce the difference (the error). This correction occurs at a rate much faster than any human operator could achieve. A standard system makes 10,000 adjustments per second automatically.

A typical closed-loop system (Color Plates 9, 11, 15) allows the operator to control load, strain, or displacement as a function of time and can be adjusted to control other parameters as well. This makes it possible to perform many different kinds of tests, such as tension, compression, torsion, creep, stress relaxation, fatigue, and fracture.

Impact Testing. The most common impact testing machines utilize either a pendulum hammer or a dropped weight. In the pendulum system a hammer is released from a known height and strikes a small notched specimen, causing it to fracture. The hammer proceeds to some final height. The difference between the initial and final heights of the hammer is directly proportional to the energy absorbed by the specimen. For the Charpy test the specimen is mounted horizontally with the ends supported so that

the pendulum will strike the specimen in midspan, opposite the notch. In the Izod test the specimen bottom is mounted in a vertical cantilever support so that the pendulum will strike the specimen at a specific distance above the notch, near the unsupported top end.

A large variety of the drop-weight tests are also available to investigate the behaviors of materials and packages during impact.

Hardness Testing. The major hardness tests are the Brinell, Rockwell, Vickers, and Shore scleroscope tests.

The Brinell hardness test uses a hardened steel ball indenter that is pushed into the material under a specified force. The diameter of the indentation left in the surface of the material is measured and a Brinell hardness number is calculated from this diameter.

The Rockwell hardness test differs from the Brinell test in that it uses a 120° diamond cone with a spherical tip for hard metals and a 1/16-in. steel ball for soft metals. The Rockwell tester gives a direct readout of the hardness number. The Rockwell scale consists of a number of different letter designators (B, C, etc.) based on the depth of penetration into the test material.

The Vickers hardness test uses a small pyramidal diamond indenter and a specified load. The diagonal length of the indentation is measured and used to obtain the Vickers hardness number.

The Shore scleroscope uses a weight that is dropped on the specimen to determine the hardness. This hardness number is determined from the rebound height of the weight.

1.6 Structural Integrity and Durability

Bela I. Sandor

The engineer is often concerned about the long-term behavior and durability of machines and structures. Designs based just on statics, dynamics, and basic mechanics of materials are typically able to satisfy only minimal performance and reliability requirements. For realistic service conditions, there may be numerous degradations to consider. A simple and common approach is to use safety factors based on experience and judgment. The degradations could become severe and require sophisticated analyses if unfavorable interactions occur. For example, fatigue with corrosion or high temperatures is difficult to predict accurately, and much more so when corrosion is occurring at a high temperature.

There are many kinds of degradations and interactions between them, and a large (and still growing) technical literature is available in most of these areas. The present coverage cannot possibly do justice to the magnitude of the most serious problems and the available resources to deal with them. Instead, the material here is to highlight some common problems and provide fundamental concepts to prepare for more serious efforts. The reader is encouraged to study the technical literature (including that by technical societies such as ASM, ASME, ASNT, ASTM, SAE), attend specialized short courses, and seek consulting advice (ASM, ASTM, Teltech) as necessary.

Finite Element Analysis. Stress Concentrations

The most common problem in creating a machine or structure with good strength-to-weight ratio is to identify its critical locations and the corresponding maximum stresses or strains and to adjust the design optimally. This is difficult if a member's geometry, including the geometry and time-dependence of the loading, is complex. The modern analytical tool for addressing such problems is finite element analysis (FEA) or finite element modeling (FEM).

Finite Element Analysis

The finite element (FE) method was developed by engineers using physical insight. In all applications the analyst seeks to calculate a *field quantity*: in stress analysis it is the displacement field or the stress field; in thermal analysis it is the temperature field or the heat flux; and so on. Results of the greatest interest are usually peak values of either the field quantity or its gradients. The FE method is a way of getting a *numerical* solution to a *specific* problem. An FEA does not produce a formula as a solution, nor does it solve a class of problems. Also, the solution is approximate unless the problem is so simple that a convenient exact formula is already available. Furthermore, it is important to validate the numerical solution instead of trusting it blindly.

The power of the FE method is its versatility. The structure analyzed may have arbitrary shape, arbitrary supports, and arbitrary loads. Such generality does not exist in classical analytical methods. For example, temperature-induced stresses are usually difficult to analyze with classical methods, even when the structure geometry and the temperature field are both simple. The FE method treats thermal stresses as readily as stresses induced by mechanical load, and the temperature distribution itself can be calculated by FE. However, it is easy to make mistakes in describing a problem to the computer program. Therefore *it is essential that the user have a good understanding of the problem and the modeling* so that errors in computed results can be detected by judgment.

Stress Concentrations

Geometric discontinuities cause localized stress increases above the average or far-field stress. A stress raiser's effect can be determined quantitatively in several ways, but not always readily. The simplest method, if applicable, is to use a known theoretical **stress concentration factor**, K_t, to calculate the peak stress from the nominal, or average, value,

$$\sigma_{\max} = K_t \sigma_{ave} \tag{1.6.1}$$

This is illustrated in Figure 1.6.1. The area under the true stress distribution always equals the area under the nominal stress level,

$$\int_A \sigma_{true}\, dA = \int_A \sigma_{ave}\, dA = \sigma_{ave} A \tag{1.6.2}$$

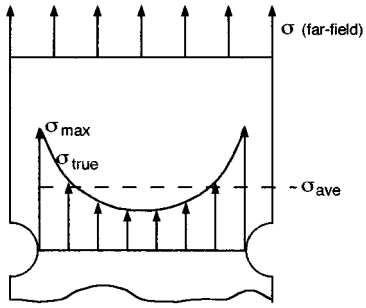


FIGURE 1.6.1 Stress distribution (simplistic) in a notched member under uniaxial load.

The factor K_t depends mainly on the geometry of the notch, not on the material, except when the material deforms severely under load. K_t values are normally obtained from plots such as in Figure 1.6.2 and are strictly valid only for ideally elastic, stiff members. K_t values can also be determined by FEA or by several experimental techniques. There are no K_t values readily available for sharp notches and cracks, but one can always assume that such discontinuities produce the highest stress concentrations, sometimes factors of tens. This is the reason for brittle, high-strength materials being extremely sensitive even to minor scratches. In fatigue, for example, invisible toolmarks may lead to premature, unexpected failures in strong steels.

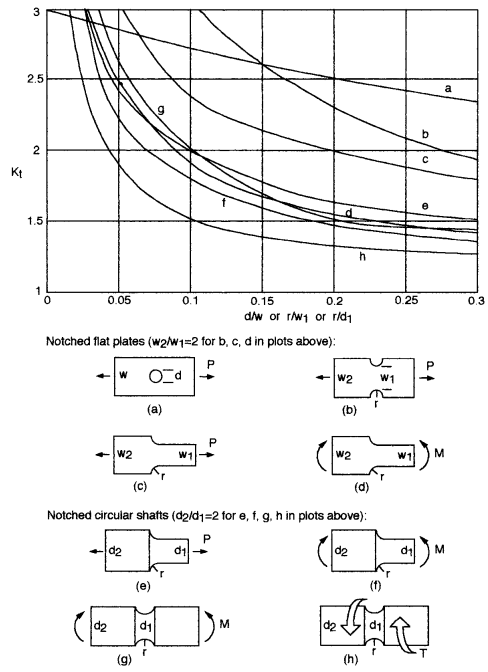


FIGURE 1.6.2 Samples of elastic stress concentration factors. (Condensed from Figures 10.1 and 10.2, Dowling, N. E. 1993. *Mechanical Behavior of Materials*. Prentice-Hall, Englewood Cliffs, NJ. With permission.)

There are many other factors that may seem similar to K_t, but they should be carefully distinguished. The first is the true stress concentration factor K_σ, defined as

$$K_\sigma = \frac{\sigma_{max}}{\sigma_{ave}} \tag{1.6.3}$$

which means that $K_\sigma = K_t$ (by Equation 1.6.1) for ideally elastic materials. K_σ is most useful in the case of ductile materials that yield at the notch tip and lower the stress level from that indicated by K_t.

Similarly, a true strain concentration factor, K_ε, is defined as

$$K_\varepsilon = \frac{\varepsilon_{max}}{\varepsilon_{ave}} \tag{1.6.4}$$

where $\varepsilon_{ave} = \sigma_{ave}/E$.

Furthermore, a large number of **stress intensity factors** are used in fracture mechanics, and these (such as K, K_c, K_I, etc.) are easily confused with K_t and K_σ, but their definitions and uses are different as seen in the next section.

Fracture Mechanics

Notches and other geometric discontinuities are common in solid materials, and they tend to facilitate the formation of cracks, which are in turn more severe stress raisers. Sharp cracks and their further growth are seldom simple to analyze and predict, because the actual stresses and strains at a crack tip are not known with the required accuracy. In fact, this is the reason the classical failure theories (maximum normal stress, or Rankine, theory; maximum shear stress, or Tresca, theory; distortion energy, or von Mises or octahedral shear stress, theory), elegantly simple as they are, are not sufficiently useful in dealing with notched members. A powerful modern methodology in this area is fracture mechanics, which was originated by A. A. Griffith** in 1920 and has grown in depth and breadth enormously in recent decades. The space here is not adequate to even list all of the significant references in this still expanding area. The purpose here is to raise the engineer's awareness to a quantitative, practically useful approach in dealing with stress concentrations as they affect structural integrity and durability.

Brittle and Ductile Behaviors. Embrittlements

Brittleness and ductility are often the first aspects of fracture considerations, but they often require some qualifications. Simplistically, a material that fractures in a tension test with 0% reduction of area (RA) is perfectly brittle (and very susceptible to fracture at stress raisers), while one with 100% RA is perfectly ductile (and quite tolerant of discontinuities). Between these extremes fall most engineering materials, with the added complication that embrittlement is often made possible by several mechanisms or environmental conditions. For example, temperature, microstructure, chemical environment, internal gases, and certain geometries are common factors in embrittlement. A few of these will be discussed later.

** The Griffith criterion of fracture states that a crack may propagate when the decrease in elastic strain energy is at least equal to the energy required to create the new crack surfaces. The available elastic strain energy must also be adequate to convert into other forms of energy associated with the fracture process (heat from plastic deformation, kinetic energy, etc.). The critical nominal stress for fracture according to the Griffith theory is proportional to $1/\sqrt{\text{crack length}}$. This is significant since crack length, even inside a member, is easier to measure nondestructively than stresses at a crack tip. Modern, practical methods of fracture analysis are sophisticated engineering tools on a common physical and mathematical basis with the Griffith theory.

Linear-Elastic Fracture Mechanics (LEFM)

A major special case of fracture mechanics is when little or no plastic deformations occur at the critical locations of notches and cracks. It is important that even intrinsically ductile materials may satisfy this condition in common circumstances.

Modes of Deformation. Three basic modes of deformation (or crack surface displacement) of cracked members are defined as illustrated schematically in Figure 1.6.3. Each of these modes is very common, but Mode I is the easiest to deal with both analytically and experimentally, so most data available are for Mode I.

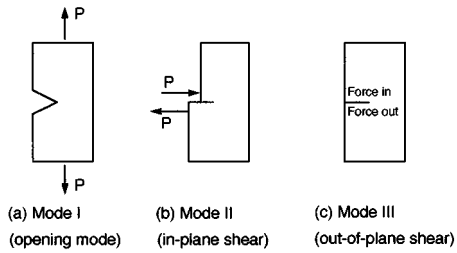


FIGURE 1.6.3 Modes of deformation.

Stress Intensity Factors. The stresses on an infinitesimal element near a crack tip under Mode I loading are obtained from the theory of linear elasticity. Referring to Figure 1.6.4,

$$\sigma_x = \frac{K_I}{\sqrt{2\pi r}} f_1(\theta) + \ldots$$

$$\sigma_y = \frac{K_I}{\sqrt{2\pi r}} f_2(\theta) + \ldots \qquad (1.6.5)$$

$$\tau_{xy} = \frac{K_I}{\sqrt{2\pi r}} f_3(\theta) + \ldots$$

$$\tau_{xz} = \tau_{zx} = 0$$

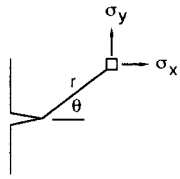


FIGURE 1.6.4 Coordinates for fracture analysis.

There are two special cases of σ_z:

$\sigma_z = 0$ for plane stress (thin members)
$\sigma_z = \nu(\sigma_x + \sigma_y)$ for plane strain, with $\varepsilon_z = 0$ (thick members)

The factor K in these and similar expressions characterizes the intensity or magnitude of the stress field near the crack tip. It is thus called the stress intensity factor, which represents a very useful concept, but different from that of the well-known stress concentration factor. K_I is a measure of the severity of a crack, and most conveniently it is expressed as

$$K_I = \sigma\sqrt{\pi a}\, f(\text{geometry}) \tag{1.6.6}$$

where a is the crack length and f is a function of the geometry of the member and of the loading (typically, $f \cong 1 \pm 0.25$). Sometimes f includes many terms, but all stress intensity factors have the same essential features and units of stress $\sqrt{\text{length}}$. In any case, expressions of K for many common situations are available in the literature, and numerical methods are presented for calculating special K values. Differential thermography via dynamic thermoelasticity is a powerful, efficient modern method for the measurement of actual stress intensity factors under a variety of complex conditions (Section 1.6, "Experimental Stress Analysis and Mechanical Testing"; Figure 1.6.12; Color Plates 8 and 11 to 14).

Fracture Toughness of Notched Members

The stress intensity factor, simply K for now, is analogous to a stress-strain curve, as in Figure 1.6.5. K increases almost linearly from 0 at $\sigma = 0$, to a value K_c at a critical (fracture) event. K_c is called the *fracture toughness* of a particular member tested. It does depend on the material, but it is not a reliable material property because it depends on the size of the member too much. This is illustrated in Figure 1.6.6 for plates of the same material but different thicknesses.

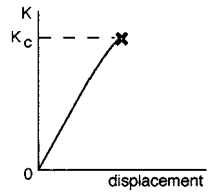


FIGURE 1.6.5 K_c = fracture toughness of a particular member.

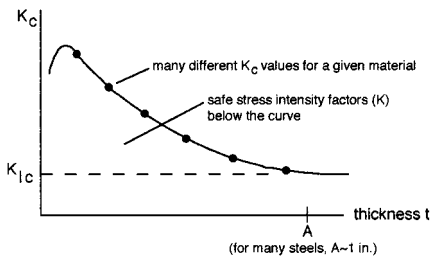


FIGURE 1.6.6 K_{Ic} = plane strain fracture toughness of material.

At very small thickness, K_c tends to drop. More significantly, K_c approaches a lower limiting value at large thickness (>*A*). This worst-case value of K_c is called K_{Ic}, the *plane strain fracture toughness* in Mode I. It may be considered a pseudomaterial property since it is independent of geometry at least over a range of thicknesses. It is important to remember that the thickness effect can be rather severe. An intrinsically ductile metal may fracture in an apparently brittle fashion if it is thick enough and has a notch.

Fracture Toughness Data. Certain criteria about crack sharpness and specimen dimensions must be satisfied in order to obtain reliable basic K_{Ic} data (see *ASTM Standards*). K_{Ic} data for many engineering materials are available in the technical literature. A schematic overview of various materials' K_{Ic} values is given in Figure 1.6.7. Note that particular expected values are not necessarily attained in practice. Poor material production or manufacturing shortcomings and errors could result in severely lowered toughness. On the other hand, special treatments or combinations of different but favorably matched materials (as in composites) could substantially raise the toughness.

Besides the thickness effect, there are a number of major influences on a given material's toughness, and they may occur in favorable or unfavorable combinations. Several of these are described here schematically, showing general trends. Note that some of the actual behavior patterns are not necessarily as simple or well defined as indicated.

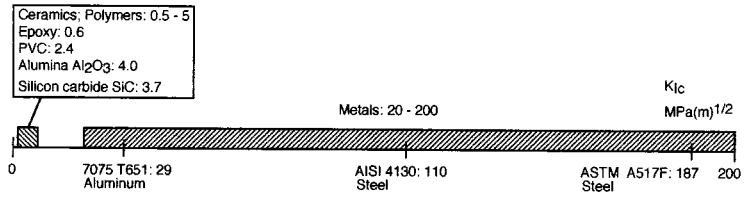


FIGURE 1.6.7 Plane strain fracture toughness ranges (approximate).

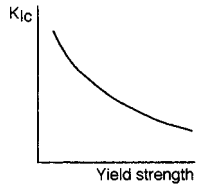


FIGURE 1.6.8 Yield strength effect on toughness.

Yield Strength. High yield strength results in a low fracture toughness (Figure 1.6.8), and therefore it should be chosen carefully, understanding the consequences.

Temperature. Two kinds of temperature effect on toughness should be mentioned here. They both may appear, at least for part of the data, as in Figure 1.6.9, with high temperature causing increased toughness. One temperature effect is by the increased ductility at higher temperature. This tends to lower the yield strength (except in low-carbon steels that strain-age at moderately elevated temperatures, about 100 to 500°C), increase the plastic zone at the notch tip, and effectively blunt the stress concentration. Another effect, the distinct temperature-transition behavior in low-carbon steels (BCC metals, in general; easily shown in Charpy tests), is caused by microstructural changes in the metal and is relatively complex in mechanism.

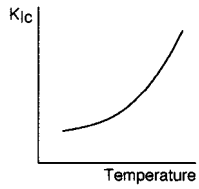


FIGURE 1.6.9 Temperature effect on toughness.

Loading Rate. The higher the rate of loading, the lower the fracture toughness in most cases. Note that toughness results obtained in notch-impact or explosion tests are most relevant to applications where the rate of loading is high.

Microstructural Aspects. In some cases apparently negligible variations in chemical composition or manufacturing processes may have a large effect on a material's fracture toughness. For example, carbon, sulfur, and hydrogen contents may be significant in several embrittling mechanisms. Also, the common mechanical processing of cold or hot working (rolling, extruding, forging) influences the grain structure (grain size and texture) and the corresponding toughness. Neutron radiation also tends to cause microscopic defects, increasing the yield strength and consequently lowering the ductility and toughness of the material.

Overview of Toughness Degradations. There is a multitude of mechanisms and situations that must be considered singly and in realistic combinations, as illustrated schematically in Figure 1.6.10 (review Figure 1.6.6 for relevant toughness definitions).

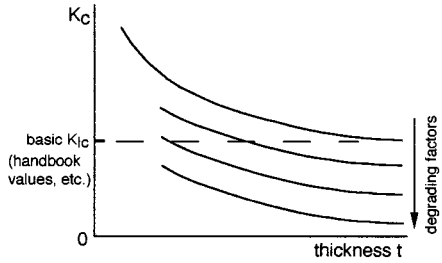


FIGURE 1.6.10 Trends of toughness degradations.

Degrading factors
Some chemical compositions
Sharper notch
Greater thickness
Faster loading
Lower temperature
Higher yield strength
Hostile chemical environment
Liquid metal embrittlement
Tensile residual stress
Neutron irradiation
Microstructural features
Moisture
Gases in solid solution
Surface hardening

Note: The toughness can drop essentially to zero in some cases.

Crack Propagation

Crack growth may be classified as either stable (subcritical) or unstable (critical). Often stable cracks become unstable in time, although the opposite behavior, cracks decelerating and even stopping, is sometimes possible. Unstable cracks under load control are extremely dangerous because they propagate at speeds nearly 40% of the speed of sound in that particular solid. This means, for example in steels, a crack growth speed of about 1 mi/sec. Thus, warnings and even electronically activated, automated countermeasures during the unstable propagation are useless. The only reasonable course is to provide, by design and proper manufacture, preventive measures such as ductile regions in a structure where cracks become stable and slow to grow, allowing for inspection and repair.

There are three kinds of stable crack growth, each important in its own right, with interactions between them possible. Under steady loads, environmentally assisted crack growth (also called stress corrosion cracking) and creep crack growth are commonly found. Under cyclic loading fatigue crack growth is likely to occur. In each case the **rate of crack growth** tends to accelerate in time or with progressive cycles of load if the loads are maintained while the cracks reduce the load-bearing cross-sectional area. This common situation, caused by increasing true stresses, is illustrated schematically in Figure 1.6.11, where a_0 is an initial flaw's size, da/dN and da/dt are the fatigue and creep crack growth rates, respectively, and a_c is the critical crack size. The rate of stable crack growth is controlled by the stress intensity factor. This will be discussed later.

Design and Failure Analysis Using Stress Intensity Concepts

The concept of stress intensity of cracked members is highly useful and practical. Three major possibilities are outlined here with respect to the essential framework of

$$K \propto \text{stress} \sqrt{\text{crack length}} \tag{1.6.7}$$

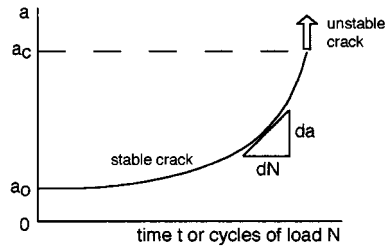


FIGURE 1.6.11 Crack growth rates under constant load.

Here K may be either an operating stress intensity factor or a K_{Ic} value, a material property (the units are the same). In design, the idea is to fix one or two quantities by some initial constraints of the case, then work with the results according to Equation 1.6.7.

1. Operating stress and material (K_{Ic}) are predetermined. This forces one to measure crack length and set the maximum allowable size of cracks.
2. Operating stress and detectable crack size are predetermined. This forces one to choose an appropriate material with the required K_{Ic} value.
3. The material (K_{Ic} value) and the detectable crack size are predetermined. This forces one to limit the operating stress accordingly.

Similar thinking can be used in failure analysis and corresponding design iteration. For example, the critical crack size at the end of the stable propagation (and start of the unstable, high-speed growth) can often be determined by looking at the broken parts. The material property, K_{Ic}, can also be estimated from the parts at hand, and thus the stress that caused the failure can be calculated. It can be determined if the stress was within normal bounds or was an overload from misuse of the equipment. These are powerful, quantitative methods that are useful in improving designs and manufacturing.

Special Methods

There are many other important and useful methods in fracture mechanics that cannot even be listed here. For example, there are several methods in the area of elastic-plastic fracture mechanics. Within this area, mainly applicable to thin members of ductile materials, the J-integral approach alone has been covered in a large number of books and journal articles.

Nondestructive Evaluation

Since all of fracture mechanics is based on knowing the crack size and its location and orientation, nondestructive evaluation (NDE) is a major part of quantitative, predictive work in this area. Many techniques of NDE are available, and some are still rapidly evolving. Two major categories of NDE methods are defined here:

1. *Geometry-based methods.* At best, the size, shape, location, and orientation of a flaw are measured. Considerable additional effort is needed to estimate the effect of the flaw on structural integrity and durability. Common methods involve acoustic, magnetic, microwave, optical (including thermal), or X-ray instruments.
2. *Stress-based methods*. A flaw's effect on the stress-strain field is directly measured, which is often much more important than just finding that flaw (a flaw of a given geometry may be benign or malignant, depending on the stress field of the neighborhood). Only a few optical methods are readily available for stress-based NDE; the most effective one for laboratory and field applications is thermoelastic stress analysis by infrared means (Figure 1.6.12; Color Plates 8, 11 to 14; Section 1.5, "Experimental Stress Analysis and Mechanical Testing").

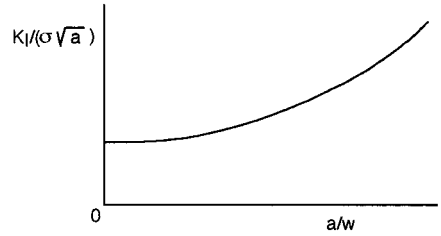


FIGURE 1.6.12 Practical fracture mechanics with NDE: nearly instantaneous measurement of crack size *and* the actual stress intensity factor via advanced thermoelastic stress analysis. The member's loading (including boundary conditions) need not be known to obtain reliable data using this method.

Creep and Stress Relaxation

Creep and stress relaxation are related time- and temperature-dependent phenomena, with creep occurring under load control and stress relaxation under deformation control. In both cases the material's temperature is a governing factor regarding what happens. Specifically, for most metals, the creep and relaxation regimes are defined as high homologous (relative, dimensionless) temperatures, normally those above half the melting point in absolute temperature for each metal. Thus, solder at room temperature creeps significantly under load, while steel and aluminum do not. However, some creep and relaxation may occur even at low homologous temperatures, and they are not always negligible. For polymers, the creep regime is above the glass transition temperature. This is typically not far from room temperature. Figures 1.6.13 and 1.6.14 show trends of creep and stress relaxation in the large-scale phenomenon region.

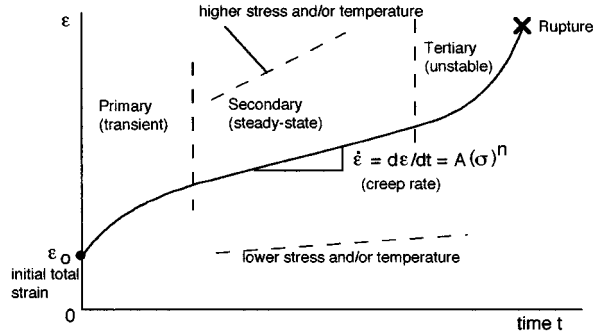


FIGURE 1.6.13 Creep under constant load. $d\varepsilon/dt = A(\sigma)^n$. A and n are material parameters.

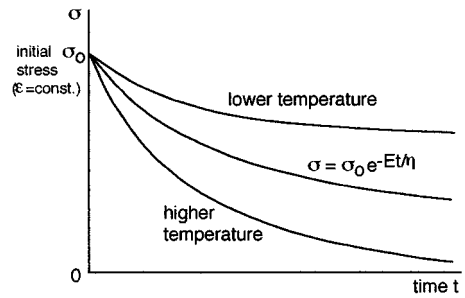


FIGURE 1.6.14 Stress relaxation under constant deformation. $\sigma = \sigma_0 e^{-Et/\eta}$. E and η are material parameters.

Stress vs. rupture life curves for creep may be nearly linear when plotted on log-log coordinates (Figure 1.6.15).

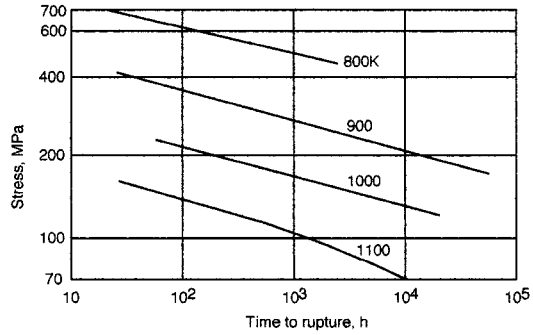


FIGURE 1.6.15 Approximate stress vs. rupture lives of S-590 alloy as functions of temperature. (After Figure 15.8, Dowling, N. E. 1993. *Mechanical Behavior of Materials.* Prentice-Hall, Englewood Cliffs, NJ. With permission.)

Mechanical Models of Viscoelastic Behaviors

Creep and stress relaxation appear to be combinations of behaviors of viscous liquids and elastic solids. The so-called viscoelastic phenomena are commonly modeled by simple mechanical components, springs and dashpots, as in Figure 1.6.16. The Maxwell model and related others are based on such elements.

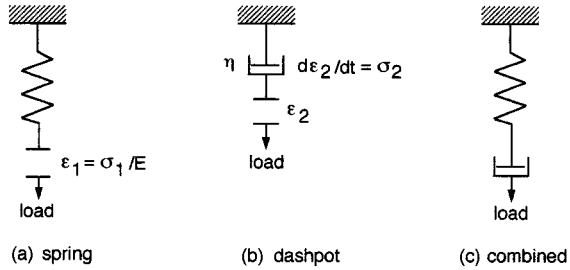


FIGURE 1.6.16 Viscoelastic elements.

The Maxwell model for creep under constant stress σ_0 is

$$\varepsilon = \varepsilon_1 + \varepsilon_2 = \frac{\sigma}{E} + \int_0^t \sigma_0 \, dt = \frac{\sigma_0}{E} + \frac{\sigma_0 t}{\eta} \tag{1.6.8}$$

For relaxation, ε = constant and σ varies, so

$$\frac{d\varepsilon}{dt} = 0 = \frac{1}{E}\frac{d\sigma}{dt} + \frac{\sigma}{\eta} \tag{1.6.9}$$

$$\int_{\sigma_0}^{\sigma} \frac{d\sigma}{\sigma} = -\frac{E}{\eta}\int_0^t dt, \quad \sigma = \sigma_0 e^{-Et/\eta}$$

Time-Temperature Parameters. Life Estimation

It is often necessary to extrapolate from laboratory creep test data, which are limited in time (from days to years), to real service lives, which tend to be from years to several decades. Time-temperature parameters are useful for this purpose. Three common parameters are outlined here. Note that no such parameter is entirely reliable in all cases. They are best if used consistently in direct comparisons of materials.

Sherby-Dorn Parameter (P_{SD})

$$P_{SD} = \log\theta_r = \log t_r - 0.217Q\left(\frac{1}{T}\right) \tag{1.6.10}$$

where, for steady-state creep,

θ_r = temperature-compensated time to rupture
t_r = rupture time, hours
Q = activation energy = constant
T = temperature, K

Stress-life data at high T and low t_r are needed to plot P_{SD} vs. stress, in order to predict a longer t_r at a lower T.

Larson-Miller Parameter (P_{LM})

This approach is analogous to the Sherby-Dorn approach, but is based on different assumptions and equations.

$$P_{LM} = 0.217Q = T\left(\log t_r + C\right) \tag{1.6.11}$$

where $C = -\log\theta_r \cong 20$ for steels. For using temperature in degrees Fahrenheit (as in most of the data),

$$P_{LM}\big|_{\circ F} = 1.8P_{LM}\big|_{K} \tag{1.6.12}$$

Manson-Haferd Parameter (P_{MH})

$$P_{MH} = \frac{T - T_a}{\log t_r - \log t_a} \tag{1.6.13}$$

where T_a and t_a are temperature and time constants representing a point of convergence for a family of data points. As above, for different temperature scales,

$$P_{MH}\big|_{\circ F} = 1.8P_{MH}\big|_{K} \tag{1.6.14}$$

Overview. The greater the extrapolation using any parameter, the greater the likelihood of error. A factor of ten or less extrapolation in life is often reasonable. At very large extrapolations there may be different damage mechanisms from that of the tests, as well as unpredictable service loading and environmental conditions.

Fatigue

Fatigue is a process of damage evolving in a material due to repeated loads, also called cyclic loads. This is a common degradation that affects virtually all solid materials, and thus it is often the main (or a contributing) factor in the failure of vehicles, machinery, structures, appliances, toys, electronic devices, and surgical implants. Many apparently well-designed and -fabricated items that fail inexplicably have problems rooted in the fatigue area.

Nearly two centuries of fatigue studies and engineering efforts have resulted in a huge, and still expanding, technical literature. This brief review can cover only a few major topics, some old but valuable items of wisdom, and practical modern methods. Three important approaches are presented: the stress-based (useful for long lives), strain-based (useful for short lives), and fracture mechanics methods.

Definitions

Constant-amplitude, stress- or strain-controlled cycling is common in testing and some service situations. Figure 1.6.17 shows the stress (σ) quantities in such cycling. Similar notations are used for strains. In completely reversed stress $\sigma_m = 0$ and $R = -1$. Zero-to-tension (a special case of pulsating tension) has $\sigma_{min} = 0$ and $R = 0$.

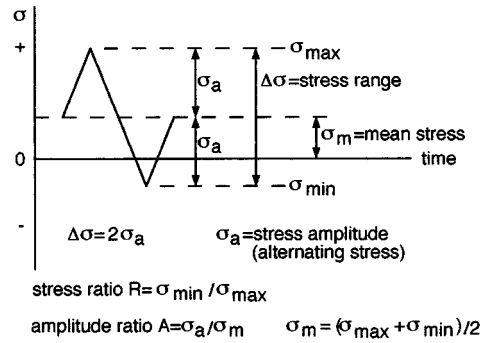


FIGURE 1.6.17 Notation for constant-amplitude stress cycling.

Material Properties in Cyclic Loading

The mechanical properties of some materials are gradually changed by cyclic plastic strains. The changes that occur are largest early in the fatigue life and become negligible beyond about 20 to 50% of the life. The most important material properties that could change significantly this way are the flow properties (yield strength, proportional limit, strain hardening exponent), while the modulus of elasticity is little affected. For metals, three initial conditions can be defined using the strain hardening exponent n as a key parameter. The concept of a cyclic stress-strain curve, as opposed to that in monotonic (static) loading, is also used to distinguish possible material behaviors in fatigue, as follows.

- **Stable**: $0.15 < n < 0.2$ (approx.)

  The monotonic and cyclic stress-strain curves are the same for most practical purposes (though seldom coincident).

  Examples: 7075-T6 Al; 4142 steel (550 BHN)
- **Cycle-Dependent Softening**: $n < 0.15$ (approx.) (means initially hard, cold-worked material)

  The cyclic stress-strain curve falls significantly below the monotonic curve, which means a gradually decreasing deformation resistance as cyclic loading progresses. The cyclic yield strength may be less than half the tensile yield strength in some cases.

  Examples: 4340 steel (350 BHN); 4142 steel (400 BHN)
- **Cycle-Dependent Hardening**: $n > 0.2$ (approx.) (means initially soft, annealed material)

  The cyclic stress-strain curve is significantly above the monotonic curve, which means a gradually increasing deformation resistance as cyclic loading progresses.

  Examples: 2024-T4 Al; 4142 steel (670 BHN)

Note that the hardest steels tend to further harden in cyclic loading. Thus, a given steel (such as 4142) may be stable, softening, or hardening, depending on its initial hardness.

In the technical literature, primes are normally used to denote cyclic material properties. For example, σ'_y is the yield strength obtained from a cyclic stress-strain curve.

Stress vs. Life (S-N) Curves

The most common and historical fatigue life plots present data of stress amplitude (simplistically, S or S_a) on a linear scale vs. cycles to failure (N or N_f) on a logarithmic scale as in Figure 1.6.18.

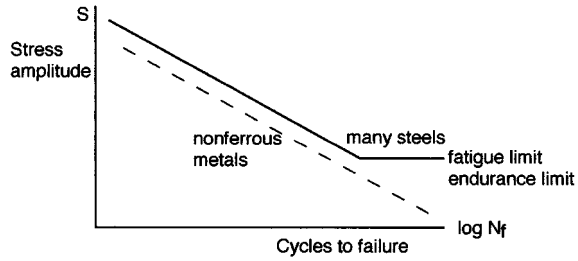


FIGURE 1.6.18 Schematic of S-N curves.

Many steels (plain carbon or low alloy) appear to have a distinct fatigue limit. For other metals that do not have such a limit (aluminum, for example), an arbitrary fatigue limit is defined as a stress amplitude corresponding to a specified life, typically 10^7 or 10^8 cycles.

Trends in S-N Curves

There are many influences on the shape and position of a material's fatigue life curve as briefly discussed below.

Ultimate Strength. It is widely believed that, at least for steels, the fatigue limit σ_e is about one half of the ultimate strength σ_u. In fact, this is a gross oversimplification, with actual values being lower or higher than that in many cases.

Mean Stress, Residual Stress. Several main points are worth remembering: residual stresses (also called self-stresses) are common, and they are to be treated as mean stresses (by sign and magnitude) in fatigue; a tensile mean stress lowers the life while a compressive one increases it. Simplistically, a tensile mean stress lowers the allowable cyclic stress amplitude according to Figure 1.6.19 where

$$\sigma_m + \sigma_a \le \sigma_u \quad \text{or} \quad \sigma_y \qquad (\text{if yielding is to be prevented})$$

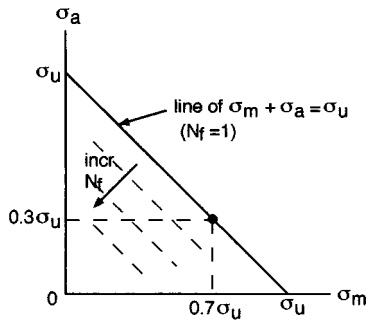


FIGURE 1.6.19 Schematic of tensile mean stress effect.

For example, if $\sigma_m = 0.7\sigma_u$, then the maximum alternating stress for one cycle is $0.3\sigma_u$. This kind of graphical relationship is called a Goodman diagram. There are several special expressions for dealing with the detrimental effects of tensile mean stresses. For example, the modified Goodman equation is

$$\frac{\sigma_a}{\sigma_e} + \frac{\sigma_m}{\sigma_u} = 1 \qquad (1.6.15)$$

where σ_e is the fatigue limit for fully reversed loading.

Sometimes curved lines represent real behavior better than the linear theory shown in Figure 1.6.19. In that case the Gerber parabola may be appropriate, in the form of

$$\frac{\sigma_a}{\sigma_e} + \left(\frac{\sigma_m}{\sigma_u}\right)^2 = 1 \qquad \text{for} \quad \sigma_m \geq 0 \tag{1.6.16}$$

Another approach worth mentioning is the Morrow expression, a mechanistically elegant and sensible one, which will be presented later.

Note that tensile mean stresses are generally detrimental and that many approaches have been proposed to deal with them, yet no single method is capable of good predictions in all cases. In practice it is best to use a particular method that has a good track record for the material and situation at hand. Constant-life diagrams are useful, elaborate derivatives of the Goodman approach, if they include a broad data base (Figure 1.6.20).

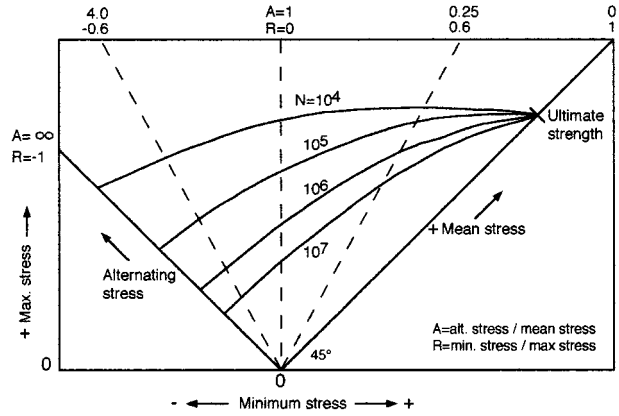


FIGURE 1.6.20 Constant-life diagram.

Notch Effects. Stress raisers can be extremely detrimental in fatigue, except when they help create localized compressive residual stresses in ductile metals, delaying crack formation and growth. These are discussed in connection with the strain-based approach.

Microstructure. Large grain size (annealed metals) lowers the fatigue strength, and small grain size (by cold working) increases it, especially at long lives, under load control.

Surface Effects. The condition of a material's surface may influence the fatigue behavior in many ways, typically in combinations.

Toolmarks are common detrimental features, especially since often they are aligned perpendicular to the principal tensile stress in axial or bending loading. An example is a shaft cut in a lathe. Note that in the case of high-strength, hard materials even invisible scratches from grinding and buffing may be stress raisers. Machining also tends to create tensile or compressive residual stresses in surface layers.

Surface treatments such as carburizing or nitriding of steels affect the fatigue life by changes in chemical composition, microstructure, hardness, or residual stress. Shot peening, surface rolling, or burnishing is done to introduce compressive residual stresses, which delay cracking in long-life service. Plating (chromium, nickel) tends to create layers of poor fatigue resistance and harmful tensile residual stresses. Shot peening after plating is a beneficial step.

Environment. Hostile chemical environments can severely reduce most materials' fatigue resistance. Common causes of problems are salt water, salt in the air, salt on the road, moisture, and even pollutants in the air. For example, sulfur in the air results in aggressive sulfuric acid on machines and structures.

Statistical Scatter. There is always statistical scatter in a material's fatigue life at any given stress level, especially at long lives. The scatter band may cover several orders of magnitude in life at a single stress

level. Because of the scatter, there is no unique fatigue life curve for any material — the curve depends not only on physical factors such as environment, but also on the number of tests done. It is not sufficient to do a handful of tests and draw a curve somewhere through the data points. As a simple rule, to have a high level of confidence (>99%) in a fatigue life curve, at least six identical tests are needed to obtain a mean value at each of several levels of stresses in the general life range of interest. A curve through these mean values is fairly representative of the average life curve (50% probability of failure), but still may not be adequate to deal with scatter. Note that the minimum number of test specimens according to the ASTM Standard E 739 is 6 to 12 for preliminary, exploratory work, or for research and development and component testing, and 12 to 24 for design allowables or reliability assessment.

Ideally, additional analysis is done, using Gaussian (normal) statistical distribution or some other model, such as the Weibull distribution. The latter is particularly informative in determining the probability of fatigue failure. The practical problem is that engineers may require very low probabilities of failure (less than 1%), but neither the necessary mathematical methods nor the data bases are available for that. A family of fatigue life curves for various probabilities of failure and other relevant considerations for one material are shown schematically in Figures 1.6.21 to 1.6.23.

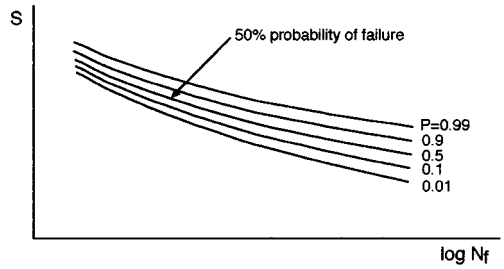


FIGURE 1.6.21 Schematic S-N curves with various probabilities of failure.

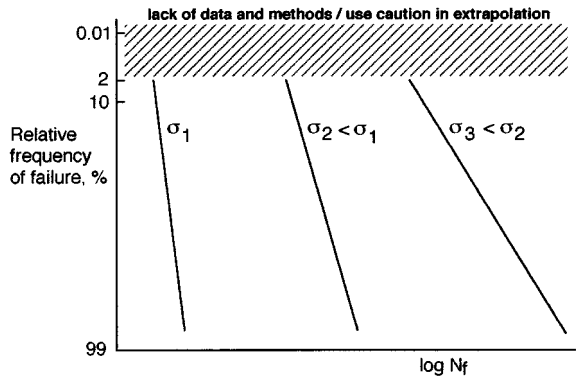


FIGURE 1.6.22 Probability aspects of fatigue depending on stress level.

Variable Amplitude Loading

Many machines, vehicles, and structures experience random or blockwise changing loading. They can be simplistically modeled for life prediction using the Palmgren-Miner rule, illustrated in Figure 1.6.24.

There are two major assumptions for this rule for completely reversed loading:

1. Every cycle at a given level of stress amplitude causes the same amount of damage, whether the cycle is early or late in the life.
2. The percentage of damage caused by a cycle of load at any level of stress is equivalent to the same percentage of damage at any other level of stress.

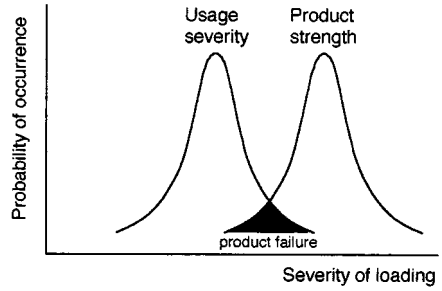


FIGURE 1.6.23 Probability aspects of fatigue depending on applied stress and product strength.

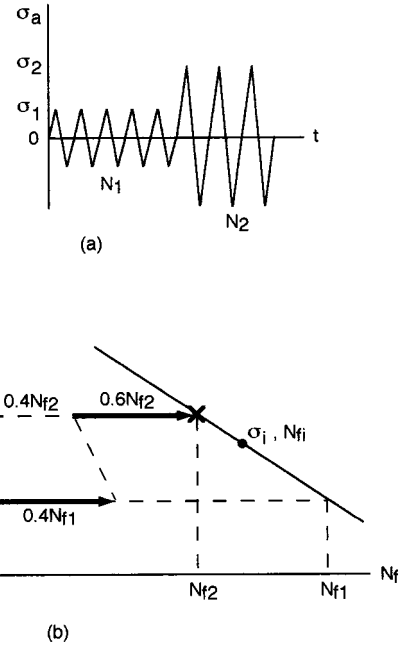


FIGURE 1.6.24 Schematic for Palmgren-Miner rule.

Thus, since 100% of the life N_{fi} is exhausted at failure at any single stress amplitude σ_i, in multilevel loading the life fractions sum to unity, as mathematically formulated here and illustrated in Figure 1.6.24,

$$\frac{N_1}{N_{f1}} + \frac{N_2}{N_{f2}} + \ldots = \sum \frac{N_i}{N_{fi}} = 1 \tag{1.6.17}$$

where N_i is the actual number of cycles at σ_i and N_{fi} is the life at σ_i.

In practice, summations of about 0.8 to 1.2 can be accepted, saying that the Palmgren-Miner rule is valid in that case. Gross deviations from summations of 1 are common, especially when the mean stress is not zero. There are modified versions of the basic rule for such cases, but they should be used with caution.

Cycle Counting. Highly irregular loading requires the use of special cycle counting methods, such as level crossing, range counting, or rainflow cycle counting. The latter is the best modern method, lending itself to efficient field data acquisition and computer work (ASTM Standard E1049; *SAE Fatigue Design Handbook*).

Multiaxial Fatigue

Complex states of stress are common in engineering components, and in fatigue analysis they may cause serious difficulties. There are many methods available, but none of them are adequate for all cases. The simplest situations that might be handled reasonably well involve fully reversed loading by in-phase or 180° out-of-phase proportional stresses at the same frequency. Multiaxial fatigue testing is difficult and expensive, so it is often desired to use uniaxial test data for predicting the multiaxial behavior. A typical approach for this is based on computing an effective stress amplitude σ_e from the amplitudes of the principal stresses σ_{1a}, σ_{2a}, σ_{3a}. With the concept of the octahedral shear yield criterion,

$$\sigma_e = \frac{1}{\sqrt{2}}\sqrt{\left(\sigma_{1a}-\sigma_{2a}\right)^2+\left(\sigma_{2a}-\sigma_{3a}\right)^2+\left(\sigma_{3a}-\sigma_{1a}\right)^2} \tag{1.6.18}$$

where in-phase stresses are positive and 180° out-of-phase stresses are negative.

The life is estimated by entering σ_e on the appropriate S-N curve. Note that mean stresses, localized or general yielding, creep, and random frequencies of loading further complicate the problem and require more sophisticated methods than outlined here.

Strain vs. Life (ε-N) Curves

A strain-based approach is necessary in fatigue when measurable inelastic strains occur. In general, total strain consists of elastic, plastic, and creep strains, with the latter two being in the category of inelastic strains,

$$\varepsilon_t = \varepsilon_e + \varepsilon_p + \varepsilon_c \tag{1.6.19}$$

When ε_p or/and ε_c are dominant, the life is relatively short and the situation is called low-cycle fatigue (LCF), as opposed to high-cycle fatigue (HCF), where ε_e is dominant. The mechanics of LCF can be understood by first considering hysteresis loops of elastic and plastic strains as defined in Figure 1.6.25.

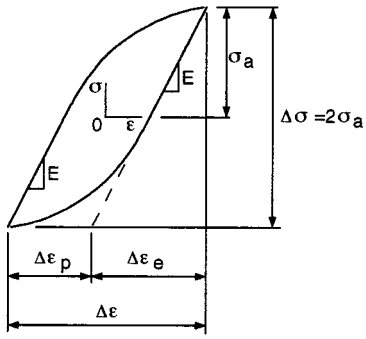


FIGURE 1.6.25 Hysteresis loop.

Simplistically, HCF means a thin loop (a straight line at very long life) and LCF means a fat loop. Strain-life plots are especially useful in the LCF regime where material properties (ε_f, σ_f) obtained in monotonic tension tests are directly useful in fatigue life prediction as shown in Figure 1.6.26. Most commonly the total strain amplitude ε_a is plotted vs. the life $2N_f$, with a corresponding equation (called Coffin-Manson equation) for fully reversed loading,

$$\varepsilon_a = \frac{\sigma_f}{E}\left(2N_f\right)^b + \varepsilon_f\left(2N_f\right)^c \tag{1.6.20}$$

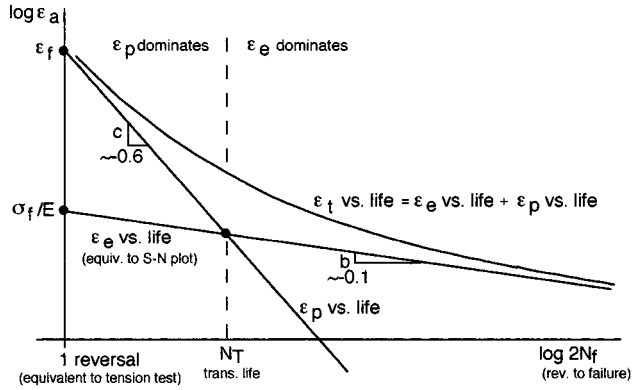


FIGURE 1.6.26 Schematic of strain vs. life curves.

It is remarkable that all metals are similar to one another in their values of the exponents b (≈ -0.1) and c (≈ -0.6), differing only in fracture strength σ_f and fracture ductility ε_f. These allow a simplistic fatigue life prediction if at least σ_f and ε_f are known.

If there is a mean stress, its effect is equivalent to an altered fracture strength. Using the Morrow approach in a simplified version,

$$\varepsilon_a = \frac{\sigma_f}{E}\left(1 - \frac{\sigma_m}{\sigma_f}\right)\left(2N_f\right)^b + \varepsilon_f\left(2N_f\right)^c \tag{1.6.21}$$

where σ_m is positive for tensile and negative for compressive mean stress.

Notch Effects

The localized plastic strains of notched members complicate the fatigue analysis considerably. It should be noted, first of all, that the theoretical stress concentration factor K_t is not entirely relevant to such members, because yielding lowers the actual peak stresses from those predicted. This leads to the definitions of the true stress and strain concentration factors,

$$K_\sigma = \frac{\text{peak stress}}{\text{ave. stress}} \qquad K_\varepsilon = \frac{\text{peak strain}}{\text{ave. strain}} \tag{1.6.22}$$

According to Neuber's rule,

$$K_t = \sqrt{K_\sigma K_\varepsilon} \tag{1.6.23}$$

which is useful for notch analysis in fatigue. This expression is strictly true for ideally elastic behavior and is qualitatively evident for elastic-plastic deformations.

Residual Stresses at Notches. An extremely important, and somewhat surprising, phenomenon can occur in notched members if they yield locally under variable-amplitude loading. If a large load (called an overload) causes yielding at a notch and is followed only by smaller loads, a residual stress of the opposite sign to the overload's sign is generated at the root of the notch. Thus, a tensile overload (such as at one side of a shaft in a straightening operation) creates a compressive residual stress, and vice versa. These stresses may remain in the member for a long time or be relaxed by other plastic strain events or by annealing. Of course, such stresses are effective mean stresses and can alter the life greatly.

Creep-Fatigue Interactions

Inelastic strains (plastic and creep strains) are the basic causes of time- and cycle-dependent damage processes. When both kinds of strains occur during the life of a particular component, complex damage interactions may arise. The simplest and most elegant approach in such a case is to sum both of the different damages linearly (as in the Palmgren-Miner summation for pure fatigue), assuming that they are equivalent to one another. In other words, assume that X percentage of creep life exhausted is equivalent to the same X percentage of fatigue life exhausted. Thus, a linear expression involving time and cycle fractions can be stated,

$$\underset{\text{pure creep}}{\sum \frac{t_i}{t_{ri}}} + \underset{\text{pure fatigue}}{\sum \frac{n_j}{N_{fj}}} = \underset{\text{at failure}}{1} \tag{1.6.24}$$

where t_i = actual time spent at stress level i in creep, t_{ri} = total time to rupture at stress level i, n_j = actual number of cycles at stress level j, and N_{fj} = cycles to failure at stress level j.

This idealized linear expression is plotted as a dashed line in Figure 1.6.27; in contrast, a more realistic ASME code and possible severe degradations are also plotted.

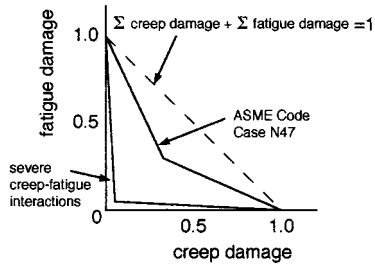


FIGURE 1.6.27 Schematic of creep-fatigue interactions. The bilinear damage rule is recommended in the ASME Boiler and Pressure Vessel Code, Section III, Code Case N47.

There are many other methods (such as damage rate equations; strain-range partitioning) to deal with creep-fatigue problems, but none of them are adequate for all situations. The difficulty is mainly because of the need to account for important, complex details of the loading cycle (frequency, hold times, temperature, and deformation wave shape).

Fracture Mechanics Method in Fatigue

Cyclic loading can cause crack growth with or without the presence of a hostile chemical environment. The rate of crack growth depends on the stress intensity factor $K \propto \sigma\sqrt{a}$. Investigations of this dependence have led to the development of powerful techniques in design and failure analysis. The fatigue crack growth behavior is quantified by the Paris equation,

$$\frac{da}{dN} = C(\Delta K)^m \tag{1.6.25}$$

where da/dN = crack growth rate
C, m = material constants
$\Delta K = K_{max} - K_{min}$ = stress intensity factor range
$K_{max} \propto \sigma_{max}$
$K_{min} \propto \sigma_{min}$

Typical data for a wide range of crack growth rates have patterns as in Figure 1.6.28, where ΔK_{th} is a threshold value akin to a fatigue limit. The linear part of the curve is useful for life prediction and failure analysis.

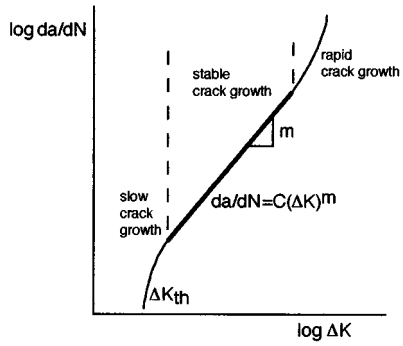


FIGURE 1.6.28 Schematic of fatigue crack propagation data.

Abridged Example of a Modern Fatigue Analysis

Many of the concepts mentioned above are applied in Sandia National Laboratories' "User's Manual for FAROW: Fatigue and Reliability of Wind Turbine Components," SAND94-2460, November 1994. FAROW is a computer program for the probabilistic analysis of large wind turbines, using structural reliability techniques to calculate the mean time to failure, probability of failure before a target lifetime, relative importance of each of the random inputs, and the sensitivity of the reliability to all input parameters. The method is useful whether extensive data are available or not (showing how much can be gained by reducing the uncertainty in each input). It helps one understand the fatigue reliability of a component and indicates how to improve the reliability. The sample figures (Figures 1.6.29 to 1.6.32) illustrate some of the key data and results for the machines and materials considered.

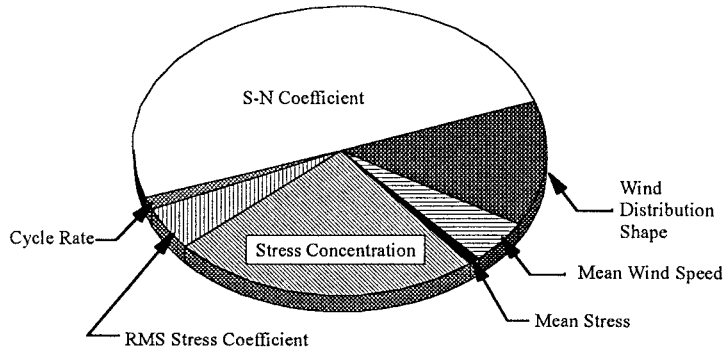


FIGURE 1.6.29 Relative importance factors as fractions of the total influence on the probability of failure. (Courtesy Sandia National Laboratories, Albuquerque, NM.)

Note especially a large discrepancy between mean lifetime and probability of failure in a few years. A mean lifetime of 600 years was calculated for a critical component, using the median values for all the random variables considered and using the constant values for all the other input parameters. However, the probability of the component failing in less than 5 years is estimated at 7.6% (Figure 1.6.32). This shows the uncertainty even in sophisticated fatigue life calculations because of reasonable uncertainty in the inputs and the sensitivity of fatigue life to parameter variation.

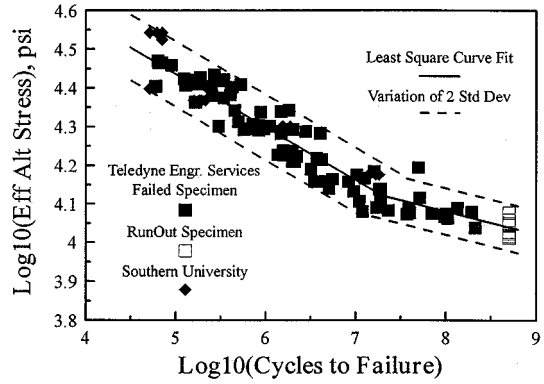


FIGURE 1.6.30 Fatigue life data for 6063 Al. (Courtesy Sandia National Laboratories, Albuquerque, NM.)

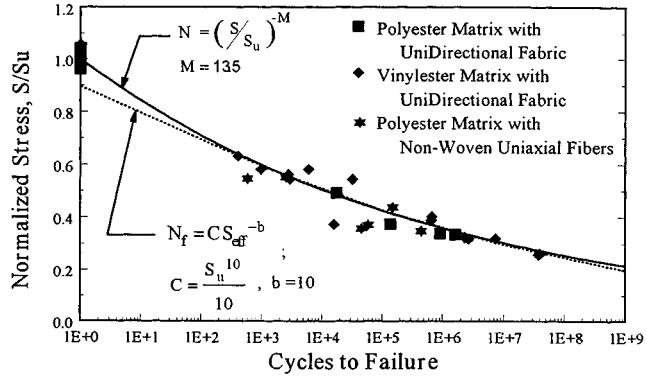


FIGURE 1.6.31 Fatigue life data for uniaxial fiberglass composite. (Courtesy Sandia National Laboratories, Albuquerque, NM.)

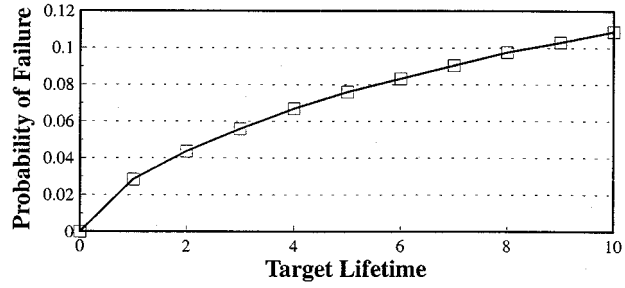


FIGURE 1.6.32 Example FAROW results for probability of premature failure as a function of target lifetime. (Courtesy Sandia National Laboratories, Albuquerque, NM.)

1.7 Comprehensive Example of Using Mechanics of Solids Methods

Richard C. Duveneck, David A. Jahnke, Christopher J. Watson, and Bela I. Sandor

A concise overview of an engineering project is presented to illustrate the relevance and coordinated application of several concepts and methods in this chapter. The sketchy outline is limited in breadth and depth, emphasizes modern methods, and is not aiming for completeness in any particular area.

The Project

Analyze the currently used A-shaped arm of the suspension system of a small, special-purpose ground vehicle. The goal is to redesign the component to save weight and, more importantly, reduce the cost of manufacturing, while assuring the product's reliability over its expected service life.

Concepts and Methods

Statics

Vectors
Free-body diagrams. Equilibrium
Two-force member: shock absorber
Frame components
Beams. Bending moments
Moments of inertia
Center of mass

Dynamics

Velocity, acceleration
Rigid-body dynamics
General plane motion
Relative motion

Vibrations

Natural frequency
Damping. Logarithmic decrement

Mechanics of Materials

Stress and strain. Transformation equations. Principal stresses. Maximum shear stress
Material properties. Material selection
Bending stresses. Beam optimization
Strain gages. Mechanical testing with closed-loop equipment

Durability

Stress concentrations. Finite element analysis
Cumulative fatigue damage. Cycle counting in random loading. Mean stresses. Goodman diagrams. Life prediction
Thermoelastic stress analysis

Illustrations

A few aspects of the project are graphically illustrated in Color Plate 16 and Figures 1.7.1 to 1.7.3.

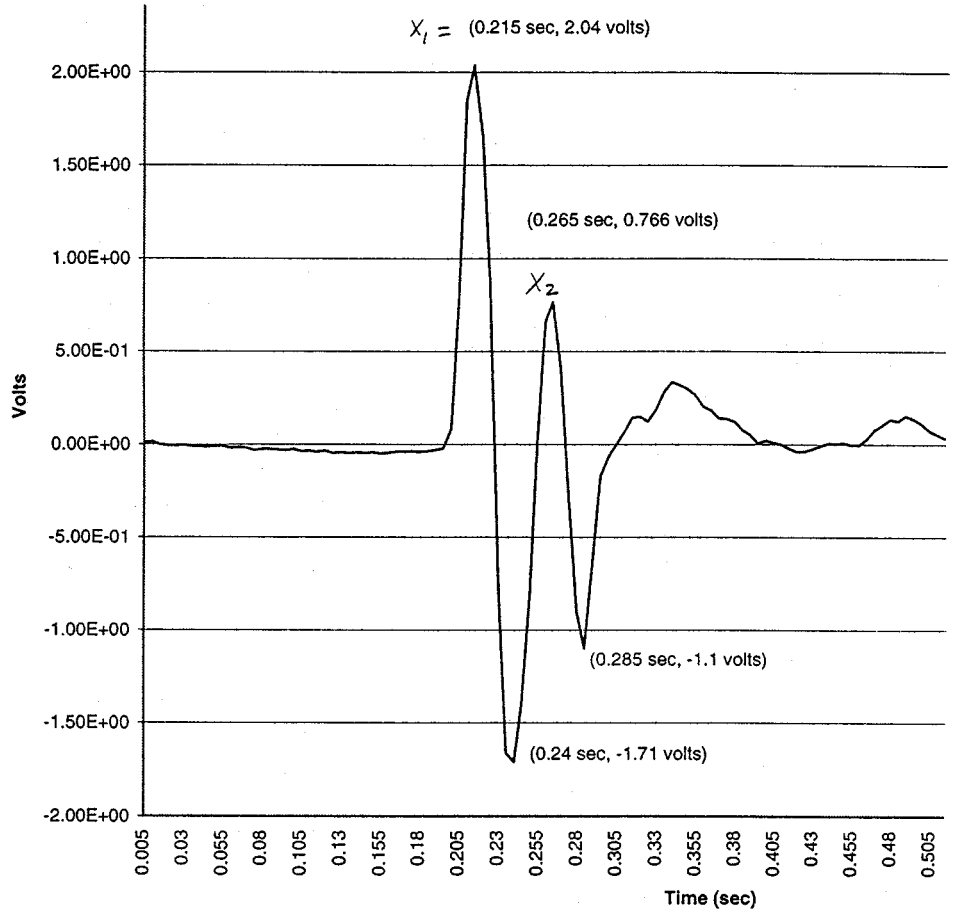


FIGURE 1.7.1 Accelerometer data from front suspension system of vehicle. Logarithmic decrement $\partial = \ln(x_1/x_2)$; damping ratio $\zeta = 0.16$.

Defining Terms

STATICS

Equilibrium: A concept used to determine unknown forces and moments. A rigid body is in equilibrium when the equivalent force-couple system of the external forces acting on it is zero. The general conditions of equilibrium are expressed in vector form ($\sum\mathbf{F} = 0$, $\sum\mathbf{M}_O = \sum[\mathbf{r} \times \mathbf{F}] = 0$) or scalar form ($\sum F_x = 0$, $\sum F_y = 0$, $\sum F_z = 0$, $\sum M_x = 0$, $\sum M_y = 0$, $\sum M_z = 0$).

Equivalent force-couple system: Any system of forces and moments acting on a rigid body can be reduced to a resultant force and a resultant moment. Transformations of a force-couple system involving chosen points of reference are easy to make. These are useful for determining unknown forces and moments and the critical locations in structural members.

Free-body diagram: A method of modeling and simplifying a problem for the efficient use of the equilibrium equations to determine unknown forces and moments. A body or group of bodies is imagined to be isolated from all other bodies, and all significant external forces and moments (known or unknown) are shown to act on the free-body model.

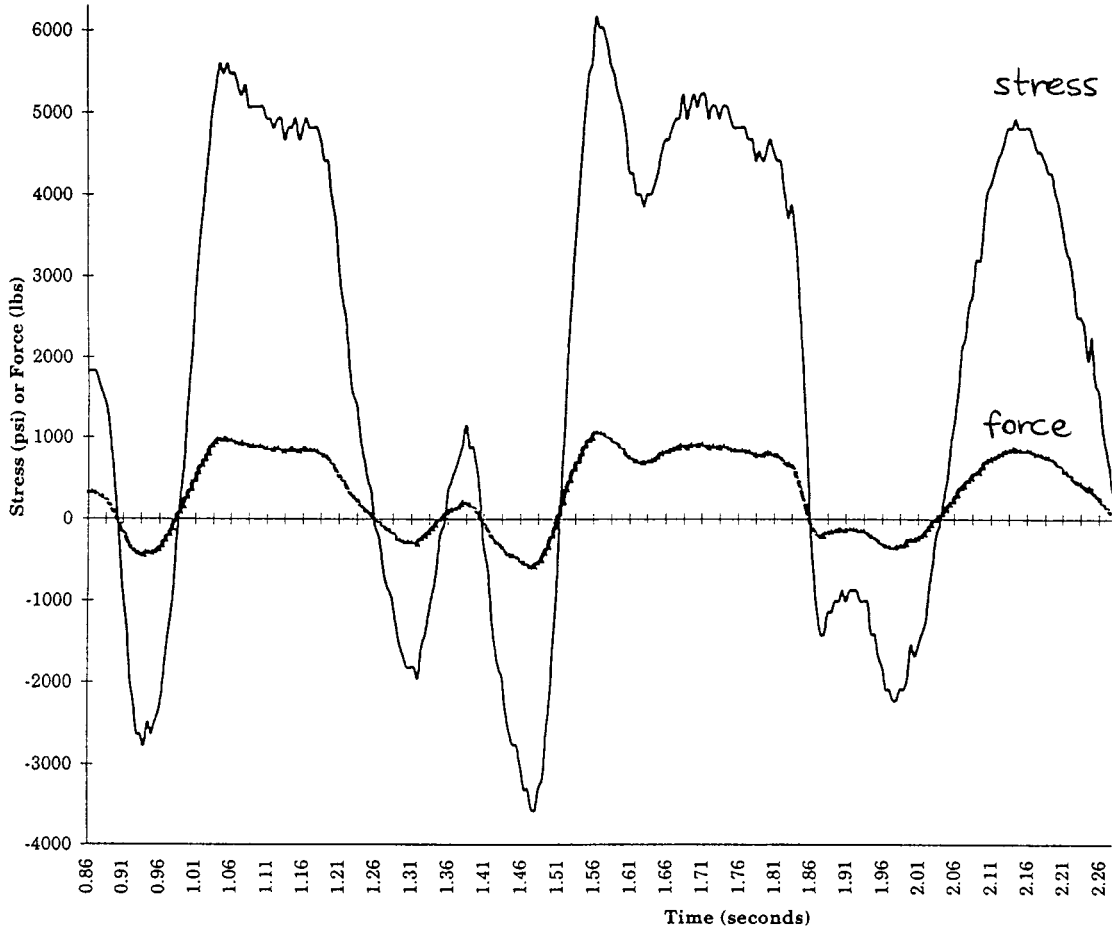


FIGURE 1.7.2 Axial stress and force vs. time in shock absorber shaft.

DYNAMICS

Equations of motion: Expressions of the acceleration of a body related to the forces acting on the body. The basic equation of motion for a particle of mass m is $\sum \mathbf{F} = m\mathbf{a}$. Many other equations of motion may be stated, depending on the dimensions of the body and its motion (such as two- or three-dimensional motion) and the coordinate system chosen.

Kinematics: The analysis of motion based on geometry and time-dependent aspects. Forces may or may not be associated with the motion, but the analysis does not involve considerations of forces. The parameters of interest in kinematics are position, displacement, velocity, acceleration, and time.

Kinetics: The analysis of motion based on kinematics and the effects of forces on masses.

VIBRATIONS

Forced vibration: Involves an exciting force applied periodically during the motion. A forced vibration may also be described in terms of the displacement of a foundation or primary mass that supports the vibrating system.

Free vibration: Occurs when only two kinds of forces are acting on a mass: (a) the elastic restoring force within the system and (b) the force of gravity or other constant forces that cause no displacement from the equilibrium configuration of the system.

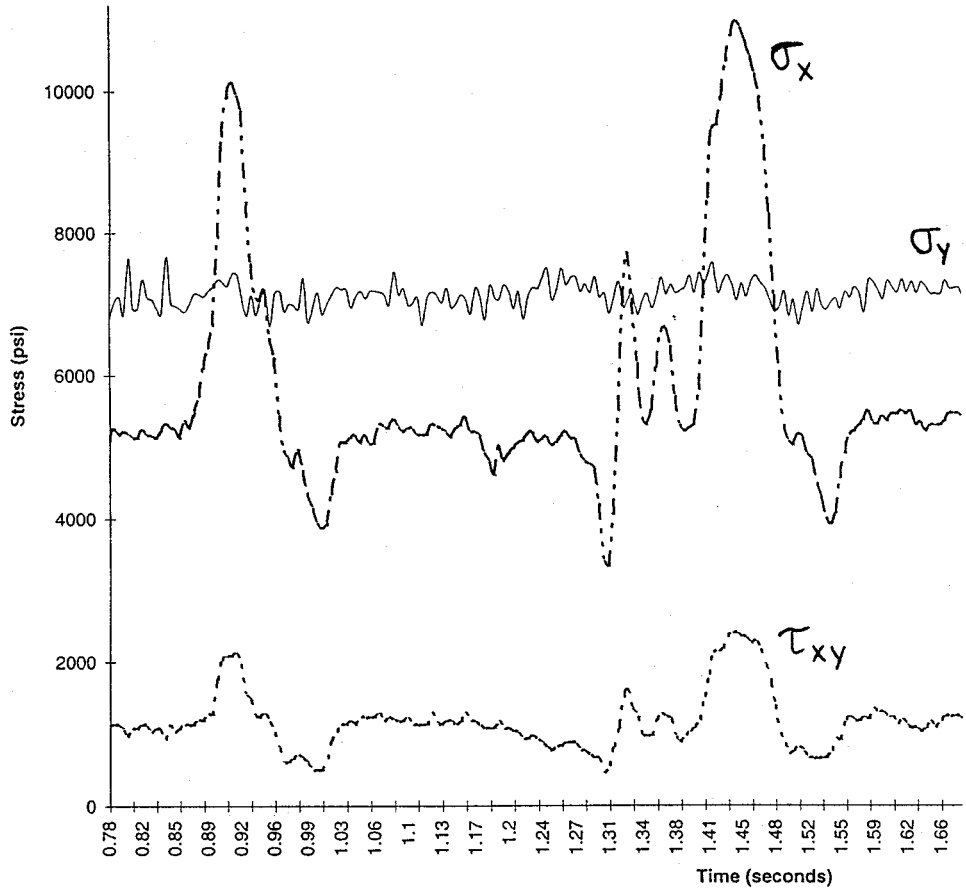


FIGURE 1.7.3 Stresses σ_x, σ_y, and τ_{xy} measured at one point of the A-arm by strain gages as the vehicle travels over bumps.

Resonance: A critical aspect of forced vibrations; it occurs when the forcing frequency equals the system's natural frequency. In this condition the amplitude of the displacements becomes infinite in theory, or dangerously large in practice when the damping is small. Near-resonance conditions may also be undesirable.

MECHANICS OF MATERIALS

Flexure formula: Used to calculate the bending stresses in beams. Must be applied with modifications if there are inelastic deformations, unsymmetric bending, or for composite beams and curved beams.

Hooke's law: Applicable for calculating uniaxial or multiaxial stress-strain responses when the material acts entirely elastically. Involves the modulus of elasticity E and Poisson's ratio v.

Principal stresses: The maximum and minimum normal stresses at a point, on an infinitesimal element. An important related quantity is the absolute maximum shear stress. These quantities can be determined (given an arbitrary state of applied stress) from stress transformation equations or from their graphical solution, Mohr's circle. Principal strains are determined in a similar way.

Stress-strain diagram: Shows the stress-strain response and many important mechanical properties for a material. These properties depend greatly on the material's chemical composition and several other factors of fabrication and service conditions. Monotonic (tension or compression) and cyclic loading conditions may result in grossly different mechanical behaviors even for a given material.

STRUCTURAL INTEGRITY AND DURABILITY

Rate of crack growth: A measure of damage evolution and remaining life of a member. In fatigue, the crack propagation rate *da/dN* depends on the stress intensity factor range ΔK and material properties. This relationship is the basis of the powerful, well-established damage-tolerant design method.

Stress concentration factor: The localized stress-raising effect of a geometric discontinuity. There are many, potentially confusing, forms of quantifying this effect. The most prominent factors are distinguished concisely:

a. Theoretical stress concentration factor, $K_t = \sigma_{max}/\sigma_{ave}$
   Depends on geometry of notch, not on material
   Has no engineering units
b. True stress concentration factor, $K_\sigma = \sigma_{max}/\sigma_{ave}$
   Depends on geometry of notch and material; $K_\sigma = K_t$ for perfectly elastic material, $K_\sigma < K_t$ for ductile material
   Has no engineering units
c. True strain concentration factor, $K_\varepsilon = \varepsilon_{max}/\varepsilon_{ave}$, $\varepsilon_{ave} = \sigma_{ave}/E$
   Depends on geometry of notch and material; $K_\varepsilon = K_t$ for perfectly elastic material, $K_\varepsilon > K_t$ for ductile material
   Has no engineering units

Stress intensity factor: A measure of the severity of a crack, or the intensity of the stress field near the crack tip. There are many, potentially confusing, forms of this factor, having identical engineering units of stress $\sqrt{\text{length}}$, but a variety of definitions and applications. A few are listed concisely:

a. Opening-mode stress intensity factor, K_I
   Depends on geometry of a crack and applied stress, not on material
   Units of stress $\sqrt{\text{length}}$
b. Plane strain fracture toughness, K_{IC}
   Depends on material but not on geometry above a certain thickness, and not on applied stress
   Units of stress $\sqrt{\text{length}}$
c. Stress intensity factor range, $\Delta K = K_{max} - K_{min}$
   Depends on geometry of a crack and applied cyclic stress, not on material
   Units of stress $\sqrt{\text{length}}$

References

STATICS AND DYNAMICS

Hibbeler, R. C. 1995. *Engineering Mechanics: Statics and Dynamics*, 7th ed. Prentice-Hall, Englewood Cliffs, NJ.

Sandor, B. I. 1987. *Engineering Mechanics Statics and Dynamics*, 2nd ed. Prentice-Hall, Englewood Cliffs, NJ.

VIBRATIONS

Harris, C. M. and Crede, C. E. 1988. *Shock and Vibration Handbook*, 3rd ed. McGraw-Hill, New York.

James, M. L., Smith, G. M., Wolford, J. C., and Whaley, P. W. 1994. *Vibration of Mechanical and Structural Systems*, 2nd ed. Harper Collins College Publishers, New York.

Wowk, V. 1991. *Machinery Vibration: Measurement and Analysis*. McGraw-Hill, New York.

Wowk, V. 1993. *Machinery Vibration: Balancing*. McGraw-Hill, New York.

MECHANICS OF MATERIALS

Dally, J. W. and Riley, W. F. 1991. *Experimental Stress Analysis*, 3rd ed. McGraw-Hill, New York.

Hibbeler, R. C. 1997. *Mechanics of Materials*, 3rd ed. Prentice-Hall, Englewood Cliffs, NJ.

Jawad, M. H. and Farr, J. R. 1989. *Structural Analysis and Design of Process Equipment*, 2nd ed. John Wiley & Sons, New York.

Kobayashi, A. S. (ed). 1993. *Handbook on Experimental Mechanics*, 2nd ed. Society for Experimental Mechanics, Bethel, CT.

Young, W. C. 1989. *Roark's Formulas for Stress and Strain*, 6th ed. McGraw-Hill, New York.

STRUCTURAL INTEGRITY AND DURABILITY

Anderson, T. L. 1994. *Fracture Mechanics: Fundamentals and Applications*, 2nd ed., CRC Press, Boca Raton, FL.

Boyer, J. E. 1986. *Atlas of Fatigue Curves*. American Society for Metals, Metals Park, OH.

Cook, R. D. 1995. *Finite Element Modeling for Stress Analysis*. John Wiley & Sons, New York.

Dowling, N. E. 1993. *Mechanical Behavior of Materials*. Prentice-Hall, Englewood Cliffs, NJ.

Fuchs, H. O. and Stephens, R. I. 1980. *Metal Fatigue in Engineering*. John Wiley & Sons, New York.

Gallagher, J. P. (ed). 1983. *Damage Tolerant Design Handbook*, 4 vols. Metals and Ceramics Information Ctr., Battelle Columbus Labs, Columbus, OH.

Murakami, Y. (ed). 1987. *Stress Intensity Factors Handbook*, 2 vols. Pergamon Press, Oxford, U.K.

Rice, R. C. (ed). 1988. *Fatigue Design Handbook*, 2nd ed. SAE Publ. No. AE-10. Society of Automotive Engineers, Warrendale, PA.

Further Information

Many technical societies are active in various areas of mechanics of solids, and they are excellent, steady sources of long-accepted and new information, some of which is available within hours. They also organize committee work, conferences, symposia, short courses, and workshops; establish codes and standards; and publish books, papers, journals, and proceedings, covering the latest developments in numerous specialties. A short list of societies is given here; note that they tend to have international breadth, regardless of the name. It is wise to belong to several relevant societies and at least scan their announcements.

ASM International (formerly American Society for Metals) (800-336-5152)
ASME — American Society for Mechanical Engineers (800-843-2763)
ASNT — American Society for Nondestructive Testing (800-222-2768)
ASTM — American Society for Testing and Materials (215-299-5585)
SAE — Society of Automotive Engineers (412-776-4841)
SEM — Society for Experimental Mechanics (203-790-6373)
SES — Standards Engineering Society (513-223-2410)

As a hint of the scope and magnitude of what is available from the large technical societies, here are selected offerings of ASTM:

- ASTM Staff Access/Tel: 215-299-5585; Fax: 215-977-9679; E-mail: infoctr@local.astm.org
- *ASTM Standardization News*, a monthly magazine; regularly presents information on "the development of voluntary full consensus standards for materials, products, systems and services and the promotion of related knowledge… the research, testing and new activities of the ASTM standards-writing committees… the legal, governmental and international events impacting on the standards development process" (quotes from the masthead).
- Over 50 volumes of ASTM Standards

  Samples of standards:

  Friction, wear, and abrasion (B611 on wear resistance of carbides; G77 on ranking of materials in sliding wear)

  Fracture mechanics (E399 on fracture toughness testing of metals)

  Fatigue (E466 on axial fatigue tests of metals; D671 on flexural fatigue of plastics)
- Training courses for ASTM Standards (215-299-5480)

- ASTM International Directory of Testing Laboratories
- ASTM Directory of Scientific & Technical Consultants & Expert Witnesses
- ASTM Special Technical Publications (STP) are books of peer-reviewed papers on recent research and developments

  Samples of STPs:

  STP 1198 — *Nondestructive Testing of Pavements and Backcalculation of Moduli*, Second Volume; 1995

  STP 1231 — *Automation in Fatigue and Fracture: Testing and Analysis*; 1995.

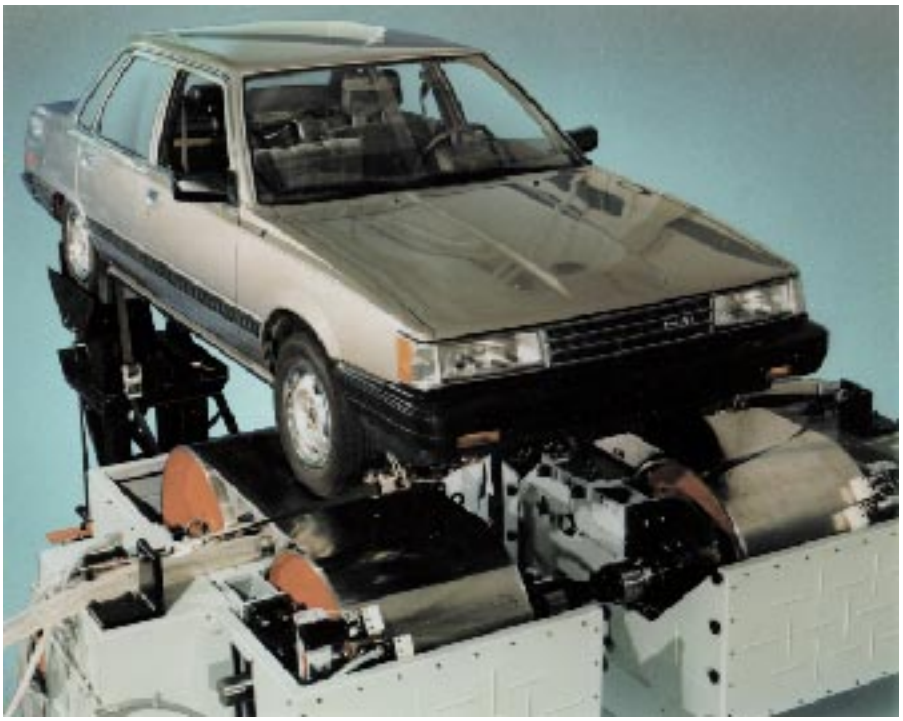


PLATE 1 Flat-Trac® Roadway Simulator, R&D 100 Award-winning system in 1993. (Photo courtesy MTS Systems Corp., Minneapolis, MN.)

PLATE 2 Spinning torque transducer with on-board preamplifier. An angular accelerometer is attached at the center of the torque cell. (Photo courtesy MTS Systems Corp., Minneapolis, MN.)

PLATE 3 Vibration screening of a circuit board using an electromagnetic shaker and a laser doppler vibration pattern imager. (Photo courtesy Ometron Inc., Sterling, VA.)

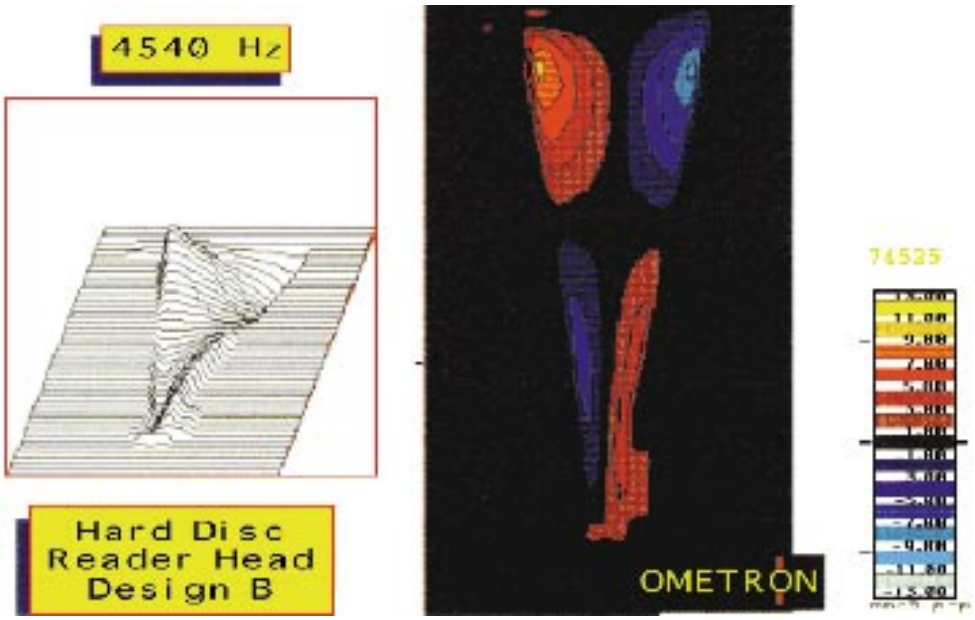


PLATE 4 Vibration patterns of a computer hard disc reader head at 4540 Hz. (Photo courtesy Ometron Inc., Sterling, VA.)

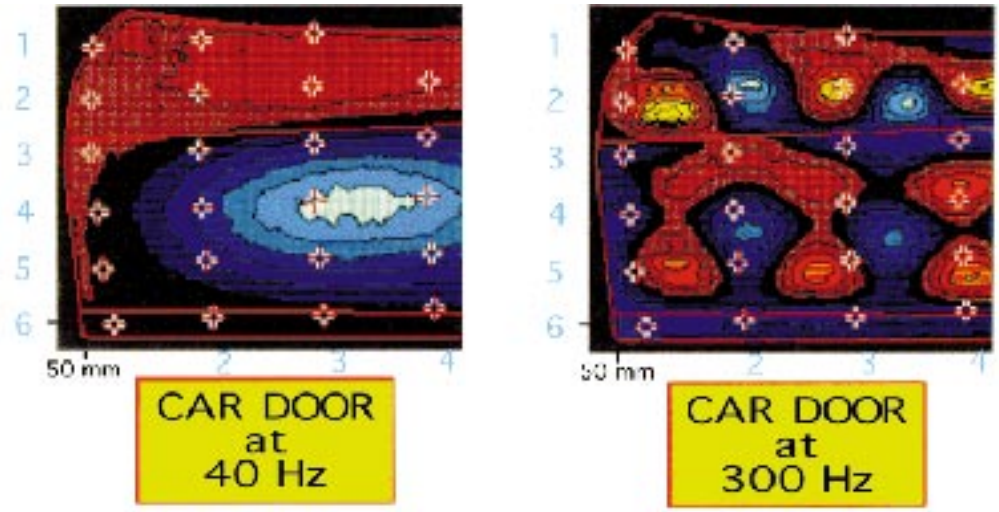


PLATE 5 Vibration patterns of a car door at 40 Hz and 300 Hz. (Photos courtesy Ometron Inc., Sterling, VA.)

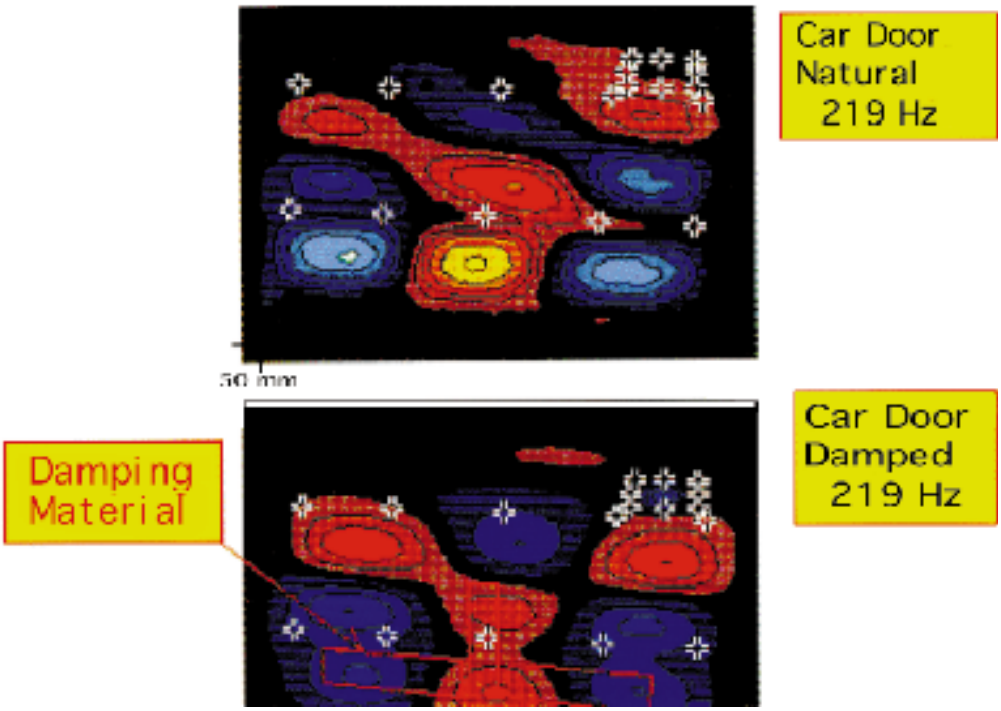


PLATE 6 Changes in the vibration patterns of a car door caused by the addition of damping material. (Photos courtesy Ometron Inc., Sterling, VA.)

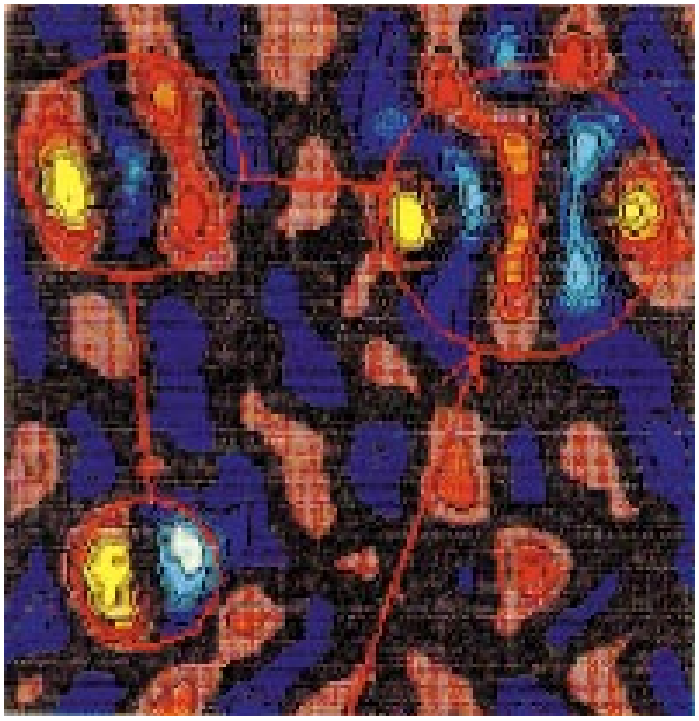


PLATE 7 Detection of delaminations in a foam-and-steel composite plate using vibration pattern imaging. (Photo courtesy Ometron Inc., Sterling, VA.)

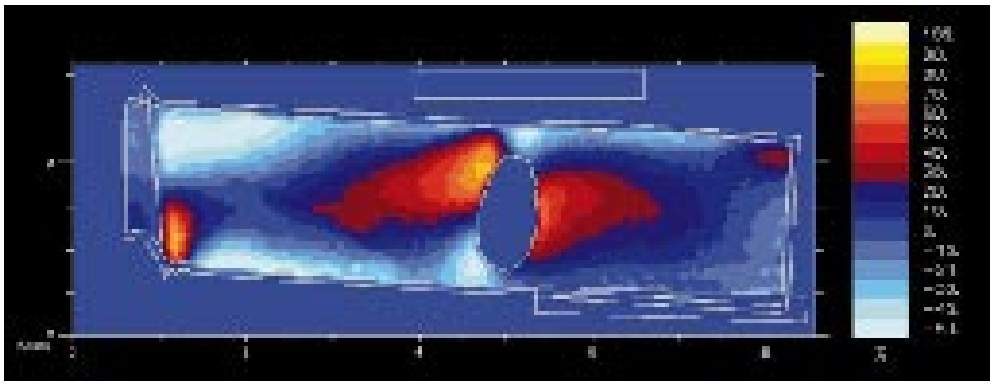


PLATE 8 Modal analysis of a vibrating turbine blade using Thermoelastic Stress Analysis. (Photo courtesy Stress Photonics Inc., Madison, WI.)

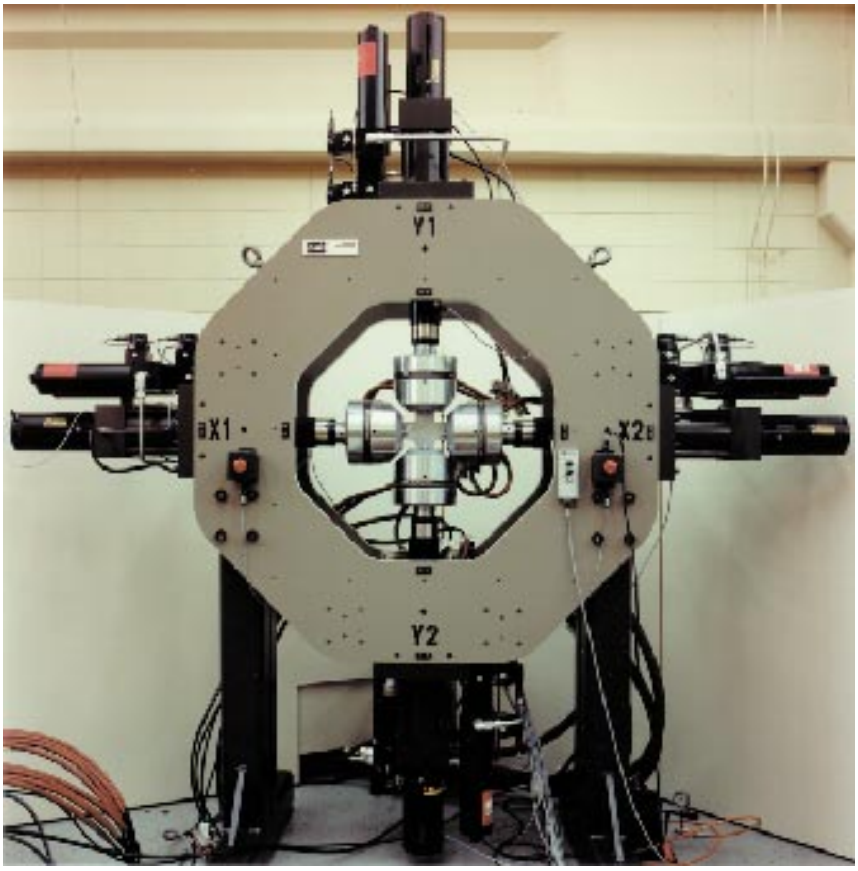


PLATE 9 Biaxial test setup. (Photo courtesy MTS Systems Corp., Minneapolis, MN.)

PLATE 10 Pressure vessel. (Photo courtesy Nooter Corp., St. Louis, MO.)

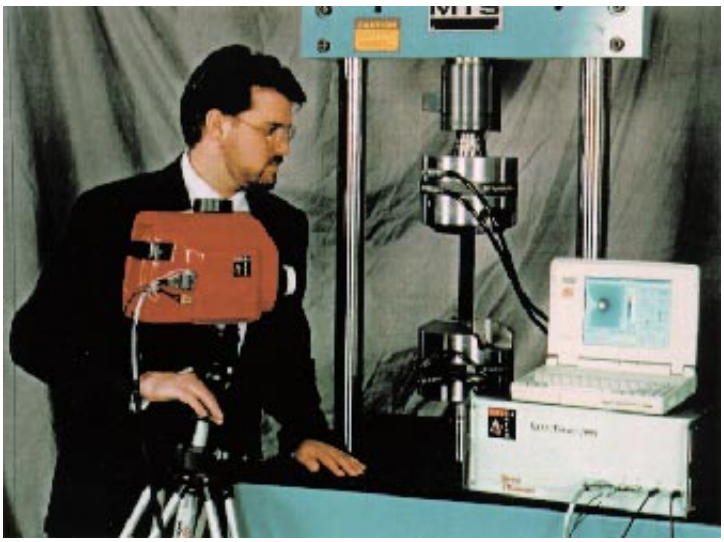


PLATE 11 Delta Therm 1000 Stress Imaging System with principal inventor Jon R. Lesniak. R&D 100 Award-winning instrument in 1994. (Photo courtesy Stress Photonics Inc., Madison, WI.)

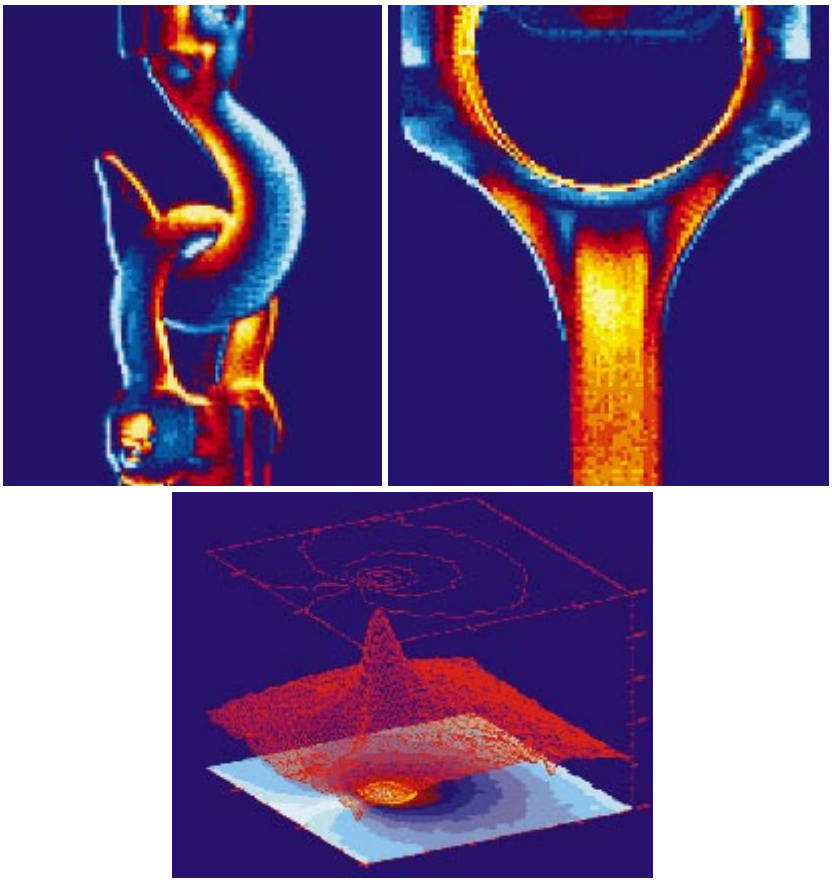


PLATE 12 TSA stress images and samples of data processing by Delta Therm instrument (Color Plate 11). (Photo courtesy Stress Photonics Inc., Madison, WI.)

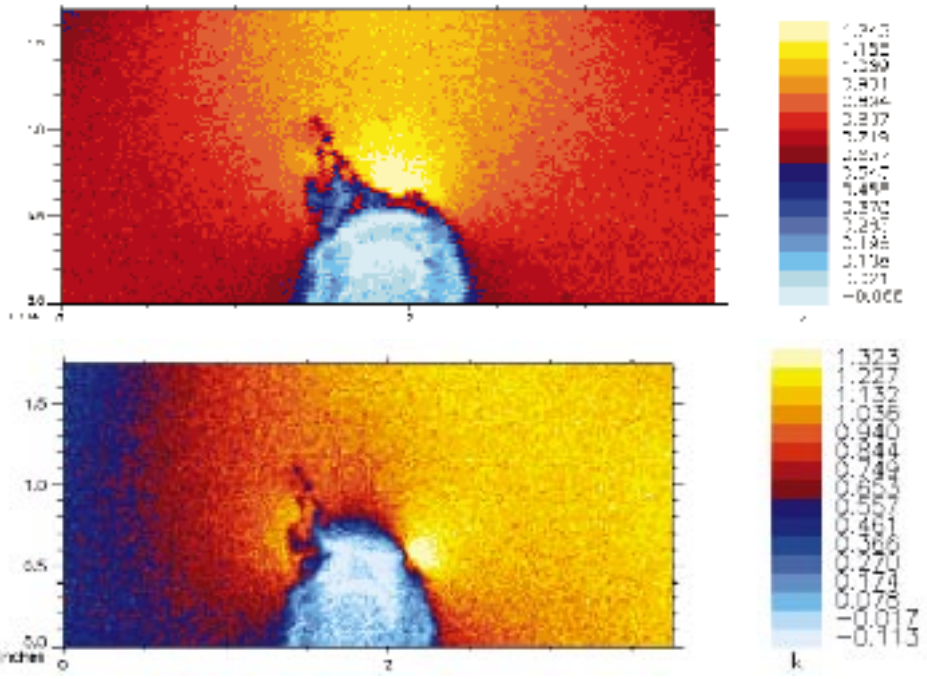


PLATE 13 TSA stress images showing damage evolution at a weld. Top: beginning of fatigue testing; yellow shows stress concentration at weld toe (no crack); dark blue spots represent lower stress at weld splatter. Bottom: gross and uneven stress redistribution to tips of crack (≈ 0.5 in. long) after 1 million cycles. (Photos courtesy Mark J. Fleming, University of Wisconsin-Madison.)

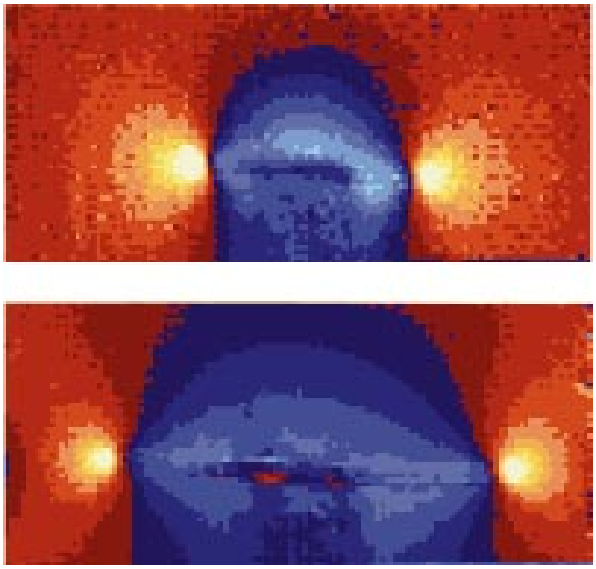


PLATE 14 Direct measurement of crack length and stress intensity factors by TSA stress imaging. Top: crack at 41,000 cycles. Bottom: crack at 94,000 cycles; light shows through the crack; blues show stress relief at crack faces and nearby. (Photos courtesy Mark J. Fleming, University of Wisconsin-Madison.)

PLATE 15 Closed-loop, electro-hydraulic mechanical testing systems. (Photo courtesy MTS Systems Corp., Minneapolis, MN.)

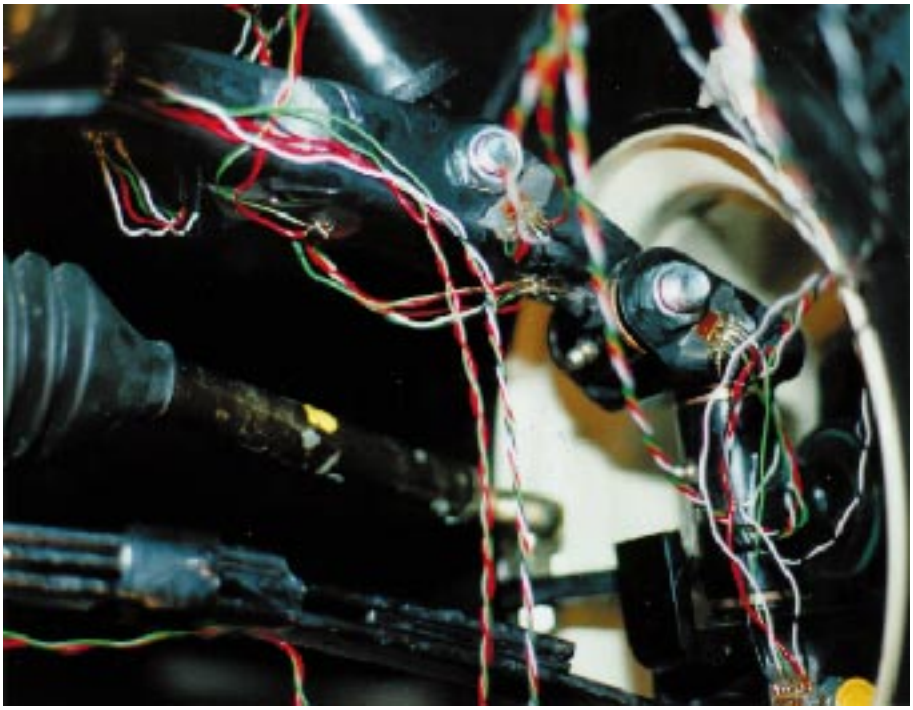


PLATE 16 Strain-gauging of a vehicle's suspension system in progress.

Moran, M.J. "Engineering Thermodynamics"
Mechanical Engineering Handbook
Ed. Frank Kreith
Boca Raton: CRC Press LLC, 1999

Engineering Thermodynamics

Michael J. Moran
Department of Mechanical Engineering
The Ohio State University

Although various aspects of what is now known as thermodynamics have been of interest since antiquity, formal study began only in the early 19th century through consideration of the motive power of *heat*: the capacity of hot bodies to produce *work*. Today the scope is larger, dealing generally with *energy and entropy*, and with relationships among the *properties* of matter. Moreover, in the past 25 years engineering thermodynamics has undergone a revolution, both in terms of the presentation of fundamentals and in the manner that it is applied. In particular, the second law of thermodynamics has emerged as an effective tool for engineering analysis and design.

2.1 Fundamentals

Classical thermodynamics is concerned primarily with the macrostructure of matter. It addresses the gross characteristics of large aggregations of molecules and not the behavior of individual molecules. The microstructure of matter is studied in kinetic theory and statistical mechanics (including quantum thermodynamics). In this chapter, the classical approach to thermodynamics is featured.

Basic Concepts and Definitions

Thermodynamics is both a branch of physics and an engineering science. The scientist is normally interested in gaining a fundamental understanding of the physical and chemical behavior of fixed, quiescent quantities of matter and uses the principles of thermodynamics to relate the *properties* of matter. Engineers are generally interested in studying *systems* and how they interact with their *surroundings*. To facilitate this, engineers have extended the subject of thermodynamics to the study of systems through which matter flows.

System

In a thermodynamic analysis, the *system is* the subject of the investigation. Normally the system is a specified quantity of matter and/or a region that can be separated from everything else by a well-defined surface. The defining surface is known as the *control surface* or *system boundary*. The control surface may be movable or fixed. Everything external to the system is the *surroundings*. A system of fixed mass is referred to as a *control mass* or as a *closed system*. When there is flow of mass through the control surface, the system is called a *control volume*, or *open system*. An *isolated* system is a closed system that does not interact in any way with its surroundings.

State, Property

The condition of a system at any instant of time is called its *state*. The state at a given instant of time is described by the properties of the system. A *property is any* quantity whose numerical value depends on the state but not the history of the system. The value of a property is determined in principle by some type of physical operation or test.

Extensive properties depend on the size or extent of the system. Volume, mass, energy, and entropy are examples of extensive properties. An extensive property is additive in the sense that its value for the whole system equals the sum of the values for its parts. *Intensive* properties are independent of the size or extent of the system. Pressure and temperature are examples of intensive properties.

A mole is a quantity of substance having a mass numerically equal to its molecular weight. Designating the molecular weight by $\mathcal{M}$ and the number of moles by n, the mass m of the substance is $m = n\mathcal{M}$. One kilogram mole, designated kmol, of oxygen is 32.0 kg and one pound mole (lbmol) is 32.0 lb. When an extensive property is reported on a unit mass or a unit mole basis, it is called a *specific* property. An overbar is used to distinguish an extensive property written on a per-mole basis from its value expressed per unit mass. For example, the volume per mole is $\bar{v}$, whereas the volume per unit mass is v, and the two specific volumes are related by $\bar{v} = \mathcal{M}v$.

Process, Cycle

Two states are identical if, and only if, the properties of the two states are identical. When any property of a system changes in value there is a change in state, and the system is said to undergo a *process*. When a system in a given initial state goes through a sequence of processes and finally returns to its initial state, it is said to have undergone a *cycle*.

Phase and Pure Substance

The term *phase* refers to a quantity of matter that is homogeneous throughout in both chemical composition and physical structure. Homogeneity in physical structure means that the matter is all *solid*, or all *liquid*, or all *vapor* (or equivalently all *gas*). *A* system can contain one or more phases. For example, a

system of liquid water and water vapor (steam) contains *two* phases. A *pure substance* is one that is uniform and invariable in chemical composition. *A* pure substance can exist in more than one phase, but its chemical composition must be the same in each phase. For example, if liquid water and water vapor form a system with two phases, the system can be regarded as a pure substance because each phase has the same composition. The nature of phases that coexist in equilibrium is addressed by the *phase rule* (Section 2.3, Multicomponent Systems).

Equilibrium

Equilibrium means a condition of balance. In thermodynamics the concept includes not only a balance of forces, but also a balance of other influences. Each kind of influence refers to a particular aspect of thermodynamic (complete) equilibrium. *Thermal* equilibrium refers to an equality of temperature, *mechanical* equilibrium to an equality of pressure, and *phase* equilibrium to an equality of chemical potentials (Section 2.3, Multicomponent Systems). *Chemical* equilibrium is also established in terms of chemical potentials (Section 2.4, Reaction Equilibrium). For complete equilibrium the several types of equilibrium must exist individually.

To determine if a system is in thermodynamic equilibrium, one may think of testing it as follows: isolate the system from its surroundings and watch for changes in its observable properties. If there are no changes, it may be concluded that the system was in equilibrium at the moment it was isolated. The system can be said to be at an *equilibrium state*. When a system is *isolated*, it cannot interact with its surroundings; however, its state can change as a consequence of spontaneous events occurring internally as its intensive properties, such as temperature and pressure, tend toward uniform values. When all such changes cease, the system is in equilibrium. At equilibrium. temperature and pressure are uniform throughout. If gravity is significant, a pressure variation with height can exist, as in a vertical column of liquid.

Temperature

A scale of temperature independent of the *thermometric substance* is called a *thermodynamic* temperature scale. The Kelvin scale, a thermodynamic scale, can be elicited from the second law of thermodynamics (Section 2.1, The Second Law of Thermodynamics, Entropy). The definition of temperature following from the second law is valid over all temperature ranges and provides an essential connection between the several *empirical* measures of temperature. In particular, temperatures evaluated using a *constant-volume gas thermometer* are identical to those of the Kelvin scale over the range of temperatures where gas thermometry can be used.

The empirical *gas scale* is based on the experimental observations that (1) at a given temperature level all gases exhibit the same value of the product $p\bar{v}$ (p is pressure and $\bar{v}$ the specific volume on a molar basis) if the pressure is low enough, and (2) the value of the product $p\bar{v}$ increases with the temperature level. On this basis the gas temperature scale is defined by

$$T = \frac{1}{\bar{R}} \lim_{p \to 0} (p\bar{v})$$

where T is temperature and $\bar{R}$ is the *universal gas constant.* The absolute temperature at the *triple point of water* (Section 2.3, *P-v-T* Relations) is fixed by international agreement to be 273.16 K on the *Kelvin* temperature scale. $\bar{R}$ is then evaluated experimentally as $\bar{R}$ = 8.314 kJ/kmol · K (1545 ft · lbf/lbmol ·°R).

The *Celsius termperature scale* (also called the centigrade scale) uses the degree Celsius (°C), which has the same magnitude as the kelvin. Thus, temperature *differences* are identical on both scales. However, the zero point on the Celsius scale is shifted to 273.15 K, as shown by the following relationship between the Celsius temperature and the Kelvin temperature:

$$T(°\text{C}) = T(\text{K}) - 273.15 \tag{2.1}$$

On the Celsius scale, the triple point of water is 0.01°C and 0 K corresponds to –273.15°C.

Two other temperature scales are commonly used in engineering in the U.S. By definition, the *Rankine scale*, the unit of which is the degree rankine (°R), is proportional to the Kelvin temperature according to

$$T(°\mathrm{R}) = 1.8T(\mathrm{K}) \tag{2.2}$$

The Rankine scale is also an absolute thermodynamic scale with an absolute zero that coincides with the absolute zero of the Kelvin scale. In thermodynamic relationships, temperature is always in terms of the Kelvin or Rankine scale unless specifically stated otherwise.

A degree of the same size as that on the Rankine scale is used in the *Fahrenheit scale*, but the zero point is shifted according to the relation

$$T(°\mathrm{F}) = T(°\mathrm{R}) - 459.67 \tag{2.3}$$

Substituting Equations 2.1 and 2.2 into Equation 2.3 gives

$$T(°\mathrm{F}) = 1.8T(°\mathrm{C}) + 32 \tag{2.4}$$

This equation shows that the Fahrenheit temperature of the *ice point* (0°C) is 32°F and of the *steam point* (100°C) is 212°F. The 100 Celsius or Kelvin degrees between the ice point and steam point corresponds to 180 Fahrenheit or Rankine degrees.

To provide a standard for temperature measurement taking into account both theoretical and practical considerations, the International Temperature Scale of 1990 (ITS-90) is defined in such a way that the temperature measured on it conforms with the thermodynamic temperature, the unit of which is the kelvin, to within the limits of accuracy of measurement obtainable in 1990. Further discussion of ITS-90 is provided by Preston-Thomas (1990).

The First Law of Thermodynamics, Energy

Energy is a fundamental concept of thermodynamics and one of the most significant aspects of engineering analysis. Energy can be *stored* within systems in various macroscopic forms: kinetic energy, gravitational potential energy, and internal energy. Energy can also be *transformed* from one form to another and *transferred* between systems. For closed systems, energy can be transferred by *work* and *heat transfer.* The total amount of energy is *conserved* in all transformations and transfers.

Work

In thermodynamics, the term *work* denotes a means for transferring energy. Work is an effect of one system on another that is identified and measured as follows: work is done by a system on its surroundings if the *sole effect* on everything external to the system *could have been* the raising of a weight. The test of whether a work interaction has taken place is not that the elevation of a weight is actually changed, nor that a force actually acted through a distance, but that the sole effect *could be* the change in elevation of a mass. The magnitude of the work is measured by the number of standard weights that could have been raised. Since the raising of a weight is in effect a force acting through a distance, the work concept of mechanics is preserved. This definition includes work effects such as is associated with rotating shafts, displacement of the boundary, and the flow of electricity.

Work done *by* a system is considered positive: $W > 0$. Work done *on* a system is considered negative: $W < 0$. The time rate of doing work, or *power,* is symbolized by $\dot{W}$ and adheres to the same sign convention.

Energy

A closed system undergoing a process that involves only work interactions with its surroundings experiences an *adiabatic* process. On the basis of experimental evidence, it can be postulated that *when*

a closed system is altered adiabatically, the amount of work is fixed by the end states of the system and is independent of the details of the process. This postulate, which is one way the *first law of thermodynamics* can be stated, can be made regardless of the type of work interaction involved, the type of process, or the nature of the system.

As the work in an adiabatic process of a closed system is fixed by the end states, an extensive property called *energy* can be defined for the system such that its change between two states is the work in an adiabatic process that has these as the end states. In engineering thermodynamics the change in the energy of a system is considered to be made up of three macroscopic contributions: the change in *kinetic energy, KE,* associated with the motion of the system *as a whole* relative to an external coordinate frame, the change in gravitational *potential energy, PE,* associated with the position of the system *as a whole* in the Earth's gravitational field, and the change in *internal energy, U,* which accounts for all other energy associated with the system. Like kinetic energy and gravitational potential energy, internal energy is an extensive property.

In summary, the change in energy between two states of a closed system in terms of the work W_{ad} of an adiabatic process between these states is

$$\left(KE_2 - KE_1\right) + \left(PE_2 - PE_1\right) + \left(U_2 - U_1\right) = -W_{ad} \tag{2.5}$$

where 1 and 2 denote the initial and final states, respectively, and the minus sign before the work term is in accordance with the previously stated sign convention for work. Since any arbitrary value can be assigned to the energy of a system at a given state 1, no particular significance can be attached to the value of the energy at state 1 or at *any* other state. Only *changes* in the energy of a system have significance.

The specific energy (energy per unit mass) is the sum of the specific internal energy, u, the specific kinetic energy, $v^2/2$, and the specific gravitational potential energy, gz, such that

$$\text{specific energy} = u + \frac{v^2}{2} + gz \tag{2.6}$$

where the velocity v and the elevation z are each relative to specified datums (often the Earth's surface) and g is the acceleration of gravity.

A property related to internal energy u, pressure p, and specific volume v is *enthalpy,* defined by

$$h = u + pv \tag{2.7a}$$

or on an extensive basis

$$H = U + pV \tag{2.7b}$$

Heat

Closed systems can also interact with their surroundings in a way that cannot be categorized as work, as, for example, a gas (or liquid) contained in a closed vessel undergoing a process while in contact with a flame. This type of interaction is called a *heat interaction,* and the process is referred to as *nonadiabatic.*

A fundamental aspect of the energy concept is that energy is conserved. Thus, since a closed system experiences precisely the same energy change during a nonadiabatic process as during an adiabatic process between the same end states, it can be concluded that the *net* energy transfer to the system in each of these processes must be the same. It follows that heat interactions also involve energy transfer.

Denoting the amount of energy transferred *to* a closed system in heat interactions by Q, these considerations can be summarized by the *closed system energy balance:*

$$\left(U_2 - U_1\right) + \left(KE_2 - KE_1\right) + \left(PE_2 - PE_1\right) = Q - W \tag{2.8}$$

The closed system energy balance expresses the conservation of energy principle for closed systems of all kinds.

The quantity denoted by Q in Equation 2.8 accounts for the amount of energy transferred to a closed system during a process by means other than work. On the basis of experiments it is known that such an energy transfer is induced only as a result of a temperature difference between the system and its surroundings and occurs only in the direction of decreasing temperature. This means of energy transfer is called an *energy transfer by heat.* The following sign convention applies:

$Q > 0$: heat transfer *to* the system

$Q < 0$: heat transfer *from* the system

The time rate of heat transfer, denoted by $\dot{Q}$, adheres to the same sign convention.

Methods based on experiment are available for evaluating energy transfer by heat. These methods recognize two basic transfer mechanisms: *conduction* and *thermal radiation.* In addition, theoretical and empirical relationships are available for evaluating energy transfer involving *combined* modes such as *convection.* Further discussion of heat transfer fundamentals is provided in Chapter 4.

The quantities symbolized by W and Q account for *transfers* of energy. The terms *work* and *heat* denote different *means* whereby energy is transferred and not *what* is transferred. Work and heat are not properties, and it is improper to speak of work or heat "contained" in a system. However, to achieve economy of expression in subsequent discussions, W and Q are often referred to simply as work and heat transfer, respectively. This less formal approach is commonly used in engineering practice.

Power Cycles

Since energy is a property, over each cycle there is no net change in energy. Thus, Equation 2.8 reads for *any* cycle

$$Q_{cycle} = W_{cycle}$$

That is, for *any* cycle the net amount of energy received through heat interactions is equal to the net energy transferred out in work interactions. A *power cycle,* or *heat* engine, is one for which a net amount of energy is transferred out by work: $W_{cycle} > 0$. This equals the net amount of energy transferred in by heat.

Power cycles are characterized both by addition of energy by heat transfer, Q_A, and inevitable rejections of energy by heat transfer, Q_R:

$$Q_{cycle} = Q_A - Q_R$$

Combining the last two equations,

$$W_{cycle} = Q_A - Q_R$$

The *thermal efficiency* of a heat engine is defined as the ratio of the net work developed to the total energy added by heat transfer:

$$\eta = \frac{W_{cycle}}{Q_A} = 1 - \frac{Q_R}{Q_A} \tag{2.9}$$

The thermal efficiency is strictly less than 100%. That is, some portion of the energy Q_A supplied is invariably rejected $Q_R \neq 0$.

The Second Law of Thermodynamics, Entropy

Many statements of the second law of thermodynamics have been proposed. Each of these can be called a statement of the second law *or* a corollary of the second law since, if one is invalid, all are invalid. In every instance where a consequence of the second law has been tested directly or indirectly by experiment it has been verified. Accordingly, the basis of the second law, like every other physical law, is experimental evidence.

Kelvin-Planck Statement

The Kelvin-Plank statement of the second law of thermodynamics refers to a *thermal reservoir.* A thermal reservoir is a system that remains at a constant temperature even though energy is added or removed by heat transfer. A reservoir is an idealization, of course, but such a system can be approximated in a number of ways — by the Earth's atmosphere, large bodies of water (lakes, oceans), and so on. Extensive properties of thermal reservoirs, such as internal energy, can change in interactions with other systems even though the reservoir temperature remains constant, however.

The Kelvin-Planck statement of the second law can be given as follows: *It is impossible for any system to operate in a thermodynamic cycle and deliver a net amount of energy by work to its surroundings while receiving energy by heat transfer from a single thermal reservoir.* In other words, a *perpetual-motion machine of the second kind* is impossible. Expressed analytically, the Kelvin-Planck statement is

$$W_{cycle} \leq 0 \qquad (\text{single reservoir})$$

where the words *single reservoir* emphasize that the system communicates thermally only with a single reservoir as it executes the cycle. The "less than" sign applies when *internal irreversibilities* are present as the system of interest undergoes a cycle and the "equal to" sign applies only when no irreversibilities are present.

Irreversibilities

A process is said to be *reversible* if it is possible for its effects to be eradicated in the sense that there is some way by which *both* the system and its surroundings can be *exactly restored* to their respective initial states. A process is *irreversible* if there is no way to undo it. That is, there is no means by which the system and its surroundings can be exactly restored to their respective initial states. A system that has undergone an irreversible process is not necessarily precluded from being restored to its initial state. However, were the system restored to its initial state, it would not also be possible to return the surroundings to their initial state.

There are many effects whose presence during a process renders it irreversible. These include, but are not limited to, the following: heat transfer through a finite temperature difference; unrestrained expansion of a gas or liquid to a lower pressure; spontaneous chemical reaction; mixing of matter at different compositions or states; friction (sliding friction as well as friction in the flow of fluids); electric current flow through a resistance; magnetization or polarization with hysteresis; and inelastic deformation. The term *irreversibility* is used to identify effects such as these.

Irreversibilities can be divided into two classes, *internal* and *external.* Internal irreversibilities are those that occur within the system, while external irreversibilities are those that occur within the surroundings, normally the immediate surroundings. As this division depends on the location of the boundary there is some arbitrariness in the classification (by locating the boundary to take in the

immediate surroundings, all irreversibilities are internal). Nonetheless, valuable insights can result when this distinction between irreversibilities is made. When internal irreversibilities are absent during a process, the process is said to be *internally reversible.* At every intermediate state of an internally reversible process of a closed system, all intensive properties are uniform throughout each phase present: the temperature, pressure, specific volume, and other intensive properties do not vary with position. The discussions to follow compare the actual and internally reversible process concepts for two cases of special interest.

For a gas as the system, the work of expansion arises from the force exerted by the system to move the boundary against the resistance offered by the surroundings:

$$W = \int_1^2 F dx = \int_1^2 pA dx$$

where the force is the product of the moving area and the pressure exerted by the system there. Noting that Adx is the change in total volume of the system,

$$W = \int_1^2 p dV$$

This expression for work applies to both actual and internally reversible expansion processes. However, for an internally reversible process p is not only the pressure at the moving boundary but also the pressure of the entire system. Furthermore, for an internally reversible process the volume equals mv, where the specific volume v has a single value throughout the system at a given instant. Accordingly, the work of an internally reversible expansion (or compression) process is

$$W = m\int_1^2 p dv \tag{2.10}$$

When such a process of a closed system is represented by a continuous curve on a plot of pressure vs. specific volume, the area *under* the curve is the magnitude of the work per unit of system mass (area a-b-c′-d′ of Figure 2.3, for example).

Although improved thermodynamic performance can accompany the reduction of irreversibilities, steps in this direction are normally constrained by a number of practical factors often related to costs. For example, consider two bodies able to communicate thermally. With a *finite* temperature difference between them, a spontaneous heat transfer would take place and, as noted previously, this would be a source of irreversibility. The importance of the heat transfer irreversibility diminishes as the temperature difference narrows; and as the temperature difference between the bodies vanishes, the heat transfer approaches *ideality.* From the study of heat transfer it is known, however, that the transfer of a finite amount of energy by heat between bodies whose temperatures differ only slightly requires a considerable amount of time, a large heat transfer surface area, or both. To approach *ideality,* therefore, a heat transfer would require an exceptionally long time and/or an exceptionally large area, each of which has cost implications constraining what can be achieved practically.

Carnot Corollaries

The two corollaries of the second law known as *Carnot* corollaries state: (1) the thermal efficiency of an irreversible power cycle is always less than the thermal efficiency of a reversible power cycle when each operates between the same two thermal reservoirs; (2) all reversible power cycles operating between the same two thermal reservoirs have the same thermal efficiency. A cycle is considered *reversible* when there are no irreversibilities within the system as it undergoes the cycle, and heat transfers between the system and reservoirs occur ideally (that is, with a vanishingly small temperature difference).

Kelvin Temperature Scale

Carnot corollary 2 suggests that the thermal efficiency of a reversible power cycle operating between two thermal reservoirs depends only on the temperatures of the reservoirs and not on the nature of the substance making up the system executing the cycle or the series of processes. With Equation 2.9 it can be concluded that the ratio of the heat transfers is also related only to the temperatures, and is independent of the substance and processes:

$$\left(\frac{Q_C}{Q_H}\right)_{\substack{rev\\cycle}} = \psi\left(T_C, T_H\right)$$

where Q_H *is* the energy transferred to the system by heat transfer from a *hot* reservoir at temperature T_H, and Q_C is the energy rejected from the system to a *cold* reservoir at temperature T_C. The words *rev cycle* emphasize that this expression applies only to systems undergoing reversible cycles while operating between the two reservoirs. Alternative temperature scales correspond to alternative specifications for the function ψ in this relation.

The *Kelvin temperature scale is* based on $\psi(T_C, T_H) = T_C/T_H$. Then

$$\left(\frac{Q_C}{Q_H}\right)_{\substack{rev\\cycle}} = \frac{T_C}{T_H} \tag{2.11}$$

This equation defines only a ratio of temperatures. The specification of the Kelvin scale is completed by assigning a numerical value to one standard reference state. The state selected is the same used to define the *gas scale:* at the triple point of water the temperature is specified to be 273.16 K. If a reversible cycle is operated between a reservoir at the reference-state temperature and another reservoir at an unknown temperature *T,* then the latter temperature is related to the value at the reference state by

$$T = 273.16\left(\frac{Q}{Q'}\right)_{\substack{rev\\cycle}}$$

where Q is the energy received by heat transfer from the reservoir at temperature *T,* and Q' is the energy rejected to the reservoir at the reference temperature. Accordingly, a temperature scale is defined that is valid over all ranges of temperature and that is independent of the thermometric substance.

Carnot Efficiency

For the special case of a reversible power cycle operating between thermal reservoirs at temperatures T_H and T_C on the Kelvin scale, combination of Equations 2.9 and 2.11 results in

$$\eta_{\max} = 1 - \frac{T_C}{T_H} \tag{2.12}$$

called the *Carnot efficiency.* This is the efficiency of *all* reversible power cycles operating between thermal reservoirs at T_H and T_C. Moreover, it is the *maximum theoretical* efficiency that any power cycle, real or ideal, could have while operating between the same two reservoirs. As temperatures on the Rankine scale differ from Kelvin temperatures only by the factor 1.8, the above equation may be applied with either scale of temperature.

The Clausius Inequality

The Clausius inequality provides the basis for introducing two ideas instrumental for quantitative evaluations of processes of systems from a second law perspective: *entropy* and *entropy generation.* The Clausius inequality states that

$$\oint \left(\frac{\delta Q}{T}\right)_b \leq 0 \tag{2.13a}$$

where δQ represents the heat transfer at a part of the system boundary during a portion of the cycle, and T is the absolute temperature at that part of the boundary. The symbol δ is used to distinguish the differentials of *nonproperties*, such as heat and work, from the differentials of properties, written with the symbol d. The subscript b indicates that the integrand is evaluated at the boundary of the system executing the cycle. The symbol $\oint$ indicates that the integral is to be performed over all parts of the boundary and over the entire cycle. The Clausius inequality can be demonstrated using the Kelvin-Planck statement of the second law, and the significance of the inequality is the same: the equality applies when there are no internal irreversibilities as the system executes the cycle, and the inequality applies when internal irreversibilities are present.

The Clausius inequality can be expressed alternatively as

$$\oint \left(\frac{\delta Q}{T}\right)_b = -S_{gen} \tag{2.13b}$$

where S_{gen} can be viewed as representing the *strength* of the inequality. The value of S_{gen} is positive when internal irreversibilities are present, zero when no internal irreversibilities are present, and can never be negative. Accordingly, S_{gen} is a measure of the irreversibilities present within the system executing the cycle. In the next section, S_{gen} is identified as the *entropy* generated (or *produced*) by internal irreversibilities during the cycle.

Entropy and Entropy Generation

Entropy

Consider two cycles executed by a closed system. One cycle consists of an internally reversible process A from state 1 to state 2, followed by an internally reversible process C from state 2 to state 1. The other cycle consists of an internally reversible process B from state 1 to state 2, followed by the same process C from state 2 to state 1 as in the first cycle. For these cycles, Equation 2.13b takes the form

$$\left(\int_1^2 \frac{\delta Q}{T}\right)_A + \left(\int_2^1 \frac{\delta Q}{T}\right)_C = -S_{gen} = 0$$

$$\left(\int_1^2 \frac{\delta Q}{T}\right)_B + \left(\int_2^1 \frac{\delta Q}{T}\right)_C = -S_{gen} = 0$$

where S_{gen} has been set to zero since the cycles are composed of internally reversible processes. Subtracting these equations leaves

$$\left(\int_1^2 \frac{\delta Q}{T}\right)_A = \left(\int_1^2 \frac{\delta Q}{T}\right)_B$$

Since A and B are arbitrary, it follows that the integral of $\delta Q/T$ has the same value for *any* internally reversible process between the two states: the value of the integral depends on the end states only. It can be concluded, therefore, that the integral defines the change in some property of the system. Selecting the symbol S to denote this property, its change is given by

$$S_2 - S_1 = \left(\int_1^2 \frac{\delta Q}{T} \right)_{\substack{int \\ rev}} \tag{2.14a}$$

where the subscript *int rev* indicates that the integration is carried out for any internally reversible process linking the two states. This extensive property is called *entropy.*

Since entropy is a property, the change in entropy of a system in going from one state to another is the same for *all* processes, both internally reversible and irreversible, between these two states. In other words, once the change in entropy between two states has been evaluated, this is the magnitude of the entropy change for *any* process of the system between these end states.

The definition of entropy change expressed on a differential basis is

$$dS = \left(\frac{\delta Q}{T} \right)_{\substack{int \\ rev}} \tag{2.14b}$$

Equation 2.14b indicates that when a closed system undergoing an internally reversible process *receives* energy by heat transfer, the system experiences an *increase* in entropy. Conversely, when energy is *removed* from the system by heat transfer, the entropy of the system *decreases.* This can be interpreted to mean that an entropy transfer is *associated* with (or accompanies) heat transfer. The direction of the entropy transfer is the same as that of the heat transfer. In an *adiabatic* internally reversible process of a closed system the entropy would remain constant. A constant entropy process is called an *isentropic* process.

On rearrangement, Equation 2.14b becomes

$$\left(\delta Q\right)_{\substack{int \\ rev}} = TdS$$

Then, for an internally reversible process of a closed system between state 1 and state 2,

$$Q_{\substack{int \\ rev}} = m \int_1^2 Tds \tag{2.15}$$

When such a process is represented by a continuous curve on a plot of temperature vs. specific entropy, the area *under* the curve is the magnitude of the heat transfer per unit of system mass.

Entropy Balance

For a cycle consisting of an actual process from state 1 to state 2, during which internal irreversibilities are present, followed by an internally reversible process from state 2 to state 1, Equation 2.13b takes the form

$$\int_1^2 \left(\frac{\delta Q}{T} \right)_b + \int_2^1 \left(\frac{\delta Q}{T} \right)_{\substack{int \\ rev}} = -S_{gen}$$

where the first integral is for the actual process and the second integral is for the internally reversible process. Since no irreversibilities are associated with the internally reversible process, the term S_{gen} accounting for the effect of irreversibilities during the cycle can be identified with the actual process only.

Applying the definition of entropy change, the second integral of the foregoing equation can be expressed as

$$S_1 - S_2 = \int_2^1 \left(\frac{\delta Q}{T}\right)_{\substack{int \\ rev}}$$

Introducing this and rearranging the equation, the *closed system entropy balance* results:

$$S_2 - S_1 = \int_1^2 \left(\frac{\delta Q}{T}\right)_b + S_{gen} \qquad (2.16)$$

entropy change	entropy transfer	entropy generation

When the end states are fixed, the entropy change on the left side of Equation 2.16 can be evaluated independently of the details of the process from state 1 to state 2. However, the two terms on the right side depend explicitly on the nature of the process and cannot be determined solely from knowledge of the end states. The first term on the right side is associated with heat transfer to or from the system during the process. This term can be interpreted as the *entropy transfer associated with (or accompanying) heat transfer.* The direction of entropy transfer is the same as the direction of the heat transfer, and the same sign convention applies as for heat transfer: a positive value means that entropy is transferred into the system, and a negative value means that entropy is transferred out.

The entropy change of a system is not accounted for solely by entropy transfer, but is also due to the second term on the right side of Equation 2.16 denoted by S_{gen}. The term S_{gen} is positive when internal irreversibilities are present during the process and vanishes when internal irreversibilities are absent. This can be described by saying that entropy is *generated* (or produced) within the system by the action of irreversibilities. The second law of thermodynamics can be interpreted as specifying that entropy is generated by irreversibilities and conserved only in the limit as irreversibilities are reduced to zero. Since S_{gen} measures the effect of irreversibilities present within a system during a process, its value depends on the nature of the process and not solely on the end states. Entropy generation is *not* a property.

When applying the entropy balance, the objective is often to evaluate the entropy generation term. However, the value of the entropy generation for a given process of a system usually does not have much significance by itself. The significance is normally determined through comparison. For example, the entropy generation within a given component might be compared to the entropy generation values of the other components included in an overall system formed by these components. By comparing entropy generation values, the components where appreciable irreversibilities occur can be identified and rank ordered. This allows attention to be focused on the components that contribute most heavily to inefficient operation of the overall system.

To evaluate the entropy transfer term of the entropy balance requires information regarding both the heat transfer and the temperature on the boundary where the heat transfer occurs. The entropy transfer term is not always subject to direct evaluation, however, because the required information is either unknown or undefined, such as when the system passes through states sufficiently far from equilibrium. In practical applications, it is often convenient, therefore, to enlarge the system to include enough of the immediate surroundings that the temperature on the boundary of the *enlarged system* corresponds to the ambient temperature, T_{amb}. The entropy transfer term is then simply Q/T_{amb}. However, as the irreversibilities present would not be just those for the system of interest but those for the enlarged system, the entropy generation term would account for the effects of internal irreversibilities within the

system *and* external irreversibilities present within that portion of the surroundings included within the enlarged system.

A form of the entropy balance convenient for particular analyses is the *rate form:*

$$\frac{dS}{dt} = \sum_j \frac{\dot{Q}_j}{T_j} + \dot{S}_{gen} \tag{2.17}$$

where *dS/dt* is the time rate of change of entropy of the system. The term $\dot{Q}_j / T_j$ represents the time rate of entropy transfer through the portion of the boundary whose instantaneous temperature is T_j. The term $\dot{S}_{gen}$ accounts for the time rate of entropy generation due to irreversibilities within the system.

For a system *isolated* from its surroundings, the entropy balance is

$$\left(S_2 - S_1\right)_{isol} = S_{gen} \tag{2.18}$$

where S_{gen} is the total amount of entropy generated within the isolated system. Since entropy is generated in all actual processes, the only processes of an isolated system that actually can occur are those for which the entropy of the isolated system increases. This is known as the *increase of entropy principle.*

2.2 Control Volume Applications

Since most applications of engineering thermodynamics are conducted on a control volume basis, the control volume formulations of the mass, energy, and entropy balances presented in this section are especially important. These are given here in the form of *overall* balances. Equations of change for mass, energy, and entropy in the form of differential equations are also available in the literature (see, e.g., Bird et al., 1960).

Conservation of Mass

When applied to a control volume, the principle of mass conservation states: *The time rate of accumulation of mass within the control volume equals the difference between the total rates of mass flow in and out across the boundary.* An important case for engineering practice is one for which inward and outward flows occur, each through one or more ports. For this case the conservation of mass principle takes the form

$$\frac{dm_{cv}}{dt} = \sum_i \dot{m}_i - \sum_e \dot{m}_e \tag{2.19}$$

The left side of this equation represents the time rate of change of mass contained within the control volume, $\dot{m}_i$ denotes the mass flow rate at an inlet, and $\dot{m}_e$ is the mass flow rate at an outlet.

The *volumetric flow rate* through a portion of the control surface with area dA is the product of the velocity component normal to the area, v_n, times the area: $v_n\,dA$. The *mass* flow rate through dA is $\rho(v_n\,dA)$. The mass rate of flow through a port of area A is then found by integration over the area

$$\dot{m} = \int_A \rho v_n \, dA$$

For *one-dimensional* flow the intensive properties are uniform with position over area A, and the last equation becomes

$$\dot{m} = \rho v A = \frac{vA}{v} \tag{2.20}$$

where v denotes the specific volume and the subscript n has been dropped from velocity for simplicity.

Control Volume Energy Balance

When applied to a control volume, the principle of energy conservation states: *The time rate of accumulation of energy within the control volume equals the difference between the total incoming rate of energy transfer and the total outgoing rate of energy transfer.* Energy can enter and exit a control volume by work and heat transfer. Energy also enters and exits with flowing streams of matter. Accordingly, for a control volume with one-dimensional flow at a single inlet and a single outlet,

$$\frac{d(U + KE + PE)_{cv}}{dt} = \dot{Q}_{cv} - \dot{W} + \dot{m}\left(u_i + \frac{v_i^2}{2} + gz_i\right) - \dot{m}\left(u_e + \frac{v_e^2}{2} + gz_e\right) \tag{2.21}$$

where the underlined terms account for the specific energy of the incoming and outgoing streams. The terms $\dot{Q}_{cv}$ and $\dot{W}$ account, respectively, for the net rates of energy transfer by heat and work over the boundary (control surface) of the control volume.

Because work is always done on or by a control volume where matter flows across the boundary, the quantity $\dot{W}$ of Equation 2.21 can be expressed in terms of two contributions: one is the work associated with the force of the fluid pressure as mass is introduced at the inlet and removed at the exit. The other, denoted as $\dot{W}_{cv}$, includes *all other* work effects, such as those associated with rotating shafts, displacement of the boundary, and electrical effects. The work rate concept of mechanics allows the first of these contributions to be evaluated in terms of the product of the pressure force, pA, and velocity at the point of application of the force. To summarize, the work term $\dot{W}$ of Equation 2.21 can be expressed (with Equation 2.20) as

$$\begin{aligned}\dot{W} &= \dot{W}_{cv} + \left(p_e A_e\right)\mathrm{v}_e - \left(p_i A_i\right)\mathrm{v}_i \\ &= \dot{W}_{cv} + \dot{m}_e\left(p_e v_e\right) - \dot{m}_i\left(p_i v_i\right)\end{aligned} \tag{2.22}$$

The terms $\dot{m}_i\,(pv_i)$ and $\dot{m}_e\,(p_e v_e)$ account for the work associated with the pressure at the inlet and outlet, respectively, and are commonly referred to as *flow work.*

Substituting Equation 2.22 into Equation 2.21, and introducing the specific enthalpy h, the following form of the control volume energy rate balance results:

$$\frac{d(U + KE + PE)_{cv}}{dt} = \dot{Q}_{cv} - \dot{W}_{cv} + \dot{m}_i\left(h_i + \frac{\mathrm{v}_i^2}{2} + \mathrm{g}z_i\right) - \dot{m}_e\left(h_e + \frac{\mathrm{v}_e^2}{2} + \mathrm{g}z_e\right) \tag{2.23}$$

To allow for applications where there may be several locations on the boundary through which mass enters or exits, the following expression is appropriate:

$$\frac{d(U + KE + PE)_{cv}}{dt} = \dot{Q}_{cv} - \dot{W}_{cv} + \sum_i \dot{m}_i\left(h_i + \frac{\mathrm{v}_i^2}{2} + \mathrm{g}z_i\right) - \sum_e \dot{m}_e\left(h_e + \frac{\mathrm{v}_e^2}{2} + \mathrm{g}z_e\right) \tag{2.24}$$

Equation 2.24 is an *accounting* rate balance for the energy of the control volume. It states that the time rate of accumulation of energy within the control volume equals the difference between the total rates of energy transfer in and out across the boundary. The mechanisms of energy transfer are heat and work, as for closed systems, and the energy accompanying the entering and exiting mass.

Control Volume Entropy Balance

Like mass and energy, entropy is an extensive property. And like mass and energy, entropy can be transferred into or out of a control volume by streams of matter. As this is the principal difference between the closed system and control volume forms, the control volume entropy rate balance is obtained by modifying Equation 2.17 to account for these entropy transfers. The result is

$$\underbrace{\frac{dS_{cv}}{dt}}_{\substack{\text{rate of}\\ \text{entropy}\\ \text{change}}} = \underbrace{\sum_j \frac{\dot{Q}_j}{T_j} + \sum_i \dot{m}_i s_i - \sum_e \dot{m}_e s_e}_{\substack{\text{rate of}\\ \text{entropy}\\ \text{transfer}}} + \underbrace{\dot{S}_{gen}}_{\substack{\text{rate of}\\ \text{entropy}\\ \text{generation}}} \tag{2.25}$$

where dS_{cv}/dt represents the time rate of change of entropy within the control volume. The terms $\dot{m}_i s_i$ and $\dot{m}_e s_e$ account, respectively, for rates of entropy *transfer* into and out of the control volume associated with mass flow. One-dimensional flow is assumed at locations where mass enters and exits. $\dot{Q}_j$ represents the time rate of heat transfer at the location on the boundary where the instantaneous temperature is T_j; and $\dot{Q}_j / T_j$ accounts for the associated rate of entropy *transfer*. $\dot{S}_{gen}$ denotes the time rate of entropy *generation* due to irreversibilities *within* the control volume. When a control volume comprises a number of components, $\dot{S}_{gen}$ is the sum of the rates of entropy generation of the components.

Control Volumes at Steady State

Engineering systems are often idealized as being at *steady state*, meaning that all properties are unchanging in time. For a control volume at steady state, the identity of the matter within the control volume change continuously, but the total amount of mass remains constant. At steady state, Equation 2.19 reduces to

$$\sum_i \dot{m}_i = \sum_e \dot{m}_e \tag{2.26a}$$

The energy rate balance of Equation 2.24 becomes, at steady state,

$$0 = \dot{Q}_{cv} - \dot{W}_{cv} + \sum_i \dot{m}_i \left(h_i + \frac{\mathrm{v}_i^2}{2} + \mathrm{g}\mathrm{z}_i \right) - \sum_e \dot{m}_e \left(h_e + \frac{\mathrm{v}_e^2}{2} + \mathrm{g}\mathrm{z}_e \right) \tag{2.26b}$$

At steady state, the entropy rate balance of Equation 2.25 reads

$$0 = \sum_j \frac{\dot{Q}_j}{T_j} + \sum_i \dot{m}_i s_i - \sum_e \dot{m}_e s_e + \dot{S}_{gen} \tag{2.26c}$$

Mass and energy are conserved quantities, but entropy is not generally conserved. Equation 2.26a indicates that the total rate of mass flow into the control volume equals the total rate of mass flow out of the control volume. Similarly, Equation 2.26b states that the total rate of energy transfer into the control volume equals the total rate of energy transfer out of the control volume. However, Equation 2.26c shows that the rate at which entropy is transferred out *exceeds* the rate at which entropy enters, the difference being the rate of entropy generation within the control volume owing to irreversibilities.

Applications frequently involve control volumes having a single inlet and a single outlet, as, for example, the control volume of Figure 2.1 where heat transfer (if any) occurs at T_b: the temperature, or a suitable average temperature, on the boundary where heat transfer occurs. For this case the mass rate balance, Equation 2.26a, reduces to $\dot{m}_i = \dot{m}_e$. Denoting the common mass flow rate by $\dot{m}$, Equations 2.26b and 2.26c read, respectively,

$$0 = \dot{Q}_{cv} - \dot{W}_{cv} + \dot{m} \left[\left(h_i - h_e \right) + \left(\frac{\mathrm{v}_i^2 - \mathrm{v}_e^2}{2} \right) + \mathrm{g}\left(\mathrm{z}_i - \mathrm{z}_e \right) \right] \tag{2.27a}$$

$$0 = \frac{\dot{Q}_{cv}}{T_b} + \dot{m}\left(s_i - s_e \right) + \dot{S}_{gen} \tag{2.28a}$$

When Equations 2.27a and 2.28a are applied to particular cases of interest, additional simplifications are usually made. The heat transfer term $\dot{Q}_{cv}$ is dropped when it is insignificant relative to other energy

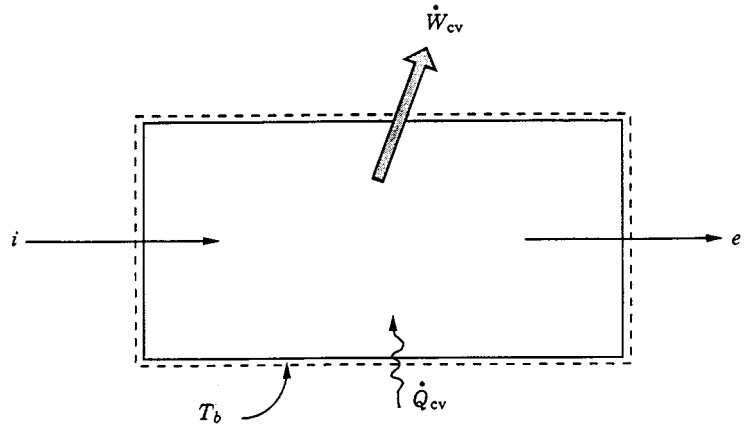


FIGURE 2.1 One-inlet, one-outlet control volume at steady state.

transfers across the boundary. This may be the result of one or more of the following: (1) the outer surface of the control volume is insulated; (2) the outer surface area is too small for there to be effective heat transfer; (3) the temperature difference between the control volume and its surroundings is small enough that the heat transfer can be ignored; (4) the gas or liquid passes through the control volume so quickly that there is not enough time for significant heat transfer to occur. The work term $\dot{W}_{cv}$ drops out of the energy rate balance when there are no rotating shafts, displacements of the boundary, electrical effects, or other work mechanisms associated with the control volume being considered. The changes in kinetic and potential energy of Equation 2.27a are frequently negligible relative to other terms in the equation.

The special forms of Equations 2.27a and 2.28a listed in Table 2.1 are obtained as follows: when there is no heat transfer, Equation 2.28a gives

$$s_e - s_i = \frac{\dot{S}_{gen}}{\dot{m}} \geq 0 \qquad \text{(2.28b)}$$

$$\text{(no heat transfer)}$$

Accordingly, when irreversibilities are present within the control volume, the specific entropy increases as mass flows from inlet to outlet. In the ideal case in which no internal irreversibilities are present, mass passes through the control volume with no change in its entropy — that is, *isentropically.*

For no heat transfer, Equation 2.27a gives

$$\dot{W}_{cv} = \dot{m}\left[\left(h_i - h_e\right) + \left(\frac{\mathrm{v}_i^2 - \mathrm{v}_e^2}{2}\right) + \mathrm{g}\left(\mathrm{z}_i - \mathrm{z}_e\right)\right] \qquad \text{(2.27b)}$$

A special form that is applicable, at least approximately, to *compressors*, *pumps*, and *turbines* results from dropping the kinetic and potential energy terms of Equation 2.27b, leaving

$$\dot{W}_{cv} = \dot{m}\left(h_i - h_e\right) \qquad \text{(2.27c)}$$

$$\text{(}\textit{compressors, pumps,}\text{ and }\textit{turbines}\text{)}$$

TABLE 2.1 Energy and Entropy Balances for One-Inlet, One-Outlet Control Volumes at Steady State and No Heat Transfer

Energy balance

$$\dot{W}_{cv} = \dot{m}\left[\left(h_i - h_e\right) + \left(\frac{v_i^2 - v_e^2}{2}\right) + g\left(z_i - z_e\right)\right] \tag{2.27b}$$

Compressors, pumps, and turbines[a]

$$\dot{W}_{cv} = \dot{m}\left(h_i - h_e\right) \tag{2.27c}$$

Throttling

$$h_e \cong h_i \tag{2.27d}$$

Nozzles, diffusers[b]

$$v_e = \sqrt{v_i^2 + 2\left(h_i - h_e\right)} \tag{2.27f}$$

Entropy balance

$$s_e - s_i = \frac{\dot{S}_{gen}}{\dot{m}} \geq 0 \tag{2.28b}$$

[a] For an ideal gas with constant c_p, Equation 1′ of Table 2.7 allows Equation 2.27c to be written as

$$\dot{W}_{cv} = \dot{m}c_p\left(T_i - T_e\right) \tag{2.27c′}$$

The power developed in an *isentropic process* is obtained with Equation 5′ of Table 2.7 as

$$\dot{W}_{cv} = \dot{m}c_p T_i\left[1 - \left(p_e/p_i\right)^{(k-1)/k}\right] \quad (s = c) \tag{2.27c″}$$

where $c_p = kR/(k-1)$.

[b] For an ideal gas with constant c_p, Equation 1′ of Table 2.7 allows Equation 2.27f to be written as

$$v_e = \sqrt{v_i^2 + 2c_p\left(T_i - T_e\right)} \tag{2.27f′}$$

The exit velocity for an *isentropic process* is obtained with Equation 5′ of Table 2.7 as

$$v_e = \sqrt{v_i^2 + 2c_p T_i\left[1 - \left(p_e/p_i\right)^{(k-1)/k}\right]} \quad (s = c) \tag{2.27f″}$$

where $c_p = kR/(k-1)$.

In *throttling devices* a significant reduction in pressure is achieved simply by introducing a restriction into a line through which a gas or liquid flows. For such devices $\dot{W}_{cv} = 0$ and Equation 2.27c reduces further to read

$$h_e \cong h_i \qquad (\textit{throttling process}) \tag{2.27d}$$

That is, upstream and downstream of the throttling device, the specific enthalpies are equal.

A *nozzle* is a flow passage of varying cross-sectional area in which the velocity of a gas or liquid increases in the direction of flow. In a *diffuser,* the gas or liquid decelerates in the direction of flow. For such devices, $\dot{W}_{cv} = 0$. The heat transfer and potential energy change are also generally negligible. Then Equation 2.27b reduces to

$$0 = h_i - h_e + \frac{v_i^2 - v_e^2}{2} \tag{2.27e}$$

Solving for the outlet velocity

$$v_e = \sqrt{v_i^2 + 2(h_i - h_e)} \qquad (2.27f)$$

$$(\textit{nozzle, diffuser})$$

Further discussion of the flow-through nozzles and diffusers is provided in Chapter 3.

The mass, energy, and entropy rate balances, Equations 2.26, can be applied to control volumes with multiple inlets and/or outlets, as, for example, cases involving heat-recovery steam generators, feedwater heaters, and counterflow and crossflow heat exchangers. Transient (or unsteady) analyses can be conducted with Equations 2.19, 2.24, and 2.25. Illustrations of all such applications are provided by Moran and Shapiro (1995).

Example 1

A turbine receives steam at 7 MPa, 440°C and exhausts at 0.2 MPa for subsequent process heating duty. If heat transfer and kinetic/potential energy effects are negligible, determine the steam mass flow rate, in kg/hr, for a turbine power output of 30 MW when (a) the steam quality at the turbine outlet is 95%, (b) the turbine expansion is internally reversible.

Solution. With the indicated idealizations, Equation 2.27c is appropriate. Solving, $\dot{m} = \dot{W}_{cv}/(h_i - h_e)$. Steam table data (Table A.5) at the inlet condition are h_i = 3261.7 kJ/kg, s_i = 6.6022 kJ/kg · K.

(a) At 0.2 MPa and x = 0.95, h_e = 2596.5 kJ/kg. Then

$$\dot{m} = \frac{30\ \text{MW}}{(3261.7 - 2596.5)\,\text{kJ/kg}}\left(\frac{10^3\,\text{kJ/sec}}{1\ \text{MW}}\right)\left(\frac{3600\ \text{sec}}{1\ \text{hr}}\right)$$

$$= 162{,}357\ \text{kg/hr}$$

(b) For an internally reversible expansion, Equation 2.28b reduces to give $s_e = s_i$. For this case, h_e = 2499.6 kJ/kg (x = 0.906), and $\dot{m}$ = 141,714 kg/hr.

Example 2

Air at 500°F, 150 lbf/in.^{2}, and 10 ft/sec expands adiabatically through a nozzle and exits at 60°F, 15 lbf/in.^{2}. For a mass flow rate of 5 lb/sec determine the exit area, in in.^{2}. Repeat for an isentropic expansion to 15 lbf/in.^{2}. Model the air as an ideal gas (Section 2.3, Ideal Gas Model) with specific heat c_p = 0.24 Btu/lb · °R (k = 1.4).

Solution. The nozle exit area can be evaluated using Equation 2.20, together with the ideal gas equation, $v = RT/p$:

$$A_e = \frac{\dot{m}v_e}{v_e} = \frac{\dot{m}(RT_e/p_e)}{v_e}$$

The exit velocity required by this expression is obtained using Equation 2.27f′ of Table 2.1,

$$v_e = \sqrt{v_i^2 + 2c_p(T_i - T_e)}$$

$$= \sqrt{\left(\frac{10\ \text{ft}}{s}\right)^2 + 2\left(0.24\frac{\text{Btu}}{\text{lb}\cdot\text{R}}\right)\left(\frac{778.17\ \text{ft}\cdot\text{lbf}}{1\ \text{Btu}}\right)(440°\text{R})\left(\frac{32.174\ \text{lb}\cdot\text{ft/sec}^2}{1\ \text{lbf}}\right)}$$

$$= 2299.5\ \text{ft sec}$$

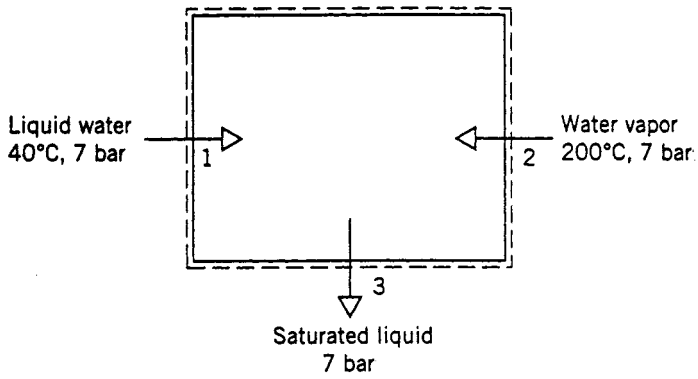


FIGURE 2.2 Open feedwater heater.

Finally, with $R = \bar{R}/\mathcal{M} = 53.33$ ft · lbf/lb · °R,

$$A_e = \frac{\left(5\frac{\text{lb}}{\text{sec}}\right)\left(53.3\frac{\text{ft}\cdot\text{lbf}}{\text{lb}\cdot°\text{R}}\right)(520°\text{R})}{\left(2299.5\frac{\text{ft}}{\text{sec}}\right)\left(15\frac{\text{lbf}}{\text{in.}^2}\right)} = 4.02 \text{ in.}^2$$

Using Equation 2.27f″ in Table 2.1 for the isentropic expansion,

$$v_e = \sqrt{(10)^2 + 2(0.24)(778.17)(960)(32.174)\left[1-\left(\frac{15}{150}\right)^{0.4/1.4}\right]}$$

$$= 2358.3 \text{ ft/sec}$$

Then $A_e = 3.92$ in.^{2}.

Example 3

Figure 2.2 provides steady-state operating data for an open feedwater heater. Ignoring heat transfer and kinetic/potential energy effects, determine the ratio of mass flow rates, $\dot{m}_1/\dot{m}_2$.

Solution. For this case Equations 2.26a and 2.26b reduce to read, respectively,

$$\dot{m}_1 + \dot{m}_2 = \dot{m}_3$$

$$0 = \dot{m}_1 h_1 + \dot{m}_2 h_2 - \dot{m}_3 h_3$$

Combining and solving for the ratio $\dot{m}_1/\dot{m}_2$,

$$\frac{\dot{m}_1}{\dot{m}_2} = \frac{h_2 - h_3}{h_3 - h_1}$$

Inserting steam table data, in kJ/kg, from Table A.5,

$$\frac{\dot{m}_1}{\dot{m}_2} = \frac{2844.8 - 697.2}{697.2 - 167.6} = 4.06$$

Internally Reversible Heat Transfer and Work

For one-inlet, one-outlet control volumes at steady state, the following expressions give the heat transfer rate and power in the absence of internal irreversibilities:

$$\left(\frac{\dot{Q}_{cv}}{\dot{m}}\right)_{\substack{int\\rev}} = \int_1^2 T ds \tag{2.29}$$

$$\left(\frac{\dot{W}_{cv}}{\dot{m}}\right)_{\substack{int\\rev}} = -\int_1^2 v\,dp + \frac{v_1^2 - v_2^2}{2} + g(z_1 - z_2) \tag{2.30a}$$

(see, e.g., Moran and Shapiro, 1995).

If there is no significant change in kinetic or potential energy from inlet to outlet, Equation 2.30a reads

$$\left(\frac{\dot{W}_{cv}}{\dot{m}}\right)_{\substack{int\\rev}} = -\int_1^2 v\,dp \qquad (\Delta ke = \Delta pe = 0) \tag{2.30b}$$

The specific volume remains approximately constant in many applications with liquids. Then Equation 30b becomes

$$\left(\frac{\dot{W}_{cv}}{\dot{m}}\right)_{\substack{int\\rev}} = -v(p_2 - p_1) \qquad (v = \text{constant}) \tag{2.30c}$$

When the states visited by a unit of mass flowing without irreversibilities from inlet to outlet are described by a continuous curve on a plot of temperature vs. specific entropy, Equation 2.29 implies that the area under the curve is the magnitude of the heat transfer per unit of mass flowing. When such an ideal process is described by a curve on a plot of pressure vs. specific volume, as shown in Figure 2.3, the magnitude of the integral $\int v dp$ of Equations 2.30a and 2.30b is represented by the area a-b-c-d *behind* the curve. The area a-b-c′-d′ *under* the curve is identified with the magnitude of the integral $\int p dv$ of Equation 2.10.

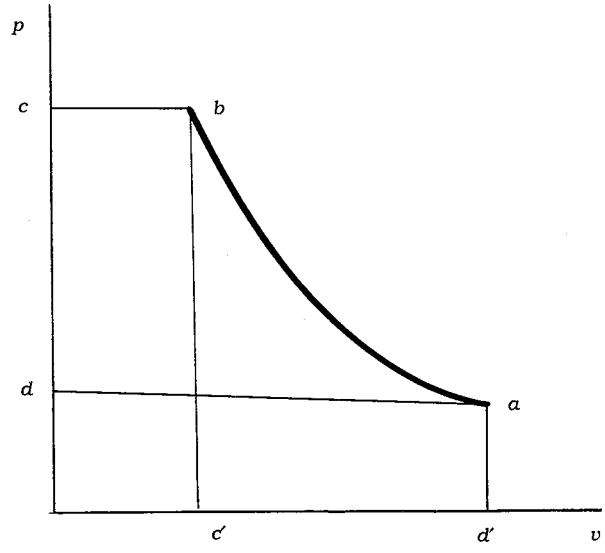


FIGURE 2.3 Internally reversible process on *p*–*v* coordinates.

2.3 Property Relations and Data

Pressure, temperature, volume, and mass can be found experimentally. The relationships between the specific heats c_v and c_p and temperature at relatively low pressure are also accessible experimentally, as are certain other property data. Specific internal energy, enthalpy, and entropy are among those properties that are not so readily obtained in the laboratory. Values for such properties are calculated using experimental data of properties that are more amenable to measurement, together with appropriate property relations derived using the principles of thermodynamics. In this section property relations and data sources are considered for *simple compressible systems*, which include a wide range of industrially important substances.

Property data are provided in the publications of the *National Institute of Standards and Technology* (formerly the U.S. Bureau of Standards), of professional groups such as the *American Society of Mechanical Engineering (ASME)*, the *American Society of Heating. Refrigerating, and Air Conditioning Engineers (ASHRAE)*, and the *American Chemical Society*, and of corporate entities such as *Dupont* and *Dow Chemical*. Handbooks and property reference volumes such as included in the list of references for this chapter are readily accessed sources of data. Property data are also retrievable from various commercial online data bases. Computer software is increasingly available for this purpose as well.

Basic Relations for Pure Substances

An energy balance in differential form for a closed system undergoing an internally reversible process in the absence of overall system motion and the effect of gravity reads

$$dU = (\delta Q)_{\substack{int\\rev}} - (\delta W)_{\substack{int\\rev}}$$

From Equation 2.14b, $(\delta Q)_{\substack{int\\rev}} = TdS$. When consideration is limited to *simple compressible systems:* systems for which the only significant work in an internally reversible process is associated with volume change, $(\delta W)_{\substack{int\\rev}} = pdV$, the following equation is obtained:

$$dU = TdS - pdV \tag{2.31a}$$

Introducing enthalpy, $H = U + pV$, the Helmholtz function, $\Psi = U - TS$, and the Gibbs function, $G = H - TS$, three additional expressions are obtained:

$$dH = TdS + Vdp \tag{2.31b}$$

$$d\Psi = -pdV - SdT \tag{2.31c}$$

$$dG = Vdp - SdT \tag{2.31d}$$

Equations 2.31 can be expressed on a per-unit-mass basis as

$$du = Tds - pdv \tag{2.32a}$$

$$dh = Tds + vdp \tag{2.32b}$$

$$d\psi = -pdv - sdT \tag{2.32c}$$

$$dg = vdp - sdT \tag{2.32d}$$

Similar expressions can be written on a per-mole basis.

Maxwell Relations

Since only properties are involved, each of the four differential expressions given by Equations 2.32 is an *exact* differential exhibiting the general form $dz = M(x, y)dx + N(x, y)dy$, where the second mixed partial derivatives are equal: $(\partial M/\partial y) = (\partial N/\partial x)$. Underlying these exact differentials are, respectively, functions of the form $u(s, v)$, $h(s, p)$, $\psi(v, T)$, and $g(T, p)$. From such considerations the *Maxwell relations* given in Table 2.2 can be established.

Example 4

Derive the Maxwell relation following from Equation 2.32a.

TABLE 2.2 Relations from Exact Differentials

Function	*Differential*	*Coefficients*	*Maxwell*
General:			
$z = z(x, y)$	$dz = M(x, y)dx + N(x, y)dy$	$\left(\frac{\partial z}{\partial x}\right)_y = M$ $\left(\frac{\partial z}{\partial y}\right)_x = N$	$\left(\frac{\partial M}{\partial y}\right)_x = \left(\frac{\partial N}{\partial x}\right)_y$
Internal energy:			
$u(s, v)$	$du = Tds - pdv$	$\left(\frac{\partial u}{\partial s}\right)_v = T$ $\left(\frac{\partial u}{\partial v}\right)_s = -p$	$\left(\frac{\partial T}{\partial v}\right)_s = -\left(\frac{\partial p}{\partial s}\right)_v$
Enthalpy:			
$h(s, p)$	$dh = Tds + vdp$	$\left(\frac{\partial h}{\partial s}\right)_p = T$ $\left(\frac{\partial h}{\partial p}\right)_s = v$	$\left(\frac{\partial T}{\partial p}\right)_s = \left(\frac{\partial v}{\partial s}\right)_p$
Helmholtz function:			
$\psi(v, T)$	$d\psi = -pdv - sdT$	$\left(\frac{\partial \psi}{\partial v}\right)_T = -p$ $\left(\frac{\partial \psi}{\partial T}\right)_v = -s$	$\left(\frac{\partial p}{\partial T}\right)_v = \left(\frac{\partial s}{\partial v}\right)_T$
Gibbs function:			
$g(T, p)$	$dg = vdp - sdT$	$\left(\frac{\partial g}{\partial p}\right)_T = v$ $\left(\frac{\partial g}{\partial T}\right)_p = -s$	$\left(\frac{\partial v}{\partial T}\right)_p = -\left(\frac{\partial s}{\partial p}\right)_T$

Solution. The differential of the function $u = u(s, v)$ is

$$du = \left(\frac{\partial u}{\partial s}\right)_v ds + \left(\frac{\partial u}{\partial v}\right)_s dv$$

By comparison with Equation 2.32a,

$$T = \left(\frac{\partial u}{\partial s}\right)_v, \quad -p = \left(\frac{\partial u}{\partial v}\right)_s$$

In Equation 2.32a, T plays the role of M and $-p$ plays the role of N, so the equality of second mixed partial derivatives gives the Maxwell relation,

$$\left(\frac{\partial T}{\partial v}\right)_s = -\left(\frac{\partial p}{\partial s}\right)_v$$

Since each of the properties T, p, v, and s appears on the right side of two of the eight coefficients of Table 2.2, four additional property relations can be obtained by equating such expressions:

$$\left(\frac{\partial u}{\partial s}\right)_v = \left(\frac{\partial h}{\partial s}\right)_p, \quad \left(\frac{\partial u}{\partial v}\right)_s = \left(\frac{\partial \psi}{\partial v}\right)_T$$

$$\left(\frac{\partial h}{\partial p}\right)_s = \left(\frac{\partial g}{\partial p}\right)_T, \quad \left(\frac{\partial \psi}{\partial T}\right)_v = \left(\frac{\partial g}{\partial T}\right)_p$$

These four relations are identified in Table 2.2 by brackets. As any three of Equations 2.32 can be obtained from the fourth simply by manipulation, the 16 property relations of Table 2.2 also can be regarded as following from this single differential expression. Several additional first-derivative property relations can be derived; see, e.g., Zemansky, 1972.

Specific Heats and Other Properties

Engineering thermodynamics uses a wide assortment of thermodynamic properties and relations among these properties. Table 2.3 lists several commonly encountered properties.

Among the entries of Table 2.3 are the specific heats c_v and c_p. These intensive properties are often required for thermodynamic analysis, and are defined as partial derivations of the functions $u(T, v)$ and $h(T, p)$, respectively,

$$c_v = \left(\frac{\partial u}{\partial T}\right)_v \tag{2.33}$$

$$c_p = \left(\frac{\partial h}{\partial T}\right)_p \tag{2.34}$$

Since u and h can be expressed either on a unit mass basis or a per-mole basis, values of the specific heats can be similarly expressed. Table 2.4 summarizes relations involving c_v and c_p. The property k, the specific heat ratio, is

$$k = \frac{c_p}{c_v} \tag{2.35}$$

TABLE 2.3 Symbols and Definitions for Selected Properties

Property	Symbol	Definition	Property	Symbol	Definition
Pressure	p		Specific heat, constant volume	c_v	$\left(\partial u/\partial T\right)_v$
Temperature	T		Specific heat, constant pressure	c_p	$\left(\partial h/\partial T\right)_p$
Specific volume	v		Volume expansivity	β	$\frac{1}{v}\left(\partial v/\partial T\right)_p$
Specific internal energy	u		Isothermal compressivity	κ	$-\frac{1}{v}\left(\partial v/\partial p\right)_T$
Specific entropy	s		Isentropic compressibility	α	$-\frac{1}{v}\left(\partial v/\partial p\right)_s$
Specific enthalpy	h	$u + pv$	Isothermal bulk modulus	B	$-v\left(\partial p/\partial v\right)_T$
Specific Helmholtz function	ψ	$u - Ts$	Isentropic bulk modulus	B_s	$-v\left(\partial p/\partial v\right)_s$
Specific Gibbs function	g	$h - Ts$	Joule-Thomson coefficient	μ_J	$\left(\partial T/\partial p\right)_h$
Compressibility factor	Z	pv/RT	Joule coefficient	η	$\left(\partial T/\partial v\right)_u$
Specific heat ratio	k	c_p/c_v	Velocity of sound	c	$\sqrt{-v^2\left(\partial p/\partial v\right)_s}$

Values for c_v and c_p can be obtained via statistical mechanics using *spectroscopic* measurements. They can also be determined macroscopically through exacting property measurements. Specific heat data for common gases, liquids, and solids are provided by the handbooks and property reference volumes listed among the Chapter 2 references. Specific heats are also considered in Section 2.3 as a part of the discussions of the *incompressible model* and the *ideal gas model*. Figure 2.4 shows how c_p for water vapor varies as a function of temperature and pressure. Other gases exhibit similar behavior. The figure also gives the variation of c_p with temperature in the limit as pressure tends to zero (the ideal gas limit). In this limit c_p increases with increasing temperature, which is a characteristic exhibited by other gases as well

The following two equations are often convenient for establishing relations among properties:

$$\left(\frac{\partial x}{\partial y}\right)_z\left(\frac{\partial y}{\partial x}\right)_z = 1 \tag{2.36a}$$

$$\left(\frac{\partial y}{\partial z}\right)_x\left(\frac{\partial z}{\partial x}\right)_y\left(\frac{\partial x}{\partial y}\right)_z = -1 \tag{2.36b}$$

Their use is illustrated in Example 5.

Example 5

Obtain Equations 2 and 11 of Table 2.4 from Equation 1.

Solution. Identifying x, y, z with s, T, and v, respectively, Equation 2.36b reads

$$\left(\frac{\partial T}{\partial v}\right)_s\left(\frac{\partial v}{\partial s}\right)_T\left(\frac{\partial s}{\partial T}\right)_v = -1$$

Applying Equation 2.36a to each of $(\partial T/\partial v)_s$ and $(\partial v/\partial s)_T$,

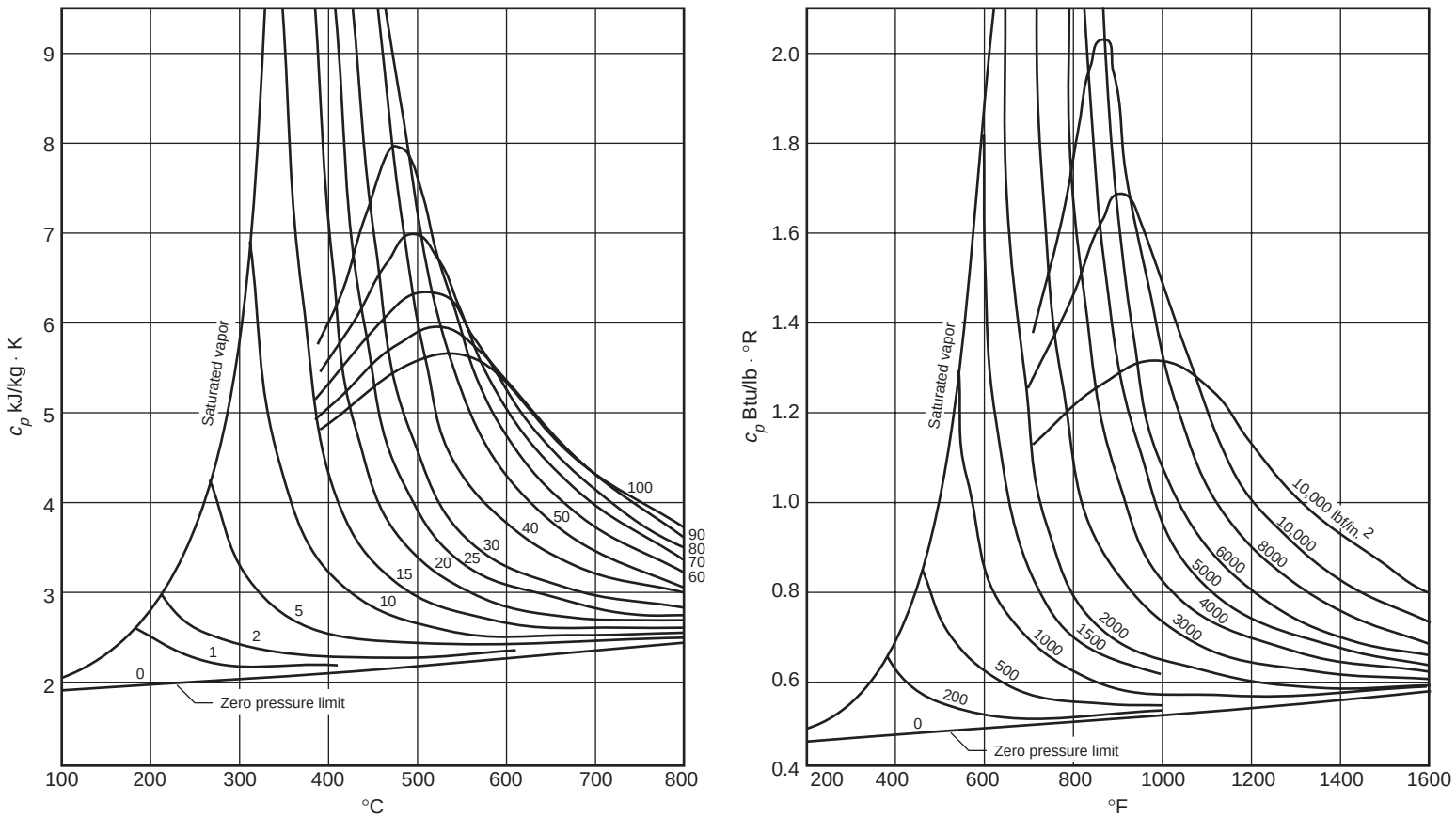


FIGURE 2.4 c_p of water vapor as a function of temperature and pressure. (Adapted from Keenan, J.H., Keyes, F.G., Hill, P.G., and Moore, J.G. 1969 and 1978. *Steam Tables — S.I. Units* (*English Units*). John Wiley & Sons, New York.)

TABLE 2.4 Specific Heat Relations[a]

$$c_v = \left(\frac{\partial u}{\partial T}\right)_v = T\left(\frac{\partial s}{\partial T}\right)_v \quad (1)$$

$$= -T\left(\frac{\partial p}{\partial T}\right)_v \left(\frac{\partial v}{\partial T}\right)_s \quad (2)$$

$$c_p = \left(\frac{\partial h}{\partial T}\right)_p = T\left(\frac{\partial s}{\partial T}\right)_p \quad (3)$$

$$= T\left(\frac{\partial v}{\partial T}\right)_p \left(\frac{\partial p}{\partial T}\right)_s \quad (4)$$

$$c_p - c_v = T\left(\frac{\partial p}{\partial T}\right)_v \left(\frac{\partial v}{\partial T}\right)_p \quad (5)$$

$$= -T\left(\frac{\partial v}{\partial T}\right)_p^2 \left(\frac{\partial p}{\partial v}\right)_T \quad (6)$$

$$= \frac{Tv\beta^2}{\kappa} \quad (7)$$

$$c_p = \frac{1}{\mu_J}\left[T\left(\frac{\partial v}{\partial T}\right)_p - v\right] \quad (8)$$

$$c_v = -\frac{1}{\eta}\left[T\left(\frac{\partial p}{\partial T}\right)_v - p\right] \quad (9)$$

$$k = \frac{c_p}{c_v} = \left(\frac{\partial v}{\partial p}\right)_T \left(\frac{\partial p}{\partial v}\right)_s \quad 10)$$

$$\left(\frac{\partial c_v}{\partial v}\right)_T = T\left(\frac{\partial^2 p}{\partial T^2}\right)_v \quad (11)$$

$$\left(\frac{\partial c_p}{\partial p}\right)_T = -T\left(\frac{\partial^2 v}{\partial T^2}\right)_p \quad (12)$$

[a] See, for example, Moran, M.J. and Shapiro, H.N. 1995. *Fundamentals of Engineering Thermodynamics*, 3rd ed. Wiley, New York, chap. 11.

$$\left(\frac{\partial s}{\partial T}\right)_v = -\frac{1}{(\partial T/\partial v)_s (\partial v/\partial s)_T} = -\left(\frac{\partial v}{\partial T}\right)_s \left(\frac{\partial s}{\partial v}\right)_T$$

Introducing the Maxwell relation from Table 2.2 corresponding to $\psi(T, v)$,

$$\left(\frac{\partial s}{\partial T}\right)_v = -\left(\frac{\partial v}{\partial T}\right)_s \left(\frac{\partial p}{\partial T}\right)_v$$

With this, Equation 2 of Table 2.4 is obtained from Equation 1, which in turn is obtained in Example 6. Equation 11 of Table 2.4 can be obtained by differentiating Equation 1 with repect to specific volume at fixed temperature, and again using the Maxwell relation corresponding to ψ.

P-v-T Relations

Considerable pressure, specific volume, and temperature data have been accumulated for industrially important gases and liquids. These data can be represented in the form $p = f(v, T)$, called an *equation of state*. Equations of state can be expressed in tabular, graphical, and analytical forms.

P-v-T Surface

The graph of a function $p = f(v, T)$ is a surface in three-dimensional space. Figure 2.5 shows the *p-v-T* relationship for water. Figure 2.5b shows the projection of the surface onto the pressure-temperature plane, called the *phase diagram*. The projection onto the *p–v* plane is shown in Figure 2.5c.

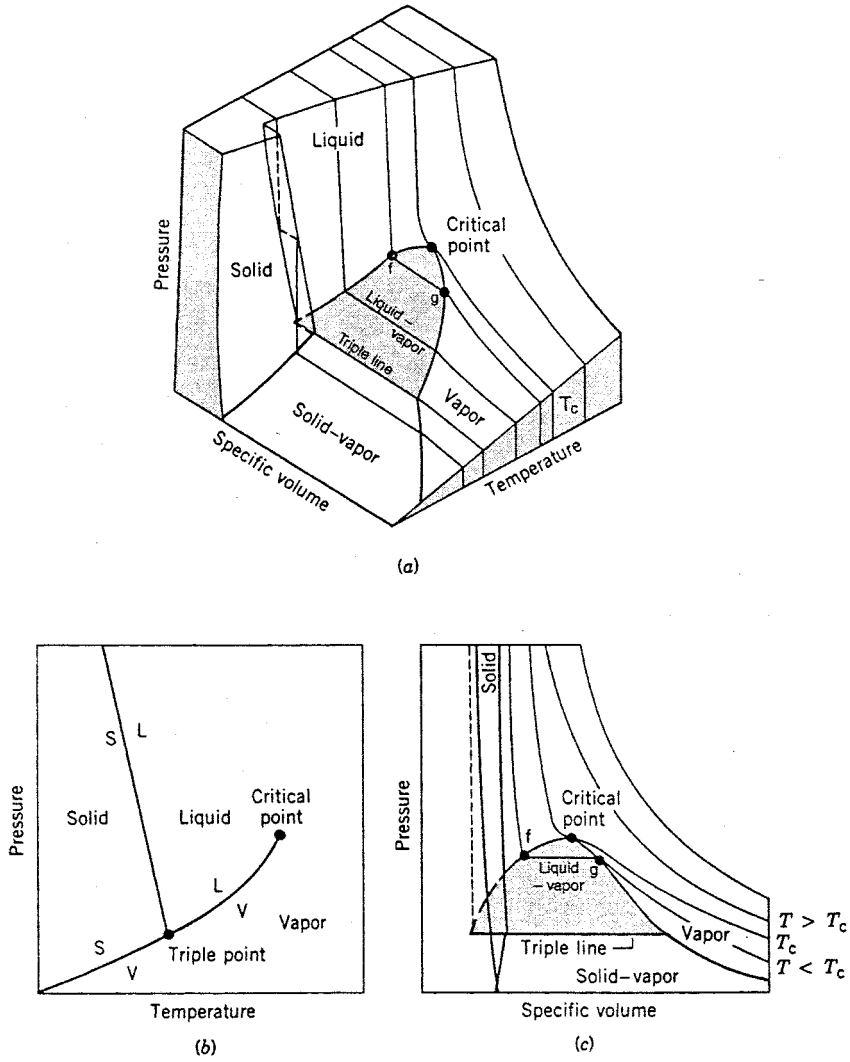


FIGURE 2.5 Pressure-specific volume-temperature surface and projections for water (not to scale).

Figure 2.5 has three regions labeled solid, liquid, and vapor where the substance exists only in a single phase. Between the single phase regions lie *two-phase* regions, where two phases coexist in equilibrium. The lines separating the single-phase regions from the two-phase regions are *saturation lines*. Any state represented by a point on a saturation line is a *saturation state*. The line separating the liquid phase and

the two-phase liquid-vapor region is the saturated liquid line. The state denoted by f is a saturated liquid state. The saturated vapor line separates the vapor region and the two-phase liquid-vapor region. The state denoted by g is a saturated vapor state. The saturated liquid line and the saturated vapor line meet at the *critical point.* At the critical point, the pressure is the *critical pressure* p_c, and the temperature is the *critical temperature* T_c. Three phases can coexist in equilibrium along the line labeled *triple line.* The triple line projects onto a point on the phase diagram. The triple point of water is used in defining the Kelvin temperature scale (Section 2.1, Basic Concepts and Definitions; The Second Law of Thermodynamics, Entropy).

When a phase change occurs during constant pressure heating or cooling, the temperature remains constant as long as both phases are present. Accordingly, in the two-phase liquid-vapor region, a line of constant pressure is also a line of constant temperature. For a specified pressure, the corresponding temperature is called the *saturation temperature.* For a specified temperature, the corresponding pressure is called the *saturation pressure.* The region to the right of the saturated vapor line is known as the *superheated vapor region* because the vapor exists at a temperature greater than the saturation temperature for its pressure. The region to the left of the saturated liquid line is known as the *compressed liquid region* because the liquid is at a pressure higher than the saturation pressure for its temperature.

When a mixture of liquid and vapor coexists in equilibrium, the liquid phase is a saturated liquid and the vapor phase is a saturated vapor. The total volume of any such mixture is $V = V_f + V_g$; or, alternatively, $mv = m_f v_f + m_g v_g$, where m and v denote mass and specific volume, respectively. Dividing by the total mass of the mixture m and letting the *mass fraction* of the vapor in the mixture, m_g/m, be symbolized by x, called the *quality,* the apparent specific volume v of the mixture is

$$\begin{aligned} v &= (1-x)v_{\mathrm{f}} + xv_{\mathrm{g}} \\ &= v_{\mathrm{f}} + xv_{\mathrm{fg}} \end{aligned} \tag{2.37a}$$

where $v_{\mathrm{fg}} = v_{\mathrm{g}} - v_{\mathrm{f}}$. Expressions similar in form can be written for internal energy, enthalpy, and entropy:

$$\begin{aligned} u &= (1-x)u_{\mathrm{f}} + xu_{\mathrm{g}} \\ &= u_{\mathrm{f}} + xu_{\mathrm{fg}} \end{aligned} \tag{2.37b}$$

$$\begin{aligned} h &= (1-x)h_{\mathrm{f}} + xh_{\mathrm{g}} \\ &= h_{\mathrm{f}} + xh_{\mathrm{fg}} \end{aligned} \tag{2.37c}$$

$$\begin{aligned} s &= (1-x)s_{\mathrm{f}} + xs_{\mathrm{g}} \\ &= s_{\mathrm{f}} + xs_{\mathrm{fg}} \end{aligned} \tag{2.37d}$$

For the case of water, Figure 2.6 illustrates the phase change from solid to liquid (melting): *a*-*b*-*c*; from solid to vapor (sublimation): *a*′-*b*′-*c*′; and from liquid to vapor (vaporization): *a*″-*b*″-*c*″. During any such phase change the temperature and pressure remain constant and thus are not independent properties. The *Clapeyron equation* allows the change in enthalpy during a phase change at fixed temperature to be evaluated from *p*-*v*-*T* data pertaining to the phase change. For vaporization, the Clapeyron equation reads

$$\left(\frac{dp}{dT}\right)_{sat} = \frac{h_{\mathrm{g}} - h_{\mathrm{f}}}{T\left(v_{\mathrm{g}} - v_{\mathrm{f}}\right)} \tag{2.38}$$

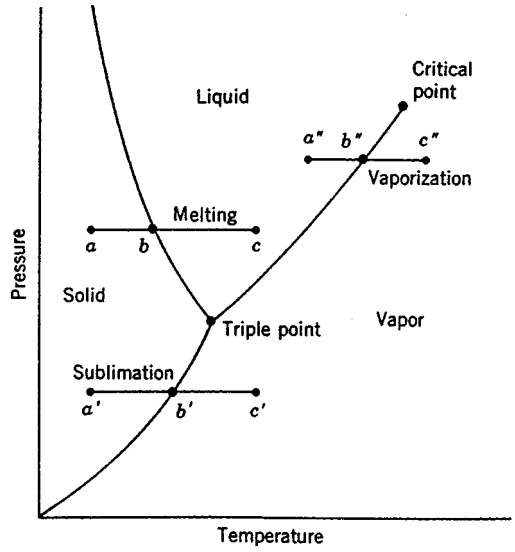


FIGURE 2.6 Phase diagram for water (not to scale).

where $(dp/dT)_{sat}$ is the slope of the saturation pressure-temperature curve at the point determined by the temperature held constant during the phase change. Expressions similar in form to Equation 2.38 can be written for sublimation and melting.

The Clapeyron equation shows that the slope of a saturation line on a phase diagram depends on the signs of the specific volume and enthalpy changes accompanying the phase change. In most cases, when a phase change takes place with an increase in specific enthalpy, the specific volume also increases, and $(dp/dT)_{sat}$ is positive. However, in the case of the melting of ice and a few other substances, the specific volume decreases on melting. The slope of the saturated solid-liquid curve for these few substances is negative, as illustrated for water in Figure 2.6.

Graphical Representations

The intensive states of a pure, simple compressible system can be represented graphically with any two independent intensive properties as the coordinates, excluding properties associated with motion and gravity. While any such pair may be used, there are several selections that are conventionally employed. These include the *p-T* and *p-v* diagrams of Figure 2.5, the *T-s* diagram of Figure 2.7, the *h-s* (Mollier) diagram of Figure 2.8, and the *p-h* diagram of Figure 2.9. The compressibility charts considered next use the compressibility factor as one of the coordinates.

Compressibility Charts

The *p-v-T* relation for a wide range of common gases is illustrated by the generalized compressibility chart of Figure 2.10. In this chart, the compressibility factor, *Z*, is plotted vs. the *reduced* pressure, p_R, *reduced* temperature, T_R, and *pseudoreduced* specific volume, v'_R, where

$$Z = \frac{p\bar{v}}{\bar{R}T} \tag{2.39}$$

and

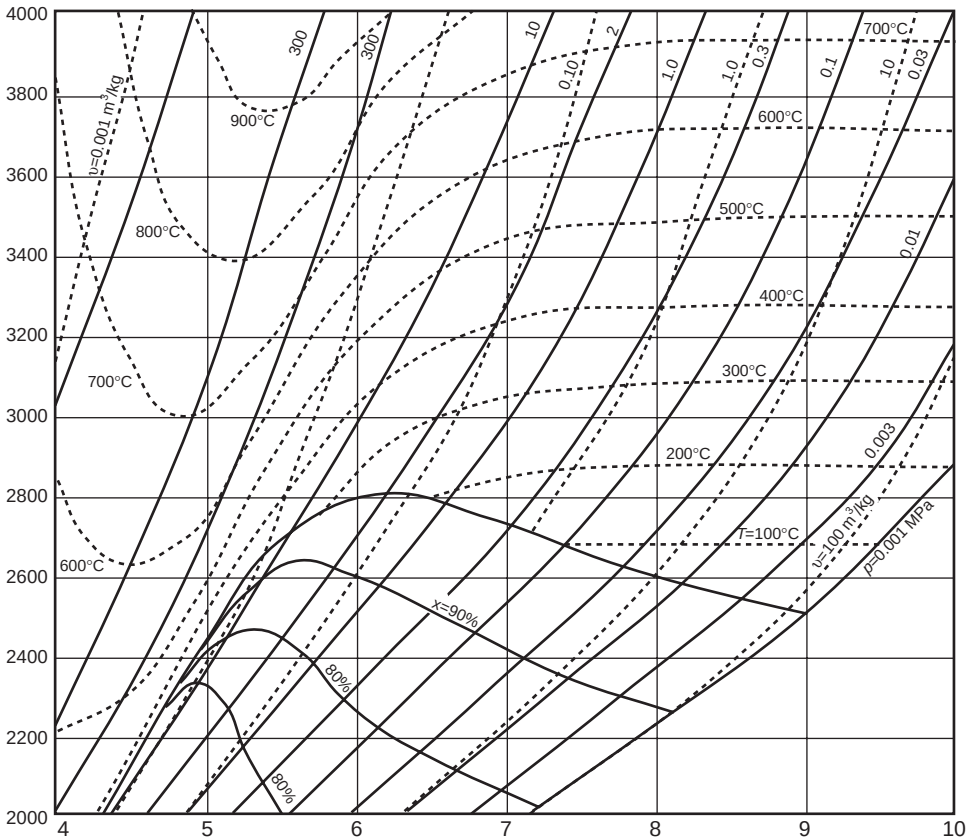


FIGURE 2.7 Temperature-entropy diagram for water. (*Source:* Jones, J.B. and Dugan, R.E. 1996. *Engineering Thermodynamics*, Prentice-Hall, Englewood Cliffs, NJ, based on data and formulations from Haar, L., Gallagher, J.S., and Kell, G.S. 1984. *NBS/NRC Steam Tables*. Hemisphere, Washington, D.C.)

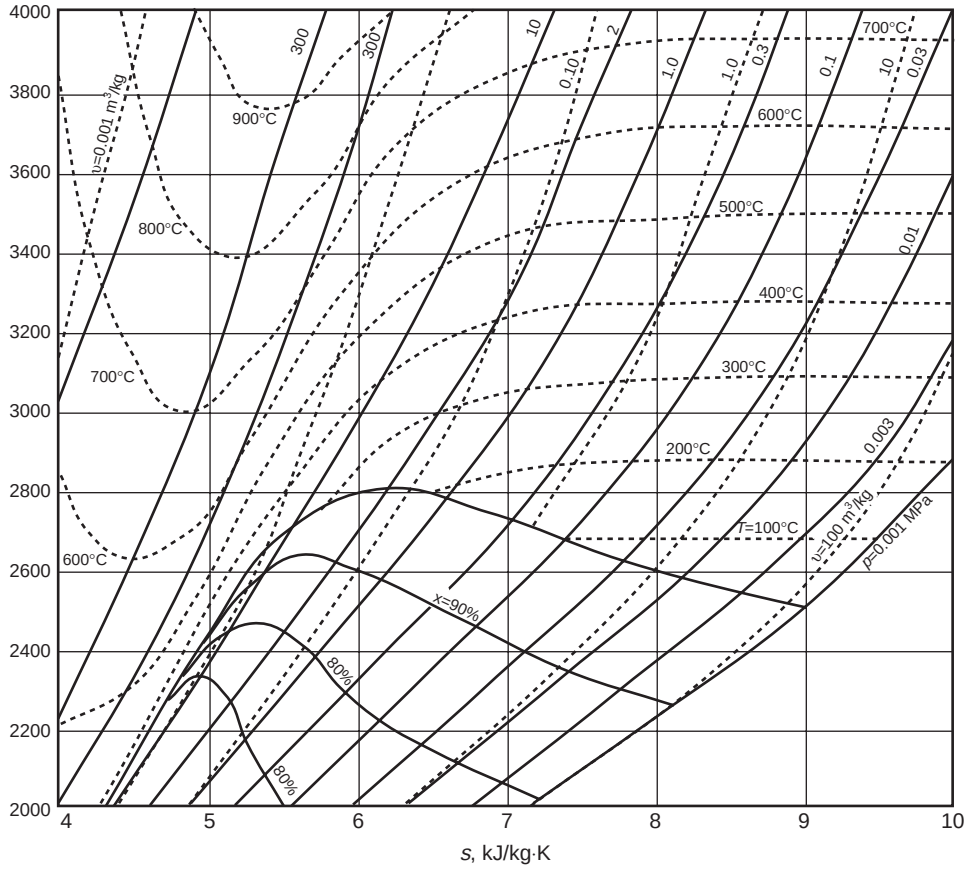


FIGURE 2.8 Enthalpy-entropy (Mollier) diagram for water. (*Source:* Jones, J.B. and Dugan, R.E. 1996. *Engineering Thermodynamics*. Prentice-Hall, Englewood Cliffs, NJ, based on data and formulations from Haar, L., Gallagher, J.S., and Kell, G.S. 1984. *NBS/NRC Steam Tables*. Hemisphere, Washington, D.C.)

$$p_R = \frac{p}{p_c}, \quad T_R = \frac{T}{T_c}, \quad v'_R = \frac{\bar{v}}{\left(\bar{R}T_c/p_c\right)} \tag{2.40}$$

In these expressions, $\bar{R}$ is the universal gas constant and p_c and T_c denote the critical pressure and temperature, respectively. Values of p_c and T_c are given for several substances in Table A.9. The reduced isotherms of Figure 2.10 represent the best curves fitted to the data of several gases. For the 30 gases used in developing the chart, the deviation of observed values from those of the chart is at most on the order of 5% and for most ranges is much less.*

Figure 2.10 gives a common value of about 0.27 for the compressibility factor at the critical point. As the critical compressibility factor for different substances actually varies from 0.23 to 0.33, the chart is inaccurate in the vicinity of the critical point. This source of inaccuracy can be removed by restricting the correlation to substances having essentially the same Z_c values. which is equivalent to including the critical compressibility factor as an independent variable: $Z = f(T_R, p_R, Z_c)$. To achieve greater accuracy

* To determine Z for hydrogen, helium, and neon above a T_R of 5, the reduced temperature and pressure should be calculated using $T_R = T/(T_c + 8)$ and $P_R = p/(p_c + 8)$, where temperatures are in K and pressures are in atm.

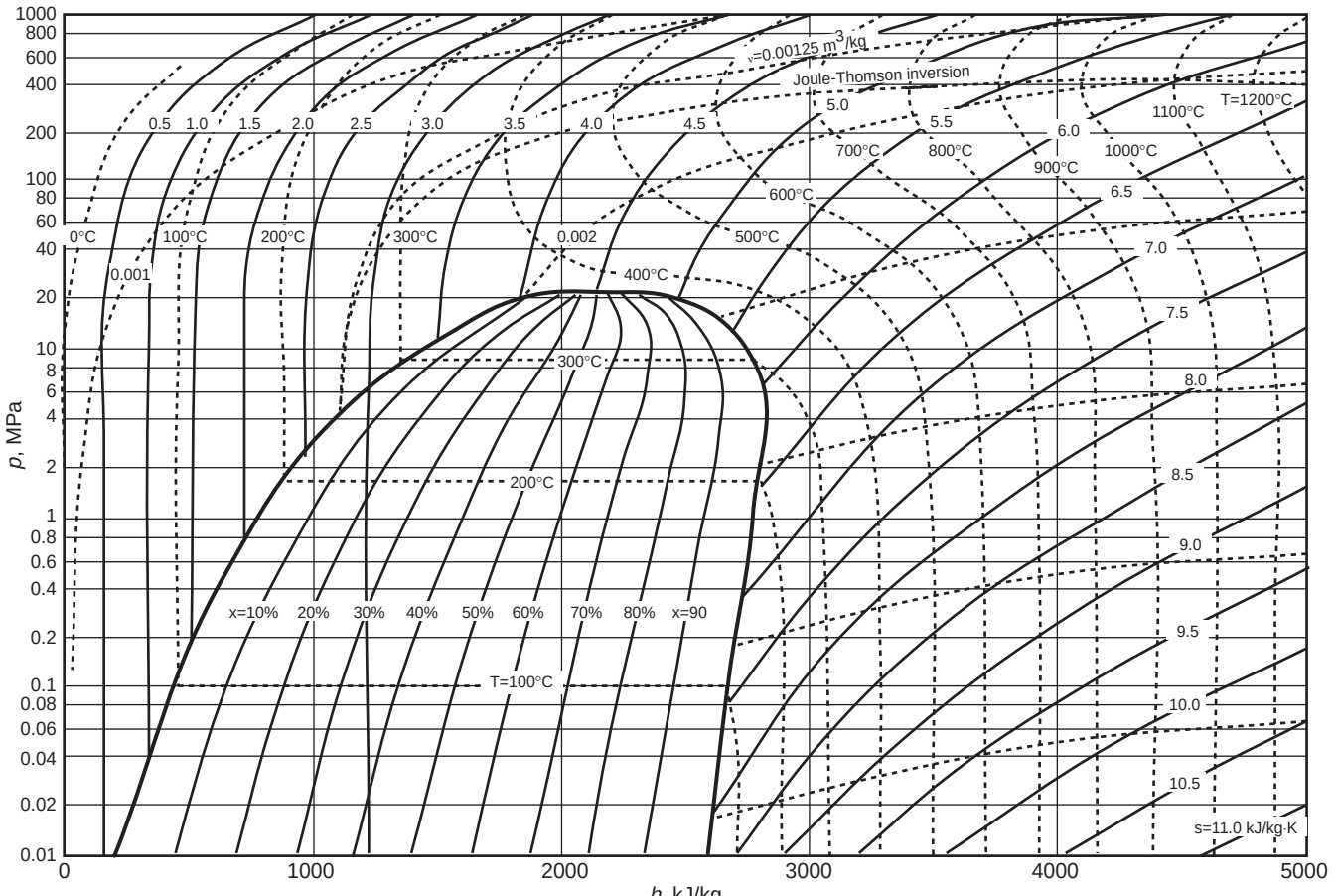


FIGURE 2.9 Pressure-enthalpy diagram for water. (*Source:* Jones, J.B. and Dugan, R.E. 1996. *Engineering Thermodynamics*. Prentice-Hall, Englewood Cliffs, NJ, based on data and formulations from Haar, L., Gallagher, J.S., and Kell, G.S. 1984. *NBS/NRC Steam Tables*. Hemisphere, Washington, D.C.)

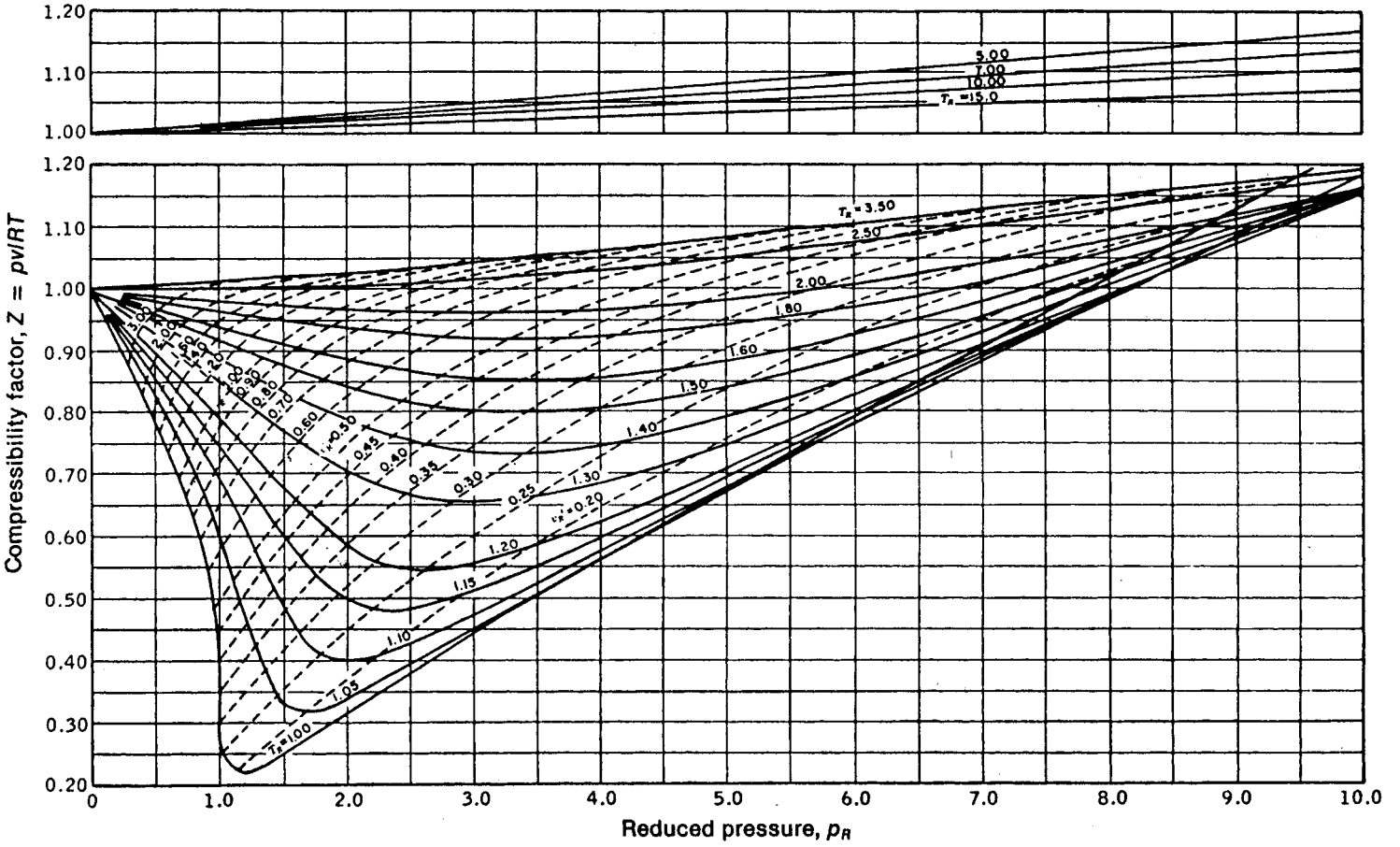


FIGURE 2.10 Generalized compressibility chart ($T_R = T/T_C$, $p_R = p/p_C$, $v'_R = \bar{v}p_C/RT_C$) for $p_R \leq 10$. (*Source:* Obert, E.F. 1960 *Concepts of Thermodynamics*. McGraw-Hill, New York.)

variables other than Z_c have been proposed as a third parameter — for example, the *acentric factor* (see, e.g., Reid and Sherwood, 1966).

Generalized compressibility data are also available in tabular form (see, e.g., Reid and Sherwood, 1966) and in equation form (see, e.g., Reynolds, 1979). The use of generalized data in any form (graphical, tabular, or equation) allows *p*, *v*, and *T* for gases to be evaluated simply and with reasonable accuracy. When accuracy is an essential consideration, generalized compressibility data should not be used as a substitute for *p*-*v*-*T* data for a given substance as provided by computer software, a table, or an equation of state.

Equations of State

Considering the isotherms of Figure 2.10, it is plausible that the variation of the compressibility factor might be expressed as an equation, at least for certain intervals of *p* and *T*. Two expressions can be written that enjoy a theoretical basis. One gives the compressibility factor as an infinite series expansion in pressure,

$$Z = 1 + \hat{B}(T)p + \hat{C}(T)p^2 + \hat{D}(T)p^3 + \ldots$$

and the other is a series in $1/\bar{v}$,

$$Z = 1 + \frac{B(T)}{\bar{v}} + \frac{C(T)}{\bar{v}^2} + \frac{D(T)}{\bar{v}^3} + \ldots$$

These expressions are known as *virial expansions*, and the coefficients $\hat{B}, \hat{C}\ \hat{D}, \ldots$ and $B, C, D \ldots$ are called *virial coefficients*. In principle, the virial coefficients can be calculated using expressions from statistical mechanics derived from consideration of the force fields around the molecules. Thus far only the first few coefficients have been calculated and only for gases consisting of relatively simple molecules. The coefficients also can be found, in principle, by fitting *p*-*v*-*T* data in particular realms of interest. Only the first few coefficients can be found accurately this way, however, and the result is a *truncated* equation valid only at certain states.

Over 100 equations of state have been developed in an attempt to portray accurately the *p*-*v*-*T* behavior of substances and yet avoid the complexities inherent in a full virial series. In general, these equations exhibit little in the way of fundamental physical significance and are mainly empirical in character. Most are developed for gases, but some describe the *p*-*v*-*T* behavior of the liquid phase, at least qualitatively. Every equation of state is restricted to particular states. The realm of applicability is often indicated by giving an interval of pressure, or density, where the equation can be expected to represent the *p*-*v*-*T* behavior faithfully. When it is not stated, the realm of applicability often may be approximated by expressing the equation in terms of the compressibility factor *Z* and the reduced properties, and superimposing the result on a generalized compressibility chart or comparing with compressibility data from the literature.

Equations of state can be classified by the number of adjustable constants they involve. The Redlich-Kwong equation is considered by many to be the best of the two-constant equations of state. It gives pressure as a function of temperature and specific volume and thus is *explicit* in pressure:

$$p = \frac{\bar{R}T}{\bar{v} - b} - \frac{a}{\bar{v}(\bar{v} + b)T^{1/2}} \tag{2.41}$$

This equation is primarily empirical in nature, with no rigorous justification in terms of molecular arguments. Values for the Redlich-Kwong constants for several substances are provided in Table A.9. Modified forms of the equation have been proposed with the aim of achieving better accuracy.

Although the two-constant Redlich-Kwong equation performs better than some equations of state having several adjustable constants, two-constant equations tend to be limited in accuracy as pressure (or density) increases. Increased accuracy normally requires a greater number of adjustable constants. For example, the Benedict-Webb-Rubin equation, which involves eight adjustable constants, has been successful in predicting the *p-v-T* behavior of *light hydrocarbons*. The Benedict-Webb-Rubin equation is also explicit in pressure,

$$p = \frac{\bar{R}T}{\bar{v}} + \left(B\bar{R}T - A - \frac{C}{T^2}\right)\frac{1}{\bar{v}^2} + \frac{\left(b\bar{R}T - a\right)}{\bar{v}^3} + \frac{a\alpha}{\bar{v}^6} + \frac{c}{\bar{v}^3T^2}\left(1 + \frac{\gamma}{\bar{v}^2}\right)\exp\left(-\frac{\gamma}{\bar{v}^2}\right) \qquad (2.42)$$

Values of the Benedict-Webb-Rubin constants for various gases are provided in the literature (see, e.g., Cooper and Goldfrank, 1967). A modification of the Benedict-Webb-Rubin equation involving 12 constants is discussed by Lee and Kessler, 1975. Many multiconstant equations can be found in the engineering literature, and with the advent of high speed computers, equations having 50 or more constants have been developed for representing the *p-v-T* behavior of different substances.

Gas Mixtures

Since an unlimited variety of mixtures can be formed from a given set of pure components by varying the relative amounts present, the properties of mixtures are reported only in special cases such as air. Means are available for predicting the properties of mixtures, however. Most techniques for predicting mixture properties are empirical in character and are not derived from fundamental physical principles. The realm of validity of any particular technique can be established by comparing predicted property values with empirical data. In this section, methods for evaluating the *p-v-T* relations for pure components are adapted to obtain plausible estimates for gas mixtures. The case of ideal gas mixtures is discussed in Section 2.3, Ideal Gas Model. In Section 2.3, Multicomponent Systems, some general aspects of property evaluation for multicomponent systems are presented.

The total number of moles of mixture, n, *is* the sum of the number of moles of the components, n_i:

$$n = n_1 + n_2 + \ldots n_j = \sum_{i=1}^{j} n_i \qquad (2.43)$$

The *relative* amounts of the components present can be described in terms of *mole fractions.* The mole fraction y_i of component i is $y_i = n_i/n$. The sum of the mole fractions of all components present is equal to unity. The apparent molecular weight $\mathcal{M}$ is the mole fraction average of the component molecular weights, such that

$$\mathcal{M} = \sum_{i=1}^{j} y_i \mathcal{M}_i \qquad (2.44)$$

The *relative* amounts of the components present also can be described in terms of *mass fractions:* m_i/m, where m_i is the mass of component i and m is the total mass of mixture.

The *p-v-T* relation for a gas mixture can be estimated by applying an equation of state to the overall mixture. The constants appearing in the equation of state are *mixture values* determined with empirical combining rules developed for the equation. For example, mixture values of the constants a and b for use in the Redlich-Kwong equation are obtained using relations of the form

$$a = \left(\sum_{i=1}^{j} y_i a_i^{1/2}\right)^2, \quad b = \sum_{i=1}^{j} y_i b_i \qquad (2.45)$$

where a_i and b_i are the values of the constants for component i. Combination rules for obtaining mixture values for the constants in other equations of state are also found in the literature.

Another approach is to regard the mixture as if it were a single pure component having critical properties calculated by one of several mixture rules. *Kay's rule is* perhaps the simplest of these, requiring only the determination of a mole fraction averaged critical temperature T_c and critical pressure p_c*:*

$$T_c = \sum_{i=1}^{j} y_i T_{c,i}, \quad p_c = \sum_{i=1}^{j} y_i p_{c,i} \tag{2.46}$$

where $T_{c,i}$ and $p_{c,i}$ are the critical temperature and critical pressure of component i, respectively. Using T_c and p_c, the mixture compressibility factor Z is obtained as for a single pure component. The unkown quantity from among the pressure p, volume V, temperature T, and total number of moles n of the gas mixture can then be obtained by solving $Z = pV/n\bar{R}T$.

Additional means for predicting the p-v-T relation of a mixture are provided by empirical mixture rules. Several are found in the engineering literature. According to the *additive pressure rule*, the pressure of a gas mixture is expressible as a sum of pressures exerted by the individual components:

$$p = p_1 + p_2 + p_3 \ldots \Big]_{T,V} \tag{2.47a}$$

where the pressures p_1, p_2, etc. are evaluated by considering the respective components to be at the volume V and temperature T of the mixture. The additive pressure rule can be expressed alternatively as

$$Z = \sum_{i=1}^{j} y_i Z_i \Big]_{T,V} \tag{2.47b}$$

where Z is the compressibility factor of the mixture and the compressibility factors Z_i are determined assuming that component i occupies the entire volume of the mixture at the temperature T.

The *additive volume rule* postulates that the volume V of a gas mixture is expressible as the sum of volumes occupied by the individual components:

$$V = V_1 + V_2 + V_3 \ldots \Big]_{p,T} \tag{2.48a}$$

where the volumes V_1, V_2, etc. are evaluated by considering the respective components to be at the pressure p and temperature T of the mixture. The additive volume rule can be expressed alternatively as

$$Z = \sum_{i=1}^{j} y_i Z_i \Big]_{p,T} \tag{2.48b}$$

where the compressibility factors Z_i are determined assuming that component i exists at the pressure p and temperature T of the mixture.

Evaluating Δh, Δu, and Δs

Using appropriate specific heat and p-v-T data, the changes in specific enthalpy, internal energy, and entropy can be determined between states of single-phase regions. Table 2.5 provides expressions for such property changes in terms of particular choices of the independent variables: temperature and pressure, and temperature and specific volume.

Taking Equation 1 of Table 2.5 as a representative case, the change in specific enthalpy between states 1 and 2 can be determined using the three steps shown in the accompanying property diagram. This requires knowledge of the variation of c_p with temperature at a fixed pressure p', and the variation of $[v - T(\partial v/\partial T)_p]$ with pressure at temperatures T_1 and T_2:

1-a: Since temperature is constant at T_1, the first integral of Equation 1 in Table 2.5 vanishes, and

$$h_a - h_1 = \int_{p_1}^{p'} \left[v - T\left(\partial v/\partial T\right)_p \right] dp$$

a-b: Since pressure is constant at p', the second integral of Equation 1 vanishes, and

$$h_b - h_a = \int_{T_1}^{T_2} c_p\left(T, p'\right) dT$$

b-2: Since temperature is constant at T_2, the first integral of Equation 1 vanishes, and

$$h_2 - h_b = \int_{p'}^{p_2} \left[v - T\left(\partial v/\partial T\right)_p \right] dp$$

Adding these expressions, the result is $h_2 - h_1$. The required integrals may be performed numerically or analytically. The analytical approach is expedited when an equation of state explicit in specific volume is known.

Similar considerations apply to Equations 2 to 4 of Table 2.5. To evaluate $u_2 - u_1$ with Equation 3, for example, requires the variation of c_v with temperature at a fixed specific volume v', and the variation of $[T(\partial p/\partial T)_v - p]$ with specific volume at temperatures T_1 and T_2. An analytical approach to performing the integrals is expedited when an equation of state explicit in pressure is known.

As changes in specific enthalpy and internal energy are related through $h = u + pv$ by

$$h_2 - h_1 = \left(u_2 - u_1\right) + \left(p_2 v_2 - p_1 v_1\right) \tag{2.49}$$

only one of $h_2 - h_1$ and $u_2 - u_1$ need be found by integration. The other can be evaluated from Equation 2.49. The one found by integration depends on the information available: $h_2 - h_1$ would be found when an equation of state explicit in v and c_p as a function of temperature at some fixed pressure is known, $u_2 - u_1$ would be found when an equation of state explicit in p and c_v as a function of temperature at some specific volume is known.

Example 6

Obtain Equation 1 of Table 2.4 and Equations 3 and 4 of Table 2.5.

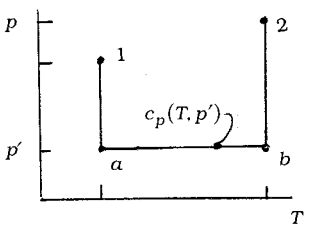
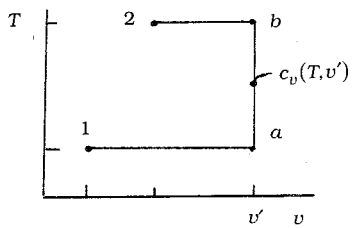
Solution. With Equation 2.33 and the Maxwell relation corresponding to $\psi(T, v)$ from Table 2.2, Equations 3′ and 4′ of Table 2.5 become, respectively,

$$du = c_v dT + \left(\frac{\partial u}{\partial v}\right)_T dv$$

$$ds = \left(\frac{\partial s}{\partial T}\right)_v dT + \left(\frac{\partial p}{\partial T}\right)_v dv$$

Introducing these expressions for du and ds in Equation 2.32a, and collecting terms,

TABLE 2.5 Δh, Δu, Δs Expressions

	temperature and pressure	temperature and specific volume
Independent properties:		
Preferred data:	$v(T,p)$, $c_p(T,\ p)$	$p(T,v)$, $c_v(T,\ v)$
Property diagram:		
Property expressions:		
	$h(T,p)$:	$u(T,\ v)$:
	$dh = \left(\frac{\partial h}{\partial T}\right)_p dT + \left(\frac{\partial h}{\partial p}\right)_T dp$ (1′)	$du = \left(\frac{\partial u}{\partial T}\right)_v dT + \left(\frac{\partial u}{\partial v}\right)_T dv$ (3′)
	$c_p \qquad \left[v - T\left(\frac{\partial v}{\partial T}\right)_p\right]$	$c_v \qquad \left[T\left(\frac{\partial p}{\partial T}\right)_v - p\right]$
	$\Delta h = \int c_p dT + \int \left[v - T\left(\frac{\partial v}{\partial T}\right)_p\right] dp$ (1)	$\Delta u = \int c_v dT + \int \left[T\left(\frac{\partial p}{\partial T}\right)_v - p\right] dv$ (3)
	$s(T,p)$:	$s(T,v)$:
	$ds = \left(\frac{\partial s}{\partial T}\right)_p dT + \left(\frac{\partial s}{\partial p}\right)_T dp$ (2′)	$ds = \left(\frac{\partial s}{\partial T}\right)_v dT + \left(\frac{\partial s}{\partial v}\right)_T dv$ (4′)
	$\frac{c_p}{T} \qquad -\left(\frac{\partial v}{\partial T}\right)_p$	$\frac{c_v}{T} \qquad \left(\frac{\partial p}{\partial T}\right)_v$
	$\Delta s = \int \frac{c_p}{T} dT - \int \left(\frac{\partial v}{\partial T}\right)_p dp$ (2)	$\Delta s = \int \frac{c_v}{T} dT + \int \left(\frac{\partial p}{\partial T}\right)_v dv$ (4)

$$\left[T\left(\frac{\partial s}{\partial T}\right)_v - c_v\right]dT = \left[\left(\frac{\partial u}{\partial v}\right)_T + p - T\left(\frac{\partial p}{\partial T}\right)_v\right]dv$$

Since T and v are independent, the coefficients of dT and dv must vanish, giving, respectively,

$$\left(\frac{\partial s}{\partial T}\right)_v = \frac{c_v}{T}$$

$$\left(\frac{\partial u}{\partial v}\right)_T = T\left(\frac{\partial p}{\partial T}\right)_v - p$$

The first of these corresponds to Equation 1 of Table 2.4 and Equation 4 of Table 2.5. The second of the above expressions establishes Equation 3 of Table 2.5. With similar considerations, Equation 3 of Table 2.4 and Equations 1 and 2 of Table 2.5 may be obtained.

Fundamental Thermodynamic Functions

A fundamental thermodynamic function is one that provides a complete description of the thermodynamic state. The functions $u(s, v)$, $h(s, p)$, $\psi(T, v)$, and $g(T, p)$ listed in Table 2.2 are fundamental thermodynamic functions.

In principle, all properties of interest can be determined from a fundamental thermodynamic function by differentiation and combination. Taking the function $\psi(T, v)$ as a representative case, the properties v and T, being the independent variables, are specified to fix the state. The pressure p and specific entropy s at this state can be determined by differentiation of $\psi(T, v)$, as shown in Table 2.2. By definition, $\psi = u - Ts$, so specific internal energy is obtained as

$$u = \psi + Ts$$

with u, p, and v known, the specific enthalpy can be found from the definition $h = u + pv$. Similarly, the specific Gibbs function is found from the definition $g = h - Ts$. The specific heat c_v can be determined by further differentiation $c_v = (\partial u/\partial T)_v$.

The development of a fundamental function requires the selection of a functional form in terms of the appropriate pair of independent properties and a set of adjustable coefficients that may number 50 or more. The functional form is specified on the basis of both theoretical and practical considerations. The coefficients of the fundamental function are determined by requiring that a set of selected property values and/or observed conditions be statisfied in a least-squares sense. This generally involves property data requiring the assumed functional form to be differentiated one or more times, for example *p-v-T* and specific heat data. When all coefficients have been evaluated, the function is tested for accuracy by using it to evaluate properties for which accepted values are known such as *velocity of sound* and *Joule-Thomson* data. Once a suitable fundamental function is established, extreme accuracy in and consistency among the thermodynamic properties are possible. The properties of water tabulated by Keenan et al. (1969) and by Haar et al. (1984) have been calculated from representations of the Helmholtz function.

Thermodynamic Data Retrieval

Tabular presentations of pressure, specific volume, and temperature are available for practically important gases and liquids. The tables normally include other properties useful for thermodynamic analyses, such as internal energy, enthalpy, and entropy. The various *steam tables* included in the references of this chapter provide examples. Computer software for retrieving the properties of a wide range of substances is also available, as, for example, the ASME Steam Tables (1993) and Bornakke and Sonntag (1996).

Increasingly, textbooks come with computer disks providing thermodynamic property data for water, certain refrigerants, and several gases modeled as ideal gases — see, e.g., Moran and Shapiro (1996).

The sample *steam table* data presented in Table 2.6 are representative of data available for substances commonly encountered in mechanical engineering practice. Table A.5 and Figures 2.7 to 2.9 provide *steam table* data for a greater range of states. The form of the tables and figures, and how they are used are assumed to be familiar. In particular, the use of *linear interpolation* with such tables is assumed known.

Specific internal energy, enthalpy, and entropy data are determined relative to arbitrary datums and such datums vary from substance to substance. Referring to Table 2.6a, the datum state for the specific internal energy and specific entropy of water is seen to correspond to saturated liquid water at 0.01°C (32.02°F), the triple point temperature. The value of each of these properties is set to zero at this state. If calculations are performed involving only differences in a particular specific property, the datum cancels. When there are changes in chemical composition during the process, special care should be exercised. The approach followed when composition changes due to chemical reaction is considered in Section 2.4.

Liquid water data (see Table 2.6d) suggests that at fixed temperature the variation of specific volume, internal energy, and entropy with pressure is slight. The variation of specific enthalpy with pressure at fixed temperature is somewhat greater because pressure is explicit in the definition of enthalpy. This behavior for *v*, *u*, *s*, and *h* is exhibited generally by liquid data and provides the basis for the following set of equations for estimating property data at liquid states from saturated liquid data:

$$v(T,p) \approx v_f(T) \tag{2.50a}$$

$$u(T,p) \approx u_f(T) \tag{2.50b}$$

$$h(T,p) \approx h_f(T) + v_f\underline{\left[p - p_{sat}(T)\right]} \tag{2.50c}$$

$$s(T,p) \approx s_f(T) \tag{2.50d}$$

As before, the subscript f denotes the saturated liquid state at the temperature *T*, and p_{sat} *is* the corresponding saturation pressure. The underlined term of Equation 2.50c is often negligible, giving $h(T, p) \approx h_f(T)$, which is used in Example 3 to evaluate h_1.

In the absence of saturated liquid data, or as an alternative to such data, the *incompressible model* can be employed:

$$\text{Incompressible model: } \begin{cases} v = \text{constant} \\ u = u(T) \end{cases} \tag{2.51}$$

This model is also applicable to solids. Since internal energy varies only with temperature, the specific heat c_v is also a function of only temperature: $c_v(T) = du/dT$. Although specific volume is constant, enthalpy varies with both temperature and pressure, such that

$$h(T,p) = u(T) + pv \tag{2.52}$$

Differentiation of Equation 2.52 with respect to temperature at fixed pressure gives $c_p = c_v$. The common specific heat is often shown simply as *c*. Specific heat and density data for several liquids and solids are

TABLE 2.6 Sample Steam Table Data

(a) Properties of Saturated Water (Liquid-Vapor): Temperature Table

		Specific Volume (m^3/kg)		Internal Energy (kJ/kg)		Enthalpy (kJ/kg)			Entropy (kJ/kg · K)	
Temp (°C)	Pressure (bar)	Saturated Liquid ($v_f \times 10^3$)	Saturated Vapor (v_g)	Saturated Liquid (u_f)	Saturated Vapor (u_g)	Saturated Liquid (h_f)	Evap. (h_{fg})	Saturated Vapor (h_g)	Saturated Liquid (s_f)	Saturated Vapor (s_g)
.01	0.00611	1.0002	206.136	0.00	2375.3	0.01	2501.3	2501.4	0.0000	9.1562
4	0.00813	1.0001	157.232	16.77	2380.9	16.78	2491.9	2508.7	0.0610	9.0514
5	0.00872	1.0001	147.120	20.97	2382.3	20.98	2489.6	2510.6	0.0761	9.0257
6	0.00935	1.0001	137.734	25.19	2383.6	25.20	2487.2	2512.4	0.0912	9.0003
8	0.01072	1.0002	120.917	33.59	2386.4	33.60	2482.5	2516.1	0.1212	8.9501

(b) Properties of Saturated Water (Liquid-Vapor): Pressure Table

		Specific Volume (m^3/kg)		Internal Energy (kJ/kg)		Enthalpy (kJ/kg)			Entropy (kJ/kg · K)	
Pressure (bar)	Temp (°C)	Saturated Liquid ($v_f \times 10^3$)	Saturated Vapor (v_g)	Saturated Liquid (u_f)	Saturated Vapor (u_g)	Saturated Liquid (h_f)	Evap. (h_{fg})	Saturated Vapor (h_g)	Saturated Liquid (s_f)	Saturated Vapor (s_g)
0.04	28.96	1.0040	34.800	121.45	2415.2	121.46	2432.9	2554.4	0.4226	8.4746
0.06	36.16	1.0064	23.739	151.53	2425.0	151.53	2415.9	2567.4	0.5210	8.3304
0.08	41.51	1.0084	18.103	173.87	2432.2	173.88	2403.1	2577.0	0.5926	8.2287
0.10	45.81	1.0102	14.674	191.82	2437.9	191.83	2392.8	2584.7	0.6493	8.1502
0.20	60.06	1.0172	7.649	251.38	2456.7	251.40	2358.3	2609.7	0.8320	7.9085

TABLE 2.6 Sample Steam Table Data (continued)

(c) Properties of Superheated Water Vapor

T(°C)	v(m³/kg)	u(kJ/kg)	h(kJ/kg)	s(kJ/kg · K)	v(m³/kg)	u(kJ/kg)	h(kJ/kg)	s(kJ/kg · K)
	p = 0.06 bar = 0.006 MPa (T_{sat} 36.16°C)				p = 0.35 bar = 0.035 MPa (T_{sat} = 72.69°C)			
Sat.	23.739	2425.0	2567.4	8.3304	4.526	2473.0	2631.4	7.7158
80	27.132	2487.3	2650.1	8.5804	4.625	2483.7	2645.6	7.7564
120	30.219	2544.7	2726.0	8.7840	5.163	2542.4	2723.1	7.9644
160	33.302	2602.7	2802.5	8.9693	5.696	2601.2	2800.6	8.1519
200	36.383	2661.4	2879.7	9.1398	6.228	2660.4	2878.4	8.3237

(d) Properties of Compressed Liquid Water

T(°C)	$v \times 10^3$ (m³/kg)	u(kJ/kg)	h(kJ/kg)	s(kJ/kg · K)	$v \times 10^3$ (m³/kg)	u(kJ/kg)	h(kJ/kg)	s(kJ/kg · K)
	p = 25 bar = 2.5 MPa (T_{sat} 223.99°C)				p = 50 bar = 5.0 MPa (T_{sat} = 263.99°C)			
20	1.0006	83.80	86.30	0.2961	0.9995	83.65	88.65	0.2956
80	1.0280	334.29	336.86	1.0737	1.0268	333.72	338.85	1.0720
140	1.0784	587.82	590.52	1.7369	1.0768	586.76	592.15	1.7343
200	1.1555	849.9	852.8	2.3294	1.1530	848.1	853.9	2.3255
Sat.	1.1973	959.1	962.1	2.5546	1.2859	1147.8	1154.2	2.9202

Source: Moran, M.J. and Shapiro, H.N. 1995. *Fundamentals of Engineering Thermodynamics*, 3rd ed. Wiley, New York, as extracted from Keenan, J. H., Keyes, F.G., Hill, P.G., and Moore, J.G. 1969. *Steam Tables*. Wiley, New York.

provided in Tables B.2, C.1, and C.2. As the variation of c with temperature is slight, c is frequently taken as constant.

When the incompressible model is applied. Equation 2.49 takes the form

$$\begin{aligned} h_2 - h_1 &= \int_{T_1}^{T_2} c(T)dT + v\left(p_2 - p_1\right) \\ &= c_{ave}\left(T_2 - T_1\right) + v\left(p_2 - p_1\right) \end{aligned} \tag{2.53}$$

Also, as Equation 2.32a reduces to $du = Tds$, and $du = c(T)dT$, the change in specific entropy is

$$\begin{aligned} \Delta s &= \int_{T_1}^{T_2} \frac{c(T)}{T} dT \\ &= c_{ave} \ln \frac{T_2}{T_1} \end{aligned} \tag{2.54}$$

Ideal Gas Model

Inspection of the generalized compressibility chart, Figure 2.10, shows that when p_R is small, and for many states when T_R is large, the value of the compressibility factor Z is close to 1. In other words, for pressures that are low relative to p_c, and for many states with temperatures high relative to T_c, the compressibility factor approaches a value of 1. Within the indicated limits, it may be assumed with reasonable accuracy that $Z = 1$ — that is,

$$p\bar{v} = \bar{R}T \quad \text{or} \quad pv = RT \tag{2.55a}$$

where $R = \bar{R}/\mathcal{M}$ is the *specific* gas constant. Other forms of this expression in common use are

$$pV = n\bar{R}T, \quad pV = mRT \tag{2.55b}$$

Referring to Equation 3′ of Table 2.5, it can be concluded that $(\partial u/\partial v)_T$ vanishes identically for a gas whose equation of state is *exactly* given by Equation 2.55, and thus the specific internal energy depends only on temperature. This conclusion is supported by experimental observations beginning with the work of Joule, who showed that the internal energy of air at low density depends primarily on temperature.

These considerations allow for an *ideal gas model* of each real gas: (1) the equation of state is given by Equation 2.55 and (2) the internal energy and enthalpy are functions of temperature alone. The real gas approaches the model in the limit of low reduced pressure. At other states the actual behavior may depart substantially from the predictions of the model. Accordingly, caution should be exercised when invoking the ideal gas model lest significant error is introduced.

Specific heat data for gases can be obtained by direct measurement. When extrapolated to zero pressure, ideal gas-specific heats result. Ideal gas-specific heats also can be calculated using molecular models of matter together with data from spectroscopic measurements. Table A.9 provides ideal gas-specific heat data for a number of substances. The following ideal gas-specific heat relations are frequently useful:

$$c_p(T) = c_v(T) + R \tag{2.56a}$$

$$c_p = \frac{kR}{k-1}, \quad c_v = \frac{R}{k-1} \tag{2.56b}$$

where $k = c_p/c_v$.

With the ideal gas model, Equations 1 to 4 of Table 2.5 give Equations 1 to 4 of Table 2.7, respectively. Equation 2 of Table 2.7 can be expressed alternatively using $s°(T)$ defined by

$$s°(T) \equiv \int_0^T \frac{c_p(T)}{T} dT \tag{2.57}$$

as

$$s(T_2, p_2) - s(T_1, p_1) = s°(T_2) - s°(T_1) - R \ln \frac{p_2}{p_1} \tag{2.58}$$

Expressions similar in form to Equations 2.56 to 2.68 can be written on a molar basis.

TABLE 2.7 Ideal Gas Expressions for Δh, Δu, and Δs

Variable Specific Heats		Constant Specific Heats	
$h(T_2) - h(T_1) = \int_{T_1}^{T_2} c_p(T) dT$	(1)	$h(T_2) - h(T_1) = c_p(T_2 - T_1)$	(1′)
$s(T_2, p_2) - s(T_1, p_1) = \int_{T_1}^{T_2} \frac{c_p(T)}{T} dT - R \ln \frac{p_2}{p_1}$	(2)	$s(T_2, p_2) - s(T_1, p_1) = c_p \ln \frac{T_2}{T_1} - R \ln \frac{p_2}{p_1}$	(2′)
$u(T_2) - u(T_1) = \int_{T_1}^{T_2} c_v(T) dT$	(3)	$u(T_2) - u(T_1) = c_v(T_2 - T_1)$	(3′)
$s(T_2, v_2) - s(T_1, v_1) = \int_{T_1}^{T_2} \frac{c_v(T)}{T} dT + R \ln \frac{v_2}{v_1}$	(4)	$s(T_2, v_2) - s(T_1, v_1) = c_v \ln \frac{T_2}{T_1} + R \ln \frac{v_2}{v_1}$	(4′)
$s_2 = s_1$		$s_2 = s_1$	
$\frac{p_r(T_2)}{p_r(T_1)} = \frac{p_2}{p_1}$	(5)	$\frac{T_2}{T_1} = \left(\frac{p_2}{p_1}\right)^{(k-1)/k}$	(5′)
$\frac{v_r(T_2)}{v_r(T_1)} = \frac{v_2}{v_1}$	(6)	$\frac{T_2}{T_1} = \left(\frac{v_1}{v_2}\right)^{k-1}$	(6′)

For processes of an ideal gas between states having the same specific entropy, $s_2 = s_1$, Equation 2.58 gives

$$\frac{p_2}{p_1} = \frac{\exp\left[s°(T_2)/R\right]}{\exp\left[s°(T_1)/R\right]}$$

or with $p_r = \exp[s°(T)/R]$

$$\frac{p_2}{p_1} = \frac{p_r(T_2)}{p_r(T_1)} \qquad (s_2 = s_1) \tag{2.59a}$$

A relation between the specific volume and temperatures for two states of an ideal gas having the same specific entropy can also be developed:

$$\frac{v_2}{v_1} = \frac{v_r(T_2)}{v_r(T_1)} \qquad (s_2 = s_1) \tag{2.59b}$$

Equations 2.59 are listed in Table 2.7 as Equations 5 and 6, respectively.

Table A.8 provides a tabular display of h, u, $s°$, p_r, and v_r vs. temperature for air as an ideal gas. Tabulations of $\bar{h}$, $\bar{u}$, and $\bar{s}°$ for several other common gases are provided in Table A.2. Property retrieval software also provides such data; see, e.g., Moran and Shapiro (1996). The use of data from Table A.8 for the nozzle of Example 2 is illustrated in Example 7.

When the ideal gas-specific heats are assumed constant, Equations 1 to 6 of Table 2.7 become Equations 1′ to 6′, respectively. The specific heat c_p is taken as constant in Example 2.

Example 7

Using data from Table A.8, evaluate the exit velocity for the nozzle of Example 2 and compare with the exit velocity for an isentropic expansion to 15 lbf/in.^{2}.

Solution. The exit velocity is given by Equation 2.27f

$$\mathrm{v}_e = \sqrt{\mathrm{v}_i^2 + 2(h_i - h_e)}$$

At 960 and 520°R, Table A.8 gives, respectively, h_i = 231.06 Btu/lb and h_e = 124.27 Btu/lb. Then

$$\mathrm{v}_e = \sqrt{\left(\frac{10\ \mathrm{ft}}{s}\right)^2 + 2(231.06 - 124.27)\left(\frac{\mathrm{Btu}}{\mathrm{lb}}\right)\left(\frac{778.17\ \mathrm{ft \cdot lbf}}{1\ \mathrm{Btu}}\right)\left(\frac{32.174\ \mathrm{lb \cdot ft/sec^2}}{1\ \mathrm{lbf}}\right)}$$

$$= 2312.5\ \mathrm{ft\ sec}$$

Using Equation 2.59a and p_r data from Table A.8, the specific enthalpy at the exit for an isentropic expansion is found as follows:

$$p_r(T_e) = p_r(T_i)\frac{p_e}{p_i} = 10.61\left(\frac{15}{150}\right) = 1.061$$

Interpolating with p_r data, h_e = 119.54 Btu/lb. With this, the exit velocity is 2363.1 ft/sec. The actual exit velocity is about 2% less than the velocity for an isentropic expansion, the maximum theoretical value. In this particular application, there is good agreement in each case between velocities calculated using Table A.8 data and, as in Example 2, assuming c_p constant. Such agreement cannot be expected generally, however. See, for example, the Brayton cycle data of Table 2.15.

Polytropic Processes

An internally reversible process described by the expression pv^n = constant is called a *polytropic process* and *n is* the *polytropic exponent.* Although this expression can be applied with real gas data, it most generally appears in practice together with the use of the ideal gas model. Table 2.8 provides several expressions applicable to polytropic processes and the special forms they take when the ideal gas model is assumed. The expressions for $\int p dv$ and $\int v dp$ have application to work evaluations with Equations 2.10 and 2.30, respectively. In some applications it may be appropriate to determine n by fitting pressure-specific volume data.

Example 8 illustrates both the polytropic process and the reduction in the compressor work achievable by cooling a gas as it is compressed.

Example 8

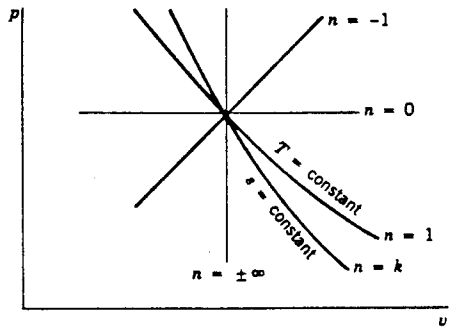
A compressor operates at steady state with air entering at 1 bar, 20°C and exiting at 5 bar. (a) If the air undergoes a polytropic process with n = 1.3, determine the work and heat transfer, each in kJ/kg of air flowing. Repeat for (b) an isothermal compression and (c) an isentropic compression.

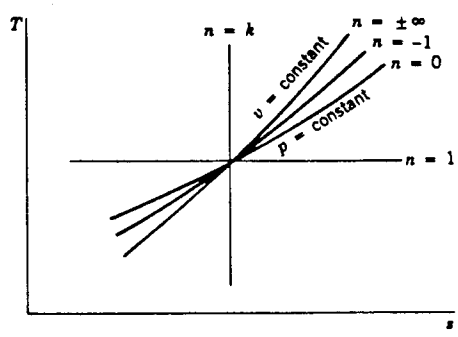
TABLE 2.8 Polytropic Processes: pv^n = Constant[a]

General		Ideal Gas[b]	
$\dfrac{p_2}{p_1}=\left(\dfrac{v_1}{v_2}\right)^n$	(1)	$\dfrac{p_2}{p_1}=\left(\dfrac{v_1}{v_2}\right)^n=\left(\dfrac{T_2}{T_1}\right)^{n/(n-1)}$	(1′)
$n = 0$: constant pressure $n = \pm\infty$: constant specific volume		$n = 0$: constant pressure $n = \pm\infty$: constant specific volume $n = 1$: constant temperature $n = k$: constant specific entropy when k is constant	
$n = 1$		$n = 1$	
$\displaystyle\int_1^2 p\,dv = p_1 v_1 \ln\frac{v_2}{v_1}$	(2)	$\displaystyle\int_1^2 p\,dv = RT\ln\frac{v_2}{v_1}$	(2′)
$\displaystyle -\int_1^2 v\,dp = -p_1 v_1 \ln\frac{p_2}{p_1}$	(3)	$\displaystyle -\int_1^2 v\,dp = -RT\ln\frac{p_2}{p_1}$	(3′)
$n \neq 1$		$n \neq 1$	
$\displaystyle\int_1^2 p\,dv = \frac{p_2v_2 - p_1v_1}{1-n} = \frac{p_1v_1}{n-1}\left[1-\left(\frac{p_2}{p_1}\right)^{(n-1)/n}\right]$	(4)	$\displaystyle\int_1^2 p\,dv = \frac{R(T_2 - T_1)}{1-n} = \frac{RT_1}{n-1}\left[1-\left(\frac{p_2}{p_1}\right)^{(n-1)/n}\right]$	(4′)
$\displaystyle -\int_1^2 v\,dp = \frac{n}{1-n}(p_2v_2 - p_1v_1) = \frac{np_1v_1}{n-1}\left[1-\left(\frac{p_2}{p_1}\right)^{(n-1)/n}\right]$	(5)	$\displaystyle -\int_1^2 v\,dp = \frac{nR}{1-n}(T_2 - T_1) = \frac{nRT_1}{n-1}\left[1-\left(\frac{p_2}{p_1}\right)^{(n-1)/n}\right]$	(5′)

[a] For polytropic processes of closed systems where volume change is the only work mode, Equations 2, 4, and 2′, 4′ are applicable with Equation 2.10 to evaluate the work. When each unit of mass passing through a one-inlet, one-exit control volume at steady state undergoes a polytropic process, Equations 3, 5, and 3′, 5′ are applicable with Equations 2.30a and 2.30b to evaluate the power. Also note that generally, $-\int_1^2 v\,dp = n\int_1^2 p\,dv$.

[b]





Solution. Using Equation 5′ of Table 2.8 together with Equation 2.30b,

$$\frac{\dot{W}_{cv}}{\dot{m}} = \frac{nRT_1}{n-1}\left[1-\left(\frac{p_2}{p_1}\right)^{(n-1)/n}\right]$$

$$= \left(\frac{1.3}{0.3}\right)\left(\frac{8.314}{28.97}\frac{\text{kJ}}{\text{kg}\cdot\text{K}}\right)(293\ \text{K})\left[1-(5)^{0.3/1.3}\right]$$

$$= -163.9\ \frac{\text{kJ}}{\text{kg}}$$

(The area behind process 1-2 of Figure 2.11, area 1-2-a-b, represents the magnitude of the work required, per unit mass of air flowing.) Also, Equation 1′ of Table 2.8 gives T_2 = 425 K.

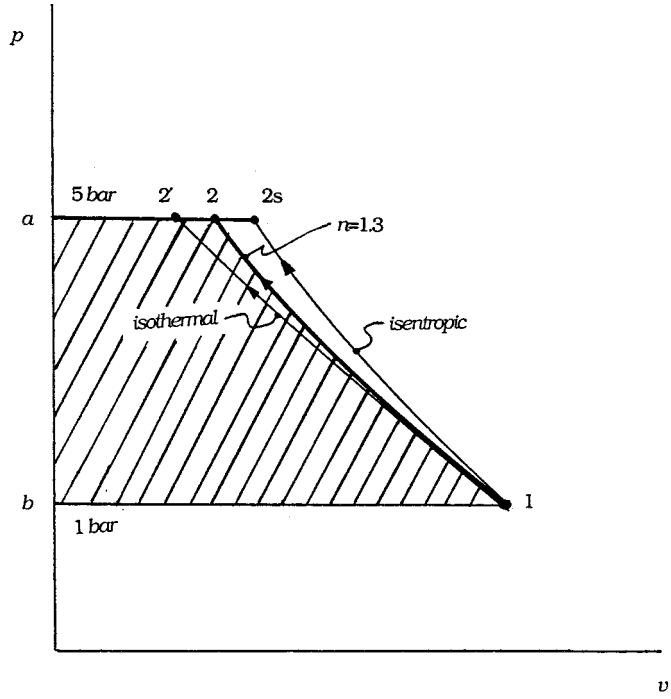


FIGURE 2.11 Internally reversible compression processes.

An energy rate balance at steady state and enthalpy data from Table A.8 gives

$$\frac{\dot{Q}_{cv}}{\dot{m}} = \frac{\dot{W}_{cv}}{\dot{m}} + h_2 - h_1$$

$$= -163.9 + (426.3 - 293.2) = -30.8\frac{\text{kJ}}{\text{kg}}$$

(b) Using Equation 3′ of Table 2.8 together with Equation 2.30b,

$$\frac{\dot{W}_{cv}}{\dot{m}} = -RT\ln\frac{p_2}{p_1}$$

$$= -\left(\frac{8.314}{28.97}\right)(293)\ln 5$$

$$= -135.3\frac{\text{kJ}}{\text{kg}}$$

Area 1-2′-*a*-*b* on Figure 2.11 represents the magnitude of the work required, per unit of mass of air flowing. An energy balance reduces to give $\dot{Q}_{cv}/\dot{m} = \dot{W}_{cv}/\dot{m} = -135.3$ kJ/kg. (c) For an isentropic compression, $\dot{Q}_{cv} = 0$ and an energy rate balance reduces to give $\dot{W}_{cv}/\dot{m} = -(h_{2s} - h_1)$, where 2*s* denotes the exit state. With Equation 2.59a and p_r data, $h_{2s} = 464.8$ kJ/kg ($T_{2s} = 463$K). Then $\dot{W}_{cv}/\dot{m} = -(464.8 - 293.2) = -171.6$ kJ/kg. Area 1-2s-*a*-*b* on Figure 2.11 represents the magnitude of the work required, per unit of mass of air flowing.

Ideal Gas Mixtures

When applied to an ideal gas mixture, the additive pressure rule (Section 2.3, *p*-*v*-*T* Relations) is known as the *Dalton model*. According to this model, each gas in the mixture acts as if it exists separately at the volume and temperature of the mixture. Applying the ideal gas equation of state to the mixture as a whole and to each component *i*, $pV = n\bar{R}T$, $p_iV = n_i\bar{R}T$, where p_i, the *partial pressure* of component *i*, is the pressure that component *i* would exert if n_i moles occupied the full volume *V* at the temperature *T*. Forming a ratio, the partial pressure of component *i* is

$$p_i = \frac{n_i}{n}p = y_i p \tag{2.60}$$

where y_i is the mole fraction of component *i*. The sum of the partial pressures equals the mixture pressure.

The internal energy, enthalpy, and entropy of the mixture can be determined as the sum of the respective properties of the component gases, provided that the contribution from each gas is evaluated at the condition at which the gas exists in the mixture. On a *molar* basis,

$$U = \sum_{i=1}^{j} n_i\bar{u}_i \quad \text{or} \quad \bar{u} = \sum_{i=1}^{j} y_i\bar{u}_i \tag{2.61a}$$

$$H = \sum_{i=1}^{j} n_i\bar{h}_i \quad \text{or} \quad \bar{h} = \sum_{i=1}^{j} y_i\bar{h}_i \tag{2.61b}$$

$$S = \sum_{i=1}^{j} n_i\bar{s}_i \quad \text{or} \quad \bar{s} = \sum_{i=1}^{j} y_i\bar{s}_i \tag{2.61c}$$

The specific heats $\bar{c}_v$ and $\bar{c}_p$ for an ideal gas mixture in terms of the corresponding specific heats of the components are expressed similarly:

$$\bar{c}_v = \sum_{i=1}^{j} y_i\bar{c}_{vi} \tag{2.61d}$$

$$\bar{c}_p = \sum_{i=1}^{j} y_i \bar{c}_{pi} \tag{2.61e}$$

When working on a *mass* basis, expressions similar in form to Equations 2.61 can be written using *mass* and *mass fractions* in place of *moles* and *mole fractions*, respectively, and using u, h, s, c_p, and c_v in place of $\bar{u}$, $\bar{h}$, $\bar{s}$, $\bar{c}_p$, and $\bar{c}_v$, respectively.

The internal energy and enthalpy of an ideal gas depend only on temperature, and thus the $\bar{u}_i$ and $\bar{h}_i$ terms appearing in Equations 2.61 are evaluated at the temperature of the mixture. Since entropy depends on *two* independent properties, the $\bar{s}_i$ terms are evaluated either at the temperature and the partial pressure p_i of component i, or at the temperature and volume of the mixture. In the former case

$$\begin{aligned} S &= \sum_{i=1}^{j} n_i \bar{s}_i(T, p_i) \\ &= \sum_{i=1}^{j} n_i \bar{s}_i(T, x_i p) \end{aligned} \tag{2.62}$$

Inserting the expressions for H and S given by Equations 2.61b and 2.61c into the Gibbs function, $G = H - TS$,

$$\begin{aligned} G &= \sum_{i=1}^{j} n_i \bar{h}_i(T) - T\sum_{i=1}^{j} n_i \bar{s}_i(T, p_i) \\ &= \sum_{i=1}^{j} n_i \bar{g}_i(T, p_i) \end{aligned} \tag{2.63}$$

where the molar-specific Gibbs function of component i is $g_i(T, p_i) = h_i(T) - Ts_i(T, p_i)$. The Gibbs function of i can be expressed alternatively as

$$\begin{aligned} \bar{g}_i(T, p_i) &= \bar{g}_i(T, p') + \bar{R}T \ln(p_i / p') \\ &= \bar{g}_i(T, p') + \bar{R}T \ln(x_i p / p') \end{aligned} \tag{2.64}$$

were p' is some specified pressure. Equation 2.64 is obtained by integrating Equation 2.32d at fixed temperature T from pressure p' to p_i.

Moist Air

An ideal gas mixture of particular interest for many practical applications is *moist air.* Moist air refers to a mixture of dry air and water vapor in which the dry air is treated as if it were a pure component. Ideal gas mixture principles usually apply to moist air. In particular, the *Dalton model is* applicable, and so the mixture pressure p is the sum of the partial pressures p_a and p_v of the dry air and water vapor, respectively.

Saturated air is a mixture of dry air and saturated water vapor. For saturated air, the partial pressure of the water vapor equals $p_{sat}(T)$, which is the saturation pressure of water corresponding to the dry-bulb (mixture) temperature T. The makeup of moist air can be described in terms of the *humidity ratio* (*specific humidity*) and the *relative humidity.* The bulb of a *wet-bulb thermometer* is covered with a wick saturated with liquid water, and the *wet-bulb* temperature of an air-water vapor mixture is the temperature indicated by such a thermometer exposed to the mixture.

When a sample of moist air is cooled at constant pressure, the temperature at which the sample becomes saturated is called the *dew point temperature*. Cooling below the dew point temperature results in the condensation of some of the water vapor initially present. When cooled to a final equilibrium state at a temperature below the dew point temperature, the original sample would consist of a gas phase of dry air and saturated water vapor in equilibrium with a liquid water phase.

Psychrometric charts are plotted with various moist air parameters, including the dry-bulb and wet-bulb temperatures, the humidity ratio, and the relative humidity, usually for a specified value of the mixture pressure such as 1 atm. Further discussion of moist air and related psychrometric principles and applications is provided in Chapter 9.

Generalized Charts for Enthalpy, Entropy, and Fugacity

The changes in enthalpy and entropy between two states can be determined in principle by correcting the respective property change determined using the ideal gas model. The corrections can be obtained, at least approximately, by inspection of the generalized enthalpy correction and entropy correction charts, Figures 2.12 and 2.13, respectively. Such data are also available in tabular form (see, e.g., Reid and Sherwood, 1966) and calculable using a generalized equation for the compressibility factor (Reynolds, 1979). Using the superscript * to identify ideal gas property values, the changes in specific enthalpy and specific entropy between states 1 and 2 are

$$\bar{h}_2 - \bar{h}_1 = \underline{\bar{h}_2^* - \bar{h}_1^*} - \underline{\bar{R}T_c\left[\left(\frac{\bar{h}^* - \bar{h}}{\bar{R}T_c}\right)_2 - \left(\frac{\bar{h}^* - \bar{h}}{\bar{R}T_c}\right)_1\right]} \tag{2.65a}$$

$$\bar{s}_2 - \bar{s}_1 = \underline{\bar{s}_2^* - \bar{s}_1^*} - \underline{\bar{R}\left[\left(\frac{\bar{s}^* - \bar{s}}{\bar{R}}\right)_2 - \left(\frac{\bar{s}^* - \bar{s}}{\bar{R}}\right)_1\right]} \tag{2.65b}$$

The first underlined term on the right side of each expression represents the respective property change assuming ideal gas behavior. The second underlined term is the correction that must be applied to the ideal gas value to obtain the actual value. The quantities $(\bar{h}^* - \bar{h})/\bar{R}T_c$ *and* $(\bar{s}^* - \bar{s})/\bar{R}$ at state 1 would be read from the respective correction chart or table or calculated, using the reduced temperature T_{R1} and reduced pressure p_{R1} corresponding to the temperature T_1 and pressure p_1 at state 1, respectively. Similarly, $(\bar{h}^* - \bar{h})/\bar{R}T_c$ *and* $(\bar{s}^* - \bar{s})/\bar{R}$ at state 2 would be obtained using T_{R2} and p_{R2}. Mixture values for T_c and p_c determined by applying Kay's rule or some other mixture rule also can be used to enter the generalized enthalpy correction and entropy correction charts.

Figure 2.14 gives the *fugacity* coefficient, *f*/*p*, as a function of reduced pressure and reduced temperature. The fugacity *f* plays a similar role in determining the specific Gibbs function for a real gas as pressure plays for the ideal gas. To develop this, consider the variation of the specific Gibbs function with pressure at fixed temperature (from Table 2.2)

$$\left.\frac{\partial g}{\partial p}\right)_T = v$$

For an ideal gas, integration at fixed temperature gives

$$g^* = RT\ln p + C(T)$$

where $C(T)$ is a function of integration. To evaluate g for a real gas, fugacity replaces pressure,

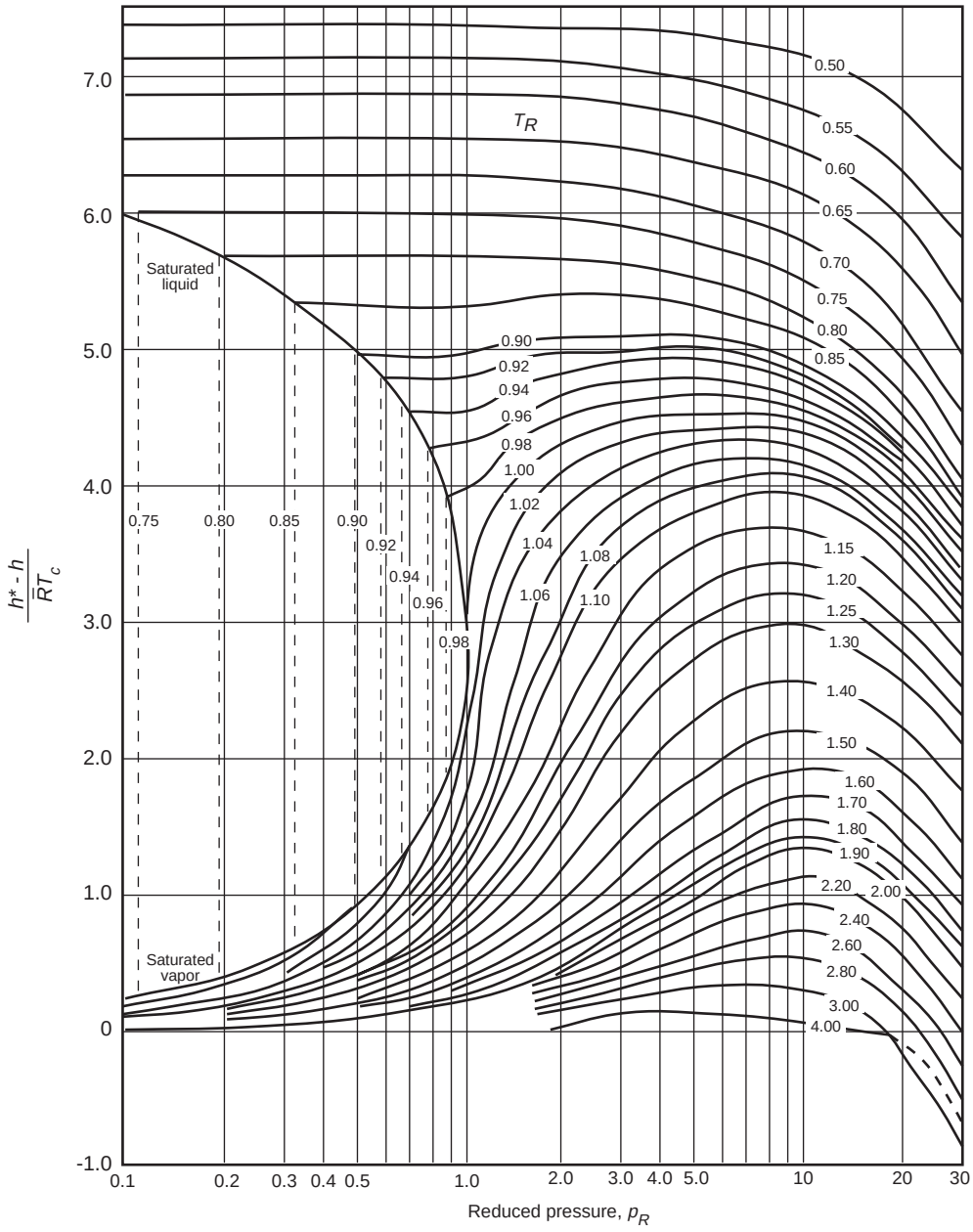


FIGURE 2.12 Generalized enthalpy correction chart. (*Source:* Adapted from Van Wylen, G. J. and Sonntag, R. E. 1986. *Fundamentals of Classical Thermodynamics*, 3rd ed., English/SI. Wiley, New York.)

$$g = RT \ln f + C(T)$$

In terms of the fugacity coefficient the departure of the real gas value from the ideal gas value at fixed temperature is then

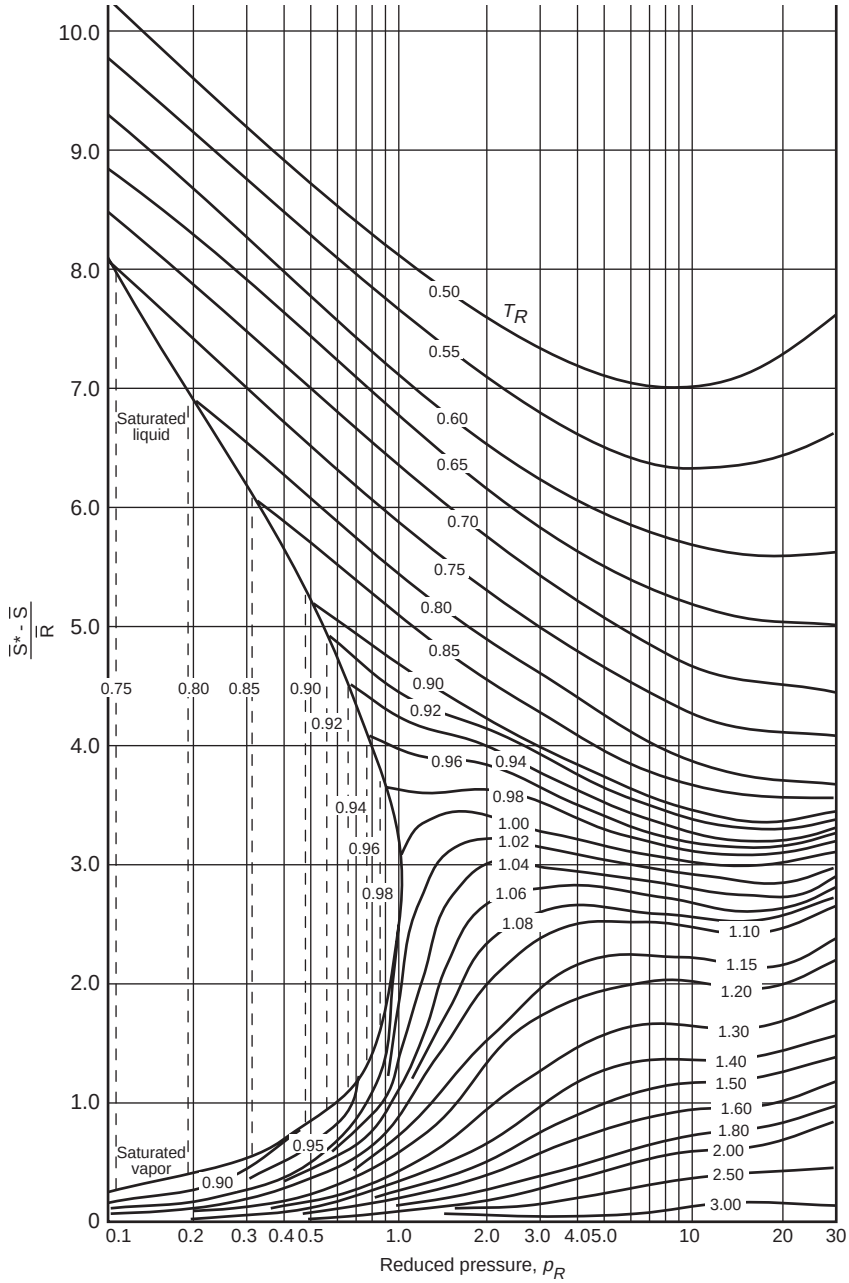


FIGURE 2.13 Generalized entropy correction chart. (*Source:* Adapted from Van Wylen, G. J. and Sonntag, R. E. 1986. *Fundamentals of Classical Thermodynamics*, 3rd ed., English/SI. Wiley, New York.)

$$g - g^* = RT \ln \frac{f}{p} \tag{2.66}$$

As pressure is reduced at fixed temperature, *f*/*p* tends to unity, and the specific Gibbs function is given by the ideal gas value.

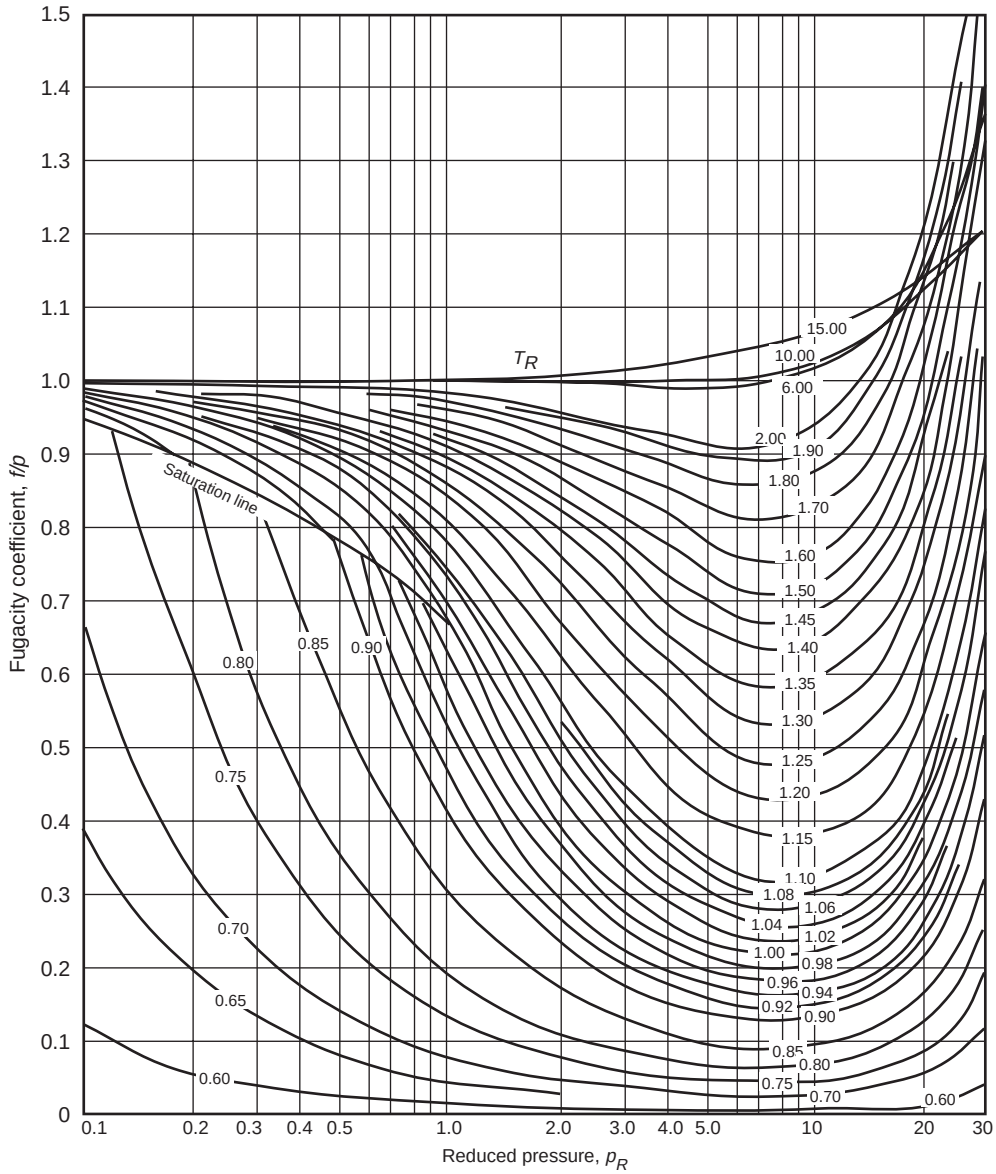


FIGURE 2.14 Generalized fugacity coefficient chart. (*Source:* Van Wylen, G. J. and Sonntag, R. E. 1986. *Fundamentals of Classical Thermodynamics*, 3rd ed., English/SI. Wiley, New York.)

Multicomponent Systems

In this section are presented some general aspects of the properties of multicomponent systems consisting of nonreacting mixtures. For a single phase *multicomponent* system consisting of j components, an extensive property X may be regarded as a function of temperature, pressure, and the number of moles of each component present in the mixture: $X = X(T, p, n_1, n_2, \ldots n_j)$. Since X is mathematically *homogeneous of degree one* in the n's, the function is expressible as

$$X = \sum_{i=1}^{j} n_i \overline{X}_i \tag{2.67}$$

where the *partial molar property* $\overline{X}_i$ is by definition

$$\overline{X}_i = \left.\frac{\partial X}{\partial n_i}\right)_{T,p,n_\ell} \tag{2.68}$$

and the subscript n_ℓ denotes that all n's except n_i are held fixed during differentiation. As $\overline{X}_i$ depends in general on temperature, pressure, and mixture composition: $\overline{X}_i(T, p, n_1, n_2, \ldots n_j)$, the partial molal property $\overline{X}_i$ is an intensive property of the mixture and not simply a property of the i*th* component.

Selecting the extensive property X to be volume, internal energy, enthalpy, entropy, and the Gibbs function, respectively, gives

$$\begin{aligned} V &= \sum_{i=1}^{j} n_i \overline{V}_i, \qquad U = \sum_{i=1}^{j} n_i \overline{U}_i \\ H &= \sum_{i=1}^{j} n_i \overline{H}_i, \qquad S = \sum_{i=1}^{j} n_i \overline{S}_i \\ G &= \sum_{i=1}^{j} n_i \overline{G}_i \end{aligned} \tag{2.69}$$

where $\overline{V}_i, \overline{U}_i, \overline{H}_i, \overline{S}_i,$ and $\overline{G}_i$ denote the respective partial molal properties.

When pure components, each initially at the same temperature and pressure, are mixed, the changes in volume, internal energy, enthalpy, and entropy on mixing are given by

$$\Delta V_{\text{mixing}} = \sum_{i=1}^{j} n_i \left(\overline{V}_i - \overline{v}_i\right) \tag{2.70a}$$

$$\Delta U_{\text{mixing}} = \sum_{i=1}^{j} n_i \left(\overline{U}_i - \overline{u}_i\right) \tag{2.70b}$$

$$\Delta H_{\text{mixing}} = \sum_{i=1}^{j} n_i \left(\overline{H}_i - \overline{h}_i\right) \tag{2.70c}$$

$$\Delta S_{\text{mixing}} = \sum_{i=1}^{j} n_i \left(\overline{S}_i - \overline{s}_i\right) \tag{2.70d}$$

where $\overline{v}_i, \overline{u}_i, \overline{h}_i,$ and $\overline{s}_i$ denote the molar-specific volume, internal energy, enthalpy, and entropy of pure component *i*.

Chemical Potential

The partial molal Gibbs function of the *i*th component of a multicomponent system is the *chemical potential*, μ_i,

$$\mu_i = \overline{G}_i = \left.\frac{\partial G}{\partial n_i}\right)_{T,p,n_\ell} \tag{2.71}$$

Like temperature and pressure, the chemical potential, μ_i is an *intensive* property.

When written in terms of chemical potentials, Equation 2.67 for the Gibbs function reads

$$G = \sum_{i=1}^{j} n_i \mu_i \tag{2.72}$$

For a *single component sysrem*, Equation 2.72 reduces to $G = n\mu$; that is, the chemical potential equals the molar Gibbs function. For an ideal gas mixture, comparison of Equations 2.63 and 2.72 suggests $\mu_i = \bar{g}_i\,(T, p_i)$; that is, the chemical potential of component *i* in an ideal gas mixture equals its Gibbs function per mole of gas *i* evaluated at the mixture temperature and the partial pressure of the *i*th gas of the mixture.

The chemical potential is a measure of the *escaping tendency* of a substance in a multiphase system: a substance tends to move from the phase having the higher chemical potential for that substance to the phase having a lower chemical potential. A necessary condition for *phase equilibrium is* that the chemical potential of each component has the same value in every phase.

The *Gibbs phase rule* gives the number *F* of independent intensive properties that may be arbitrarily specified to fix the intensive state of a system at equilibrium consisting of *N* nonreacting components present in *P* phases: $F = 2 + N - P$. *F* is called the *degrees of freedom* (or the *variance*). For water as a single component, for example, $N = 1$ and $F = 3 - P$.

- For a single phase, $P = 1$ and $F = 2$: two intensive properties can be varied independently, say temperature *and* pressure, while maintaining a single phase.
- For two phases, $P = 2$ and $F = 1$: only one intensive property can be varied independently if two phases are maintained — for example, temperature *or* pressure.
- For three phases, $P = 3$ and $F = 0$: there are no degrees of freedom; each intensive property of each phase is fixed. For a system consisting of ice, liquid water, and water vapor at equilibrium, there is a unique temperature: 0.01°C (32.02°F) and a unique pressure: 0.6113 kPa (0.006 atm).

The phase rule does not address the relative amounts that may be present in the various phases.

With $G = H - TS$ and $H = U + pV$, Equation 2.72 can be expressed as

$$U = TS - pV + \sum_{i=1}^{j} n_i \mu_i \tag{2.73}$$

from which can be derived

$$dU = TdS - pdV + \sum_{i=1}^{j} \mu_i dn_i \tag{2.74}$$

When the mixture composition is constant, Equation 2.74 reduces to Equation 2.31a.

Ideal Solution

The *Lewis-Randall rule* states that the fugacity $\bar{f}_i$ of each component *i* in an *ideal solution* is the product of its mole fraction and the fugacity of the pure component, f_i, at the same temperature, pressure, and state of aggregation (gas, liquid, or solid) as the mixture:

$$\bar{f}_i = y_i f_i \qquad \text{(Lewis-Randall rule)} \tag{2.75}$$

The following characteristics are exhibited by an ideal solution: $\overline{V}_i = \bar{v}_i$, $\overline{U}_i = \bar{u}_i$, $\overline{H}_i = \bar{h}_i$. With these, Equations 2.70a, b, and c show that there is no change in volume, internal energy, or enthalpy on mixing pure components to form an ideal solution. The *adiabatic* mixing of different pure components would result in an increase in entropy, however, because such a process is irreversible.

The volume of an ideal solution is

$$V = \sum_{i=1}^{j} n_i \bar{v}_i = \sum_{i=1}^{j} V_i \qquad \text{(ideal solution)} \tag{2.76}$$

where V_i is the volume that pure component i would occupy when at the temperature and pressure of the mixture. Comparing Equations 2.48a and 2.76, the *additive volume rule* is seen to be exact for ideal solutions. The internal energy and enthalpy of an ideal solution are

$$U = \sum_{i=1}^{j} n_i \bar{u}_i, \quad H = \sum_{i=1}^{j} n_i \bar{h}_i \qquad \text{(ideal solution)} \tag{2.77}$$

where $\bar{u}_i$ and $\bar{h}_i$ denote, respectively, the molar internal energy and enthalpy of pure component i at the temperature and pressure of the mixture. Many gaseous mixtures at low to moderate pressures are adequately modeled by the Lewis Randall rule. The ideal gas mixtures considered in Section 2.3, Ideal Gas Model, is an important special case. Some liquid solutions also can be modeled with the Lewis-Randall rule.

2.4 Combustion

The thermodynamic analysis of reactive systems is primarily an extension of principles presented in Sections 2.1 to 2.3. It is necessary, though, to modify the methods used to evaluate specific enthalpy and entropy.

Reaction Equations

In combustion reactions, rapid oxidation of combustible elements of the fuel results in energy release as combustion products are formed. The three major combustible chemical elements in most common fuels are carbon, hydrogen, and sulfur. Although sulfur is usually a relatively unimportant contributor to the energy released, it can be a significant cause of pollution and corrosion.

The emphasis in this section is on hydrocarbon fuels, which contain hydrogen, carbon, sulfur, and possibly other chemical substances. Hydrocarbon fuels may be liquids, gases, or solids such as coal. Liquid hydrocarbon fuels are commonly derived from crude oil through distillation and cracking processes. Examples are gasoline, diesel fuel, kerosene, and other types of fuel oils. The compositions of liquid fuels are commonly given in terms of mass fractions. For simplicity in combustion calculations, gasoline is often considered to be octane, C_8H_{18}, and diesel fuel is considered to be dodecane, $C_{12}H_{26}$. Gaseous hydrocarbon fuels are obtained from natural gas wells or are produced in certain chemical processes. Natural gas normally consists of several different hydrocarbons, with the major constituent being methane, CH_4. The compositions of gaseous fuels are commonly given in terms of mole fractions. Both gaseous and liquid hydrocarbon fuels can be synthesized from coal, oil shale, and tar sands. The composition of coal varies considerably with the location from which it is mined. For combustion calculations, the makeup of coal is usually expressed as an *ultimate analysis* giving the composition on a mass basis in terms of the relative amounts of chemical elements (carbon, sulfur, hydrogen, nitrogen, oxygen) and ash. Coal combustion is considered further in Chapter 8, Energy Conversion.

A fuel is said to have burned *completely* if all of the carbon present in the fuel is burned to carbon dioxide, all of the hydrogen is burned to water, and all of the sulfur is burned to sulfur dioxide. In practice, these conditions are usually not fulfilled and combustion is *incomplete.* The presence of carbon monoxide (CO) in the products indicates incomplete combustion. The products of combustion of *actual* combustion reactions and the relative amounts of the products can be determined with certainty only by experimental means. Among several devices for the experimental determination of the composition of products of combustion are the *Orsat analyzer*, *gas chromatograph*, *infrared analyzer*, and *flame ionization detector.* Data from these devices can be used to determine the makeup of the gaseous products of combustion. Analyses are frequently reported on a "dry" basis: mole fractions are determined for all gaseous products as if no water vapor were present. Some experimental procedures give an analysis including the water vapor, however.

Since water is formed when hydrocarbon fuels are burned, the mole fraction of water vapor in the gaseous products of combustion can be significant. If the gaseous products of combustion are cooled at constant mixture pressure, the *dew point temperature* (Section 2.3, Ideal Gas Model) is reached when water vapor begins to condense. Corrosion of duct work, mufflers, and other metal parts can occur when water vapor in the combustion products condenses.

Oxygen is required in every combustion reaction. Pure oxygen is used only in special applications such as cutting and welding. In most combustion applications, air provides the needed oxygen. Idealizations are often used in combustion calculations involving air: (1) all components of air other than oxygen (O_2) *are* lumped with nitrogen (N_2). On a molar basis air is then considered to be 21% oxygen and 79% nitrogen. With this idealization the molar ratio of the nitrogen to the oxygen in combustion air is 3.76; (2) the water vapor present in air may be considered in writing the combustion equation or ignored. In the latter case the combustion air is regarded as *dry*; (3) additional simplicity results by regarding the nitrogen present in the combustion air as inert. However, if high-enough temperatures are attained, nitrogen can form compounds, often termed NO_x, such as nitric oxide and nitrogen dioxide.

Even trace amounts of oxides of nitrogen appearing in the exhaust of internal combustion engines can be a source of air pollution.

The minimum amount of air that supplies sufficient oxygen for the complete combustion of all the combustible chemical elements is the *theoretical*, or *stoichiometic*, amount of air. In practice, the amount of air actually supplied may be greater than or less than the theoretical amount, depending on the application. The amount of air is commonly expressed as the *percent of theoretical air* or the *percent excess* (or *percent deficiency*) of air. The *air-fuel ratio* and its reciprocal *the fuel-air ratio*, each of which can be expressed on a mass or molar basis, are other ways that fuel-air mixtures are described. Another is the *equivalence ratio:* the ratio of the actual fuel-air ratio to the fuel-air ratio for complete combustion with the theoretical amount of air. The reactants form a *lean* mixture when the equivalence ratio is less than unity and a *rich* mixture when the ratio is greater than unity.

Example 9

Methane, CH_4, is burned with dry air. The molar analysis of the products on a dry basis is CO_2, 9.7%; CO, 0.5%; O_2, 2.95%; and N_2, 86.85%. Determine (a) the air-fuel ratio on both a molar and a mass basis, (b) the percent of theoretical air, (c) the equivalence ratio, and (d) the dew point temperature of the products, in °F, if the pressure is 1 atm.

Solution.

(a) The solution is conveniently conducted on the basis of 100 lbmol of dry products. The chemical equation then reads

$$a\mathrm{CH_4} + b\left(\mathrm{O_2} + 3.76\mathrm{N_2}\right) \rightarrow 9.7\mathrm{CO_2} + 0.5\mathrm{CO} + 2.95\mathrm{O_2} + 86.85\mathrm{N_2} + c\mathrm{H_2O}$$

where N_2 is regarded as inert. Water is included in the products together with the assumed 100 lbmol of dry products. Balancing the carbon, hydrogen, and oxygen, the reaction equation is

$$10.2\mathrm{CH_4} + 23.1\left(\mathrm{O_2} + 3.76\mathrm{N_2}\right) \rightarrow 9.7\mathrm{CO_2} + 0.5\mathrm{CO} + 2.95\mathrm{O_2} + 86.85\mathrm{N_2} + 20.4\mathrm{H_2O}$$

The nitrogen also balances, as can be verified. This checks the accuracy of both the given product analysis and the calculations conducted to determine the unknown coefficients. Exact closure cannot be expected with measured data, however. On a molar basis, the air-fuel ratio is

$$\overline{\mathrm{AF}} = \frac{23.1(4.76)}{10.2} = 10.78\frac{\mathrm{lbmol(air)}}{\mathrm{lbmol(fuel)}}$$

On a mass basis

$$\mathrm{AF} = (10.78)\left(\frac{28.97}{16.04}\right) = 19.47\frac{\mathrm{lb(air)}}{\mathrm{lb(fuel)}}$$

(b) The balanced chemical equation for the complete combustion of methane with the *theoretical* amount of air is

$$\mathrm{CH_4} + 2\left(\mathrm{O_2} + 3.76\mathrm{N_2}\right) \rightarrow \mathrm{CO_2} + 2\mathrm{H_2O} + 7.52\mathrm{N_2}$$

The theoretical air-fuel ratio on a molar basis is

$$\left(\overline{\mathrm{AF}}\right)_{theo} = \frac{2(4.76)}{1} = 9.52\frac{\mathrm{lbmol(air)}}{\mathrm{lbmol(fuel)}}$$

The percent theoretical air is then

$$\%\ \text{theoretical air} = \frac{\left(\overline{\text{AF}}\right)}{\left(\overline{\text{AF}}\right)_{theo}}$$

$$= \frac{10.78}{9.52} = 1.13(113\%)$$

(c) Equivalence ratio = $(\overline{FA})/(\overline{FA})_{theo}$ = 9.52/10.78 = 0.88. The reactants form a lean mixture.

(d) To determine the dew point temperature requires the partial pressure p_v of the water vapor. The mole fraction of the water vapor is

$$y_v = \frac{20.4}{100 + 20.4} = 0.169$$

Since p = 1 atm, p_v = 0.169 atm = 2.48 lbf/in.^{2}. With p_{sat} = 2.48 lbf/in.^{2}, the corresponding saturation temperature from the steam tables is 134°F. This is the dew point temperature.

Property Data for Reactive Systems

Tables of thermodynamic properties such as the steam tables provide values for the specific enthalpy and entropy relative to some arbitrary datum state where the enthalpy (or alternatively the internal energy) and entropy are set to zero. When a chemical reaction occurs, however, reactants disappear and products are formed, and it is generally no longer possible to evaluate $\Delta\bar{h}$ and $\Delta\bar{s}$ so that these arbitrary datums cancel. Accordingly, special means are required to assign specific enthalpy and entropy for application to reacting systems.

Property data suited for the analysis of reactive systems are available from several sources. The encyclopedic *JANAF Thermochemical Tables* is commonly used. Data for a wide range of substances are retrievable from Knacke et al. (1991), which provides both tabular data and analytical expressions readily programmable for use with personal computers of the specific heat, enthalpy, entropy, and Gibbs function. Textbooks on engineering thermodynamics also provide selected data, as, for example, Moran and Shapiro (1995).

Enthalpy of Formation

An enthalpy datum for reacting systems can be established by assigning arbitrarily a value of zero to the enthalpy of the *stable elements* at a *standard reference state* where the temperature is T_{ref} = 298.15 K (25°C) and the pressure is p_{ref}, which may be 1 bar or 1 atm depending on the data source. The term stable simply means that the particular element is chemically stable. For example, at the standard state the stable forms of hydrogen, oxygen, and nitrogen are H_2, O_2, and N_2 and not the monatomic H, O, and N.

The molar enthalpy of a *compound* at the standard state equals its *enthalpy of formation*, symbolized here by $\bar{h}_f^\circ$. The enthalpy of formation is the energy released or absorbed when the compound is formed from its elements, the compound and elements all being at T_{ref} and p_{ref}. The enthalpy of formation may be determined by application of procedures from statistical thermodynamics using observed spectroscopic data. The enthalpy of formation also can be found in principle by measuring the heat transfer in a reaction in which the compound is formed from the elements. In this chapter, the superscript ° is used to denote p_{ref}. For the case of the enthalpy of formation, the reference temperature T_{ref} is also intended by this symbol. Table 2.9 gives the values of the enthalpy of formation of various substances at 298 K and 1 atm.

The molar enthalpy of a substance at a state other than the standard state is found by adding the molar enthalpy change $\Delta\bar{h}$ between the standard state and the state of interest to the molar enthalpy of formation:

TABLE 2.9 Enthalpy of Formation, Gibbs Function of Formation, and Absolute Entropy of Various Substances at 298 K and 1 atm

$\bar{h}_f^\circ$ **and** $\bar{g}_f^\circ$ **(kJ/kmol),** $\bar{s}^\circ$ **(kJ/kmol•K)**

Substance	Formula	$\bar{h}_f^\circ$	$\bar{g}_f^\circ$	$\bar{s}^\circ$
Carbon	C(s)	0	0	5.74
Hydrogen	H_2(g)	0	0	130.57
Nitrogen	N_2(g)	0	0	191.50
Oxygen	O_2(g)	0	0	205.03
Carbon monoxide	CO(g)	–110,530	–137,150	197.54
Carbon dioxide	CO_2(g)	–393,520	–394,380	213.69
Water	H_2O(g)	–241,820	–228,590	188.72
	H_2O(l)	–285,830	–237,180	69.95
Hydrogen peroxide	H_2O_2(g)	–136,310	–105,600	232.63
Ammonia	NH_3(g)	–46,190	–16,590	192.33
Oxygen	O(g)	249,170	231,770	160.95
Hydrogen	H(g)	218,000	203,290	114.61
Nitrogen	N(g)	472,680	455,510	153.19
Hydroxyl	OH(g)	39,460	34,280	183.75
Methane	CH_4(g)	–74,850	–50,790	186.16
Acetylene	C_2H_2(g)	226,730	209,170	200.85
Ethylene	C_2H_4(g)	52,280	68,120	219.83
Ethane	C_2H_6(g)	–84,680	–32,890	229.49
Propylene	C_3H_6(g)	20,410	62,720	266.94
Propane	C_3H_8(g)	–103,850	–23,490	269.91
Butane	C_4H_{10}(g)	–126,150	–15,710	310.03
Pentane	C_5H_{12}(g)	–146,440	–8,200	348.40
Octane	C_8H_{18}(g)	–208,450	17,320	463.67
	C_8H_{18}(l)	–249,910	6,610	360.79
Benzene	C_6H_6(8)	82,930	129,660	269.20
Methyl alcohol	CH_3OH(g)	–200,890	–162,140	239.70
	CH_3OH(I)	–238,810	–166,290	126.80
Ethyl alcohol	C_2H_5OH(g)	–235,310	–168,570	282.59
	C_2H_5OH(I)	–277,690	174,890	160.70

Source: Adapted from Wark, K. 1983. *Thermodynamics*, 4th ed. McGraw-Hill, New York, as based on JANAF Thermochemical Tables, NSRDS-NBS-37, 1971; *Selected Values of Chemical Thermodynamic Properties*, NBS Tech. Note 270-3, 1968; and *API Research Project 44*, Carnegie Press, 1953.

$$\bar{h}(T,p) = \bar{h}_f^\circ + \left[\bar{h}(T,p) - \bar{h}\left(T_{ref}, p_{ref}\right)\right] = \bar{h}_f^\circ + \Delta\bar{h} \tag{2.78}$$

That is, the enthalpy of a substance is composed of $\bar{h}_f^\circ$, associated with the formation of the substance from its elements, and $\Delta\bar{h}$, associated with a change of state at constant composition. An arbitrarily chosen datum can be used to determine $\Delta\bar{h}$, since it is a *difference* at constant composition. Accordingly, $\Delta\bar{h}$ can be evaluated from sources such as the steam tables and the ideal gas tables.

The *enthalpy of combustion*, $\bar{h}_{RP}$, is the difference between the enthalpy of the products and the enthalpy of the reactants, each on a per-mole-of-fuel basis, when complete combustion occurs and both reactants and products are at the same temperature and pressure. For hydrocarbon fuels the enthalpy of combustion is negative in value since chemical internal energy is liberated in the reaction. The *heating value* of a fuel is a positive number equal to the magnitude of the enthalpy of combustion. Two heating values are recognized: the *higher* heating value and the *lower* heating value. The higher heating value

is obtained when all the water formed by combustion is a liquid; the lower heating value is obtained when all the water formed by combustion is a vapor. The higher heating value exceeds the lower heating value by the energy that would be required to vaporize the liquid water formed at the specified temperature. Heating values are typically reported at a temperature of 25°C (77°F) and a pressure of 1 bar (or 1 atm). These values also depend on whether the fuel is a liquid or a gas. A sampling is provided on a unit-mass-of-fuel basis in Table 2.10.

TABLE 2.10 Heating Values in kJ/kg of Selected Hydrocarbons at 25°C

		Higher Value[a]		Lower Value[b]	
Hydrocarbon	**Formula**	**Liquid Fuel**	**Gas. Fuel**	**Liquid Fuel**	**Gas. Fuel**
Methane	CH_4	—	55,496	—	50,010
Ethane	C_2H_6	—	51,875	—	47,484
Propane	C_3H_8	49,973	50,343	45,982	46,352
n-Butane	C_4H_{10}	49,130	49,500	45,344	45,714
n-Octane	C_8H_{18}	47,893	48,256	44,425	44,788
n-Dodecane	$C_{12}H_{26}$	47,470	47,828	44,109	44,467
Methanol	CH_3OH	22,657	23,840	19,910	21,093
Ethanol	C_3H_5OH	29,676	30,596	26,811	27,731

[a] H_2O liquid in the products.
[b] H_2O vapor in the products.

In the absence of work $\dot{W}_{cv}$ and appreciable kinetic and potential energy effects, the energy liberated on combustion is transferred from a reactor at steady state in two ways: the energy accompanying the exiting combustion products and by heat transfer. The temperature that would be achieved by the products in the limit of adiabatic operation is the *adiabatic flame* or *adiabatic combustion* temperature.

For a specified fuel and specified temperature and pressure of the reactants, the *maximum* adiabatic flame temperature is realized for complete combustion with the theoretical amount of air. Example 10 provides an illustration. The measured value of the temperature of the combustion products may be several hundred degrees below the calculated maxunum adiabatic flame temperature, however, for several reasons including the following: (1) heat loss can be reduced but not eliminated; (2) once adequate oxygen has been provided to permit complete combustion, bringing in more air dilutes the combustion products, lowering the temperature; (3) incomplete combustion tends to reduce the temperature of the products, and combustion is seldom complete; (4) as result of the high temperatures achieved, some of the combustion products may dissociate. Endothermic dissociation reactions also lower the product temperature.

Absolute Entropy

A common datum for assigning entropy values to substances involved in chemical reactions is realized through the *third law* of thermodynamics, which is based on experimental observations obtained primarily from studies of chemical reactions at low temperatures and specific heat measurements at temperatures approaching absolute zero. The third law states that the entropy of a pure crystalline substance is zero at the absolute zero of temperature, 0 K or 0°R. Substances not having a pure crystalline structure have a nonzero value of entropy at absolute zero.

The third law provides a datum relative to which the entropy of each substance participating in a reaction can be evaluated. The entropy relative to this datum is called the *absolute* entropy. The change in entropy of a substance between absolute zero and any given state can be determined from measurements of energy transfers and specific heat data or from procedures based on statistical thermodynamics and observed molecular data. Table 2.9 and Tables A.2 and A.8 provide absolute entropy data for various substances. In these tables, p_{ref} =1 atm.

When the absolute entropy is known at pressure p_{ref} and temperature T, the absolute entropy at the same temperature and *any* pressure p can be found from

$$\bar{s}(T,p) = \bar{s}\left(T,p_{ref}\right) + \left[\bar{s}(T,p) - \bar{s}\left(T,p_{ref}\right)\right] \tag{2.79}$$

For an ideal gas, the second term on the right side of Equation 2.79 can be evaluated by using Equation 2.58, giving

$$\bar{s}(T,p) = \bar{s}^{\circ}(T) - \bar{R}\ln\frac{p}{p_{ref}} \qquad \text{(ideal gas)} \tag{2.80}$$

In this expression, $\bar{s}^{\circ}(T)$ denotes the absolute entropy at temperature T and pressure p_{ref}.

The entropy of the ith component of an *ideal gas mixture* is evaluated at the mixture temperature T and the *partial* pressure p_i: $\bar{s}_i(T, p_i)$. For the ith component, Equation 2.80 takes the form

$$\begin{aligned}\bar{s}_i\left(T,p_i\right) &= \bar{s}_i^{\circ}(T) - \bar{R}\ln\frac{p_i}{p_{ref}} \\ &= \bar{s}_i^{\circ}(T) - \bar{R}\ln\frac{y_i p}{p_{ref}} \qquad \text{(ideal gas)}\end{aligned} \tag{2.81}$$

where $\bar{s}_i^{\circ}(T)$ is the absolute entropy of component i at temperature T and p_{ref}.

Example 10

Liquid octane at 25°C, 1 atm enters a well insulated reactor and reacts with dry air entering at the same temperature and pressure. For steady-state operation and negligible effects of kinetic and potential energy, determine the temperature of the combustion products for complete combustion with the theoretical amount of air, and (b) the rates of entropy generation and exergy destruction, each per kmol of fuel.

Solution. For combustion of liquid octane with the theoretical amount of air, the chemical equation is

$$C_8H_{18}(l) + 12.5O_2 + 47N_2 \rightarrow 8CO_2 + 9H_2O(g) + 47N_2$$

(a) At steady state, the control volume energy rate balance reduces to read

$$0 = \cancel{\frac{\dot{Q}_{cv}}{\dot{n}_F}}^{0} - \cancel{\frac{\dot{W}_{cv}}{\dot{n}_F}}^{0} + \sum_R n_i\left(\bar{h}_f^{\circ} + \Delta\bar{h}\right)_i - \sum_P n_e\left(\bar{h}_f^{\circ} + \Delta\bar{h}\right)_e$$

where R denotes reactants, P denotes products, and the symbols for enthalpy have the same significance as in Equation 2.78. Since the reactants enter at 25°C, the corresponding $(\Delta\bar{h})_i$ terms vanish, and the energy rate equation becomes

$$\sum_P n_e\left(\Delta\bar{h}\right)_e = \sum_R n_i\bar{h}_{fi}^{\circ} - \sum_P n_e\bar{h}_{fe}^{\circ}$$

Introducing coefficients from the reaction equation, this takes the form

$$\begin{aligned}&8\left(\Delta\bar{h}\right)_{CO_2} + 9\left(\Delta\bar{h}\right)_{H_2O(g)} + 47\left(\Delta\bar{h}\right)_{N_2} = \left[\left(\bar{h}_f^{\circ}\right)_{C_8H_{18}(l)} + 12.5\left(\bar{h}_f^{\circ}\right)_{O_2} + 47\left(\bar{h}_f^{\circ}\right)_{N_2}\right] \\ &-\left[8\left(\bar{h}_f^{\circ}\right)_{CO_2} + 9\left(\bar{h}_f^{\circ}\right)_{H_2O(g)} + 47\left(\Delta\bar{h}_f^{\circ}\right)_{N_2}\right]\end{aligned}$$

Using data from Table 2.9 to evaluate the right side,

$$8\left(\Delta\bar{h}\right)_{CO_2} + 9\left(\Delta\bar{h}\right)_{H_2O(g)} + 47\left(\Delta\bar{h}\right)_{N_2} = 5,074,630 \text{ kJ/kmol (fuel)}$$

Each $\Delta\bar{h}$ term on the left side of this equation depends on the temperature of the products, T_p, which can be solved for iteratively as T_p = 2395 K.

(b) The entropy rate balance on a per-mole-of-fuel basis takes the form

$$0 = \sum_j \frac{\dot{Q}_j}{\dot{n}_F}\cancel{T_j}^{\,0} + \bar{s}_F + \left(12.5\bar{s}_{O_2} + 47\bar{s}_{N_2}\right) - \left(8\bar{s}_{CO_2} + 9\bar{s}_{H_2O(g)} + 47\bar{s}_{N_2}\right) + \frac{\dot{S}_{gen}}{\dot{n}_F}$$

or on rearrangement,

$$\frac{\dot{S}_{gen}}{\dot{n}_F} = \left(8\bar{s}_{CO_2} + 9\bar{s}_{H_2O(g)} + 47\bar{s}_{N_2}\right) - \bar{s}_F - \left(12.5\bar{s}_{O_2} + 47\bar{s}_{N_2}\right)$$

The absolute entropy of liquid octane from Table 2.9 is 360.79 kJ/mol · K. The oxygen and nitrogen in the combustion air enter the reactor as components of an ideal gas mixture at T_{ref}, p_{ref}. With Equation 2.81, where $p = p_{ref}$, and absolute entropy data from Table 2.9,

$$\begin{aligned}\bar{s}_{O_2} &= \bar{s}^\circ_{O_2}\left(T_{ref}\right) - \bar{R}\ln y_{O_2}\\ &= 205.03 - 8.314\ln 0.21 = 218.01 \text{ kJ/kmol}\cdot\text{K}\\ \bar{s}_{N_2} &= \bar{s}^\circ_{N_2}\left(T_{ref}\right) - \bar{R}\ln y_{N_2}\\ &= 191.5 - 8.314\ln 0.79 = 193.46 \text{ kJ/kmol}\cdot\text{K}\end{aligned}$$

The product gas exits as a gas mixture at 1 atm, 2395 K with the following composition: y_{CO_2} = 8/64 = 0.125, $y_{H_2O(g)}$ = 9/64 = 0.1406, y_{N_2} = 47/64 = 0.7344. With Equation 2.81, where $p = p_{ref}$, *a*nd absolute entropy data at 2395 K from Table A.2,

$$\begin{aligned}\bar{s}_{CO_2} &= 320.173 - 8.314\ln 0.125 = 337.46 \text{ kJ/kmol}\cdot\text{K}\\ \bar{s}_{H_2O} &= 273.986 - 8.314\ln 0.1406 = 290.30 \text{ kJ/kmol}\cdot\text{K}\\ \bar{s}_{N_2} &= 258.503 - 8.314\ln 0.7344 = 261.07 \text{ kJ/kmol}\cdot\text{K}\end{aligned}$$

Inserting values, the rate of entropy generation is

$$\begin{aligned}\frac{\dot{S}_{gen}}{\dot{n}_F} &= 8(337.46) + 9(290.30) + 47(261.07) - 360.79 - 12.5(218.01) - 47(193.46)\\ &= 5404 \text{ kJ/kmol}\cdot\text{K}\end{aligned}$$

Using Equation 2.87 and assuming T_0 = 298 K, the rate of exergy destruction is $\dot{E}_D/\dot{n}_F$ = 1.61 × 10^6 kJ/kmol.

Gibbs Function of Formation

Paralleling the approach used for enthalpy, a value of zero is assigned to the Gibbs function of each stable element at the standard state. The *Gibbs function of formation* of a compound equals the change in the Gibbs function for the reaction in which the compound is formed from its elements. Table 2.9 provides Gibbs function of formation data of various substances at 298 K and 1 atm.

The Gibbs function at a state other than the standard state is found by adding to the Gibbs function of formation the change in the specific Gibbs function $\Delta\bar{g}$ between the standard state and the state of interest:

$$\bar{g}(T,p) = \bar{g}_f^\circ + \left[\bar{g}(T,p) - \bar{g}\left(T_{ref}, p_{ref}\right)\right] = \bar{g}_f^\circ + \Delta\bar{g} \tag{2.82a}$$

where

$$\Delta\bar{g} = \left[\bar{h}(T,p) - \bar{h}\left(T_{ref}, p_{ref}\right)\right] - \left[T\bar{s}(T,p) - T_{ref}\bar{s}\left(T_{ref}, p_{ref}\right)\right] \tag{2.82b}$$

The Gibbs function of component *i* in an ideal gas mixture is evaluated at the partial pressure of component *i* and the mixture temperature.

As an application, the maximum theoretical work that can be developed, per mole of fuel consumed, is evaluated for the control volume of Figure 2.15, where the fuel and oxygen each enter in separate streams and carbon dioxide and water each exit separately. All entering and exiting streams are at the same temperature *T* and pressure *p*. The reaction is complete:

$$C_aH_b + \left(a + \frac{b}{4}\right)O_2 \rightarrow aCO_2 + \frac{b}{2}H_2O$$

This control volume is similar to idealized devices such as a reversible fuel *cell* or a *van't Hoff equilibrium box.*

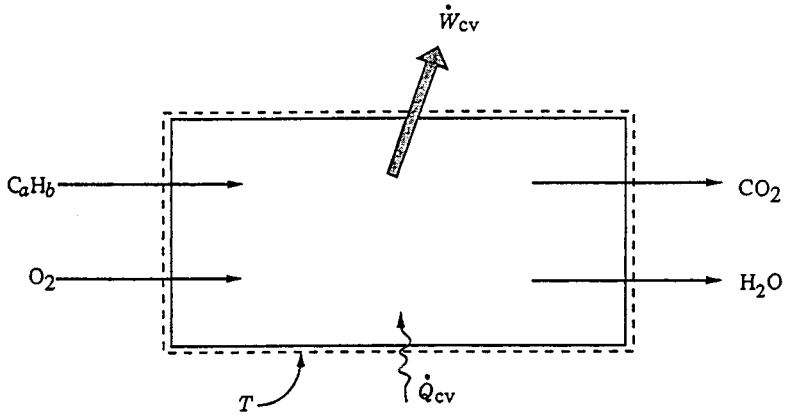


FIGURE 2.15 Device for evaluating maximum work.

For steady-state operation, the energy rate balance reduces to give

$$\frac{\dot{W}_{cv}}{\dot{n}_F} = \frac{\dot{Q}_{cv}}{\dot{n}_F} + \bar{h}_F + \left(a + \frac{b}{4}\right)\bar{h}_{O_2} - a\bar{h}_{CO_2} - \frac{b}{2}\bar{h}_{H_2O}$$

where $\dot{n}_F$ denotes the molar flow rate of the fuel. Kinetic and potential energy effects are regarded as negligible. If heat transfer occurs only at the temperature *T*, an entropy balance for the control volume takes the form

$$0 = \frac{\dot{Q}_{cv}/\dot{n}_F}{T} + \bar{s}_F + \left(a + \frac{b}{4}\right)\bar{s}_{O_2} - a\bar{s}_{CO_2} - \frac{b}{2}\bar{s}_{H_2O} + \frac{\dot{S}_{gen}}{\dot{n}_F}$$

Eliminating the heat transfer term from these expressions, an expression for the maximum theoretical value of the work developed per mole of fuel is obtained when the entropy generation term is set to zero:

$$\left(\frac{\dot{W}_{cv}}{\dot{n}_F}\right)_{\substack{int\\rev}} = \left[\bar{h}_F + \left(a + \frac{b}{4}\right)\bar{h}_{O_2} - a\bar{h}_{CO_2} - \frac{b}{2}\bar{h}_{H_2O}\right](T,p) - T\left[\bar{s}_F + \left(a + \frac{b}{4}\right)\bar{s}_{O_2} - a\bar{s}_{CO_2} - \frac{b}{2}\bar{s}_{H_2O}\right](T,p)$$

This can be written alternatively in terms of the enthalpy of combustion as

$$\left(\frac{\dot{W}_{cv}}{\dot{n}_F}\right)_{\substack{int\\rev}} = -\bar{h}_{RP}(T,p) - T\left[\bar{s}_F + \left(a + \frac{b}{4}\right)\bar{s}_{O_2} - a\bar{s}_{CO_2} - \frac{b}{2}\bar{s}_{H_2O}\right](T,p) \qquad (2.83a)$$

or in terms of Gibbs functions as

$$\left(\frac{\dot{W}_{cv}}{\dot{n}_F}\right)_{\substack{int\\rev}} = \left[\bar{g}_F + \left(a + \frac{b}{4}\right)\bar{g}_{O_2} - a\bar{g}_{CO_2} - \frac{b}{2}\bar{g}_{H_2O}\right](T,p) \qquad (2.83b)$$

Equation 2.83b is used in the solution to Example 11.

Example 11

Hydrogen (H_2) and oxygen (O_2), each at 25°C, 1 atm, enter a fuel cell operating at steady state, and liquid water exits at the same temperature and pressure. The hydrogen flow rate is 2×10^{-4} kmol/sec and the fuel cell operates isothermally at 25°C. Determine the maximum theoretical power the cell can develop, in kW.

Solution. The overall cell reaction is $H_2 + 1/2\ O_2 \rightarrow H_2O(\ell)$, and Equations 2.83 are applicable. Selecting Equation 2.83b, and using Gibbs function data from Table 2.9,

$$\left(\frac{\dot{W}_{cv}}{\dot{n}_F}\right)_{\substack{int\\rev}} = \left(\bar{g}_{H_2} + \frac{1}{2}\bar{g}_{O_2} - \bar{g}_{H_2O(\ell)}\right)(25°C,\ 1\ \text{atm})$$

$$= 0 + \frac{1}{2}(0) - (-237{,}180) = 237{,}180\ \text{kJ/kmol}$$

Then

$$\left(\dot{W}_{cv}\right)_{\substack{int\\rev}} = \left(237{,}180\,\frac{\text{kJ}}{\text{kmol}}\right)\left(2\times10^{-4}\,\frac{\text{kmol}}{\text{s}}\right)\left(\frac{\text{kW}}{1\text{kJ/s}}\right) = 47.4\ \text{kW}$$

Reaction Equilibrium

Let the objective be to determine the equilibrium composition of a system consisting of five gases A, B, C, D, and E, at a temperature T and pressure p, subject to a chemical reaction of the form

$$\nu_A A + \nu_B B \leftrightarrow \nu_C C + \nu_D D$$

where the ν's are stoichiometric coefficients. Component E is assumed to be inert and thus does not appear in the reaction equation. The equation suggests that at equilibrium the tendency of A and B to form C and D is just balanced by the tendency of C and D to form A and B.

At equilibrium, the temperature and pressure would be uniform throughout the system. Additionally, the *equation of reaction equilibrium* must be satisfied:

$$\nu_A \mu_A + \nu_B \mu_B = \nu_C \mu_C + \nu_D \mu_D \tag{2.84a}$$

where the μ's are the chemical potentials (Section 2.3, Multicomponent Systems) of A, B, C, and D in the equilibrium mixture. In principle, the composition that would be present at equilibrium for a given temperature and pressure can be determined by solving this equation.

For ideal gas mixtures, the solution procedure is simplified by using the *equilibrium constant* $K(T)$ and the following equation:

$$\begin{aligned} K(T) &= \frac{y_C^{\nu_C} y_D^{\nu_D}}{y_A^{\nu_A} y_B^{\nu_B}} \left(\frac{p}{p_{ref}} \right)^{\nu_C+\nu_D-\nu_A-\nu_B} \\ &= \frac{n_C^{\nu_C} n_D^{\nu_D}}{n_A^{\nu_A} n_B^{\nu_B}} \left(\frac{p/p_{ref}}{n} \right)^{\nu_C+\nu_D-\nu_A-\nu_B} \end{aligned} \tag{2.84b}$$

where y_A, y_B, y_C, and y_D denote the mole fractions of A, B, C, and D in the equilibrium mixture and $n = n_A + n_B + n_C + n_D + n_E$, where the n's denote the molar amounts of the gases in the mixture. Tabulations of $K(T)$ for each of several reactions of the form Equation 2.84a are provided in Table 2.11. An application of Equation 2.84b is provided in Example 12.

Example 12

One kmol of CO reacts with the theoretical amount of dry air to form an equilibrium mixture of CO_2, CO, O_2, and N_2 at 2500 K, 1 atm. Determine the amount of CO in the equilibrium mixture, in kmol.

Solution. The reaction of CO with the theoretical amount of dry air to form CO_2, CO, O_2, and N_2 is

$$\mathrm{CO} + \frac{1}{2}\mathrm{O_2} + 1.88\mathrm{N_2} \rightarrow z\mathrm{CO} + \frac{z}{2}\mathrm{O_2} + (1-z)\mathrm{CO_2} + 1.88\mathrm{N_2}$$

where z is the amount of CO, in kmol, present in the equilibrium mixture. The total number of moles n is

$$n = z + \frac{z}{2} + (1-z) + 1.88 = \frac{5.76+z}{2}$$

At equilibrium $CO_2 \leftrightarrow$ CO + 1/2 O_2; and Equation 2.84b takes the form

$$K = \frac{z(z/2)^{1/2}}{1-z} \left[\frac{p/p_{ref}}{(5.76+z)/2} \right]^{1/2}$$

where $p/p_{ref} = 1$. At 2500 K, Table 2.11 gives $K = 0.0363$. Solving iteratively, z = 0.175.

TABLE 2.11 Logarithms to the Base 10 of the Equilibrium Constant *K*

	$\log_{10} K$								
Temp (K)	**$H_2 \Leftrightarrow 2H$**	**$O_2 \Leftrightarrow 2O$**	**$N_2 \Leftrightarrow 2N$**	**$\frac{1}{2}O_2 + \frac{1}{2}N_2 \Leftrightarrow NO$**	**$H_2O \Leftrightarrow H_2 + \frac{1}{2}O_2$**	**$H_2O \Leftrightarrow OH + \frac{1}{2}H_2$**	**$CO_2 \Leftrightarrow CO + \frac{1}{2}O_2$**	**$CO_2 + H_2 \Leftrightarrow CO + H_2O$**	**Temp (°R)**
298	–71.224	–81.208	–159.600	–15.171	–40.048	–46.054	–45.066	–5.018	537
500	–40.316	–45.880	–92.672	–8.783	–22.886	–26.130	–25.025	–2.139	900
1000	–17.292	–19.614	–43.056	–4.062	–10.062	–11.280	–10.221	–0.159	1800
1200	–13.414	–15.208	–34.754	–3.275	–7.899	–8.811	–7.764	+0.135	2160
1400	–10.630	–12.054	–28.812	–2.712	–6.347	–7.021	–6.014	+0.333	2520
1600	–8.532	–9.684	–24.350	–2.290	–5.180	–5.677	–4.706	+0.474	2880
1700	–7.666	–8.706	–22.512	–2.116	–4.699	–5.124	–4.169	+0.530	3060
1800	–6.896	–7.836	–20.874	–1.962	–4.270	–4.613	–3.693	+0.577	3240
1900	–6.204	–7.058	–19.410	–1.823	–3.886	–4.190	–3.267	+0.619	3420
2000	–5.580	–6.356	–18.092	–1.699	–3.540	–3.776	–2.884	+0.656	3600
2100	–5.016	–5.720	–16.898	–1.586	–3.227	–3.434	–2.539	+0.688	3780
2200	–4.502	–5.142	–15.810	–1.484	–2.942	–3.091	–2.226	+0.716	3960
2300	–4.032	–4.614	–14.818	–1.391	–2.682	–2.809	–1.940	+0.742	4140
2400	–3.600	–4.130	–13.908	–1.305	–2.443	–2.520	–1.679	+0.764	4320
2500	–3.202	–3.684	–13.070	–1.227	–2.224	–2.270	–1.440	+0.784	4500
2600	–2.836	–3.272	–12.298	–1.154	–2.021	–2.038	–1.219	+0.802	4680
2700	–2.494	–2.892	–11.580	–1.087	–1.833	–1.823	–1.015	+0.818	4860
2800	–2.178	–2.536	–10.914	–1.025	–1.658	–1.624	–0.825	+0.833	5040
2900	–1.882	–2.206	–10.294	–0.967	–1.495	–1.438	–0.649	+0.846	5220
3000	–1.606	–1.898	–9.716	–0.913	–1.343	–1.265	–0.485	+0.858	5400
3100	–1.348	–1.610	–9.174	–0.863	–1.201	–1.103	–0.332	+0.869	5580
3200	–1.106	–1.340	–8.664	–0.815	–1.067	–0.951	–0.189	+0.878	5760
3300	–0.878	–1.086	–8.186	–0.771	–0.942	–0.809	–0.054	+0.888	5940
3400	–0.664	–0.846	–7.736	–0.729	–0.824	–0.674	+0.071	+0.895	6120
3500	–0.462	–0.620	–7.312	–0.690	–0.712	–0.547	+0.190	+0.902	6300

Source: Based on data from the JANAF Thermochemical Tables, NSRDS-NBS-37, 1971.

2.5 Exergy Analysis

The method of *exergy analysis* (*availability analysis*) presented in this section enables the location, cause, and true magnitude of energy resource waste and loss to be determined. Such information can be used in the design of new energy-efficient systems and for improving the performance of existing systems. Exergy analysis also provides insights that elude a purely first-law approach. For example, on the basis of first-law reasoning alone, the condenser of a power plant may be mistakenly identified as the component primarily responsible for the plant's seemingly low overall performance. An exergy analysis correctly reveals not only that the condenser loss is relatively unimportant (see the last two rows of the Rankine cycle values of Table 2.15), but also that the steam generator is the principal site of thermodynamic inefficiency owing to combustion and heat transfer irreversibilities within it.

When exergy concepts are combined with principles of engineering economy, the result is known as *thermoeconomics*. Thermoeconomics allows the real cost sources at the component level to be identified: capital investment costs, operating and maintenance costs, and the costs associated with the destruction and loss of exergy. Optimization of thermal systems can be achieved by a careful consideration of such cost sources. From this perspective thermoeconomics *is exergy-aided cost minimization.*

Discussions of exergy analysis and thermoeconomics are provided by Bejan et al. (1996), Moran (1989), and Moran and Shapiro (1995). In this section salient aspects are presented.

Defining Exergy

An opportunity for doing work exists whenever two systems at different states are placed in communication because, in principle, work can be developed as the two are allowed to come into equilibrium. When one of the two systems is a suitably idealized system called an *environment* and the other is some system of interest, *exergy* is the maximum theoretical useful work (shaft work or electrical work) obtainable as the systems interact to equilibrium, heat transfer occurring with the environment only. (Alternatively, exergy is the minimum theoretical useful work required to form a quantity of matter from substances present in the environment and to bring the matter to a specified state.) Exergy is a measure of the *departure* of the state of the system from that of the environment, and is therefore an attribute of the system and environment together. Once the environment is specified, however, a value can be assigned to exergy in terms of property values for the system only, so exergy can be regarded as an extensive property of the system.

Exergy can be destroyed and generally is not conserved. A limiting case is when exergy would be completely destroyed, as would occur if a system were to come into equilibrium with the environment *spontaneously* with no provision to obtain work. The capability to develop work that existed initially would be completely wasted in the spontaneous process. Moreover, since no work needs to be done to effect such a spontaneous change, the value of exergy can never be negative.

Environment

Models with various levels of specificity are employed for describing the environment used to evaluate exergy. Models of the environment typically refer to some portion of a system's surroundings, the intensive properties of each phase of which are uniform and do not change significantly as a result of any process under consideration. The environment is regarded as composed of common substances existing in abundance within the Earth's atmosphere, oceans, and crust. The substances are in their stable forms as they exist naturally, and there is no possibility of developing work from interactions — physical or chemical — between parts of the environment. Although the intensive properties of the environment are assumed to be unchanging, the extensive properties can change as a result of interactions with other systems. Kinetic and potential energies are evaluated relative to coordinates in the environment, all parts of which are considered to be at rest with respect to one another.

For computational ease, the temperature T_0 and pressure p_0 of the environment are often taken as standard-state values, such as 1 atm and 25°C (77°F). However, these properties may be specified

differently depending on the application. T_0 and p_0 might be taken as the average ambient temperature and pressure, respectively, for the location at which the system under consideration operates. Or, if the system uses atmospheric air, T_0 might be specified as the average air temperature. If both air and water from the natural surroundings are used, T_0 would be specified as the lower of the average temperatures for air and water.

Dead States

When a system is in equilibrium with the environment, the state of the system is called the *dead state*. At the dead state, the conditions of mechanical, thermal, and chemical equilibrium between the system and the environment are satisfied: the pressure, temperature, and chemical potentials of the system equal those of the environment, respectively. In addition, the system has no motion or elevation relative to coordinates in the environment. Under these conditions, there is no possibility of a spontaneous change within the system or the environment, nor can there be an interaction between them. The value of exergy is zero.

Another type of equilibrium between the system and environment can be identified. This is a restricted form of equilibrium where only the conditions of mechanical and thermal equilibrium must be satisfied. This state of the system is called the *restricted dead state*. At the restricted dead state, the fixed quantity of matter under consideration is imagined to be sealed in an envelope impervious to mass flow, at zero velocity and elevation relative to coordinates in the environment, and at the temperature T_0 and pressure p_0.

Exergy Balances

Exergy can be transferred by three means: exergy transfer associated with work, exergy transfer associated with heat transfer, and exergy transfer associated with the matter entering and exiting a control volume. All such exergy transfers are evaluated relative to the environment used to define exergy. Exergy is also destroyed by irreversibilities within the system or control volume.

Exergy balances can be written in various forms, depending on whether a closed system or control volume is under consideration and whether steady-state or transient operation is of interest. Owing to its importance for a wide range of applications, an exergy rate balance for control volumes at steady state is presented next.

Control Volume Exergy Rate Balance

At steady state, the control volume exergy rate balance takes the form

$$0 = \underbrace{\sum_j \dot{E}_{q,j} - \dot{W}_{cv} + \sum_i \dot{E}_i - \sum_e \dot{E}_e}_{\text{rates of exergy transfer}} - \underbrace{\dot{E}_D}_{\text{rate of exergy destruction}} \tag{2.85a}$$

or

$$0 = \sum_j \left(1 - \frac{T_0}{T_j}\right)\dot{Q}_j - \dot{W}_{cv} + \sum_i \dot{m}_i e_i - \sum_e \dot{m}_e e_e - \dot{E}_D \tag{2.85b}$$

$\dot{W}_{cv}$ has the same significance as in Equation 2.22: the work rate excluding the flow work. $\dot{Q}_j$ is the time rate of heat transfer at the location on the boundary of the control volume where the instantaneous temperature is T_j. The associated rate of exergy transfer is

$$\dot{E}_{q,j} = \left(1 - \frac{T_0}{T_j}\right)\dot{Q}_j \tag{2.86}$$

As for other control volume rate balances, the subscripts i and e denote inlets and outlets, respectively. The exergy transfer rates at control volume inlets and outlets are denoted, respectively, as $\dot{E}_i = \dot{m}_i e_i$ and $\dot{E}_e = \dot{m}_e e_e$. Finally, $\dot{E}_D$ accounts for the time rate of exergy destruction due to irreversibilities within the control volume. The exergy destruction rate is related to the entropy generation rate by

$$\dot{E}_D = T_0 \dot{S}_{gen} \tag{2.87}$$

The specific exergy transfer terms e_i and e_e are expressible in terms of four components: physical exergy e^{PH}, kinetic exergy e^{KN}, potential exergy e^{PT}, and chemical exergy e^{CH}:

$$e = e^{PH} + e^{KN} + e^{PT} + e^{CH} \tag{2.88}$$

The first three components are evaluated as follows:

$$e^{PH} = (h - h_0) - T_0(s - s_0) \tag{2.89a}$$

$$e^{KN} = \frac{1}{2}\mathrm{v}^2 \tag{2.89b}$$

$$e^{PT} = gz \tag{2.89c}$$

In Equation 2.89a, h_0 and s_0 denote, respectively, the specific enthalpy and specific entropy at the restricted dead state. In Equations 2.89b and 2.89c, v and z denote velocity and elevation relative to coordinates in the environment, respectively. The chemical exergy e^{CH} is considered next.

Chemical Exergy

To evaluate the chemical exergy, the exergy component associated with the departure of the chemical composition of a system from that of the environment, the substances comprising the system are referred to the properties of a suitably selected set of environmental substances. For this purpose, alternative models of the environment have been developed. For discussion, see, for example, Moran (1989) and Kotas (1995).

Exergy analysis is facilitated, however, by employing a *standard environment* and a corresponding table of *standard chemical exergies.* Standard chemical exergies are based on standard values of the environmental temperature T_0 and pressure p_0 — for example, 298.15 K (25°C) and 1 atm, respectively. A standard environment is also regarded as consisting of a set of reference substances with standard concentrations reflecting as closely as possible the chemical makeup of the natural environment. The reference substances generally fall into three groups: gaseous components of the atmosphere, solid substances from the lithosphere, and ionic and noninonic substances from the oceans. The chemical exergy data of Table 2.12 correspond to two alternative standard exergy reference environments, called here model I and model II, that have gained acceptance for engineering evaluations.

Although the use of standard chemical exergies greatly facilitates the application of exergy principles, the term *standard* is somewhat misleading since there is no one specification of the environment that

suffices for all applications. Still, chemical exergies calculated relative to alternative specifications of the environment are generally in good agreement. For a broad range of engineering applications the simplicity and ease of use of standard chemical exergies generally outweigh any slight lack of accuracy that might result. In particular, the effect of slight variations in the values of T_0 and p_0 about the values used to determine the standard chemical exergies reported in Table 2.12 can be neglected.

The literature of exergy analysis provides several expressions allowing the chemical exergy to be evaluated in particular cases of interest. The molar chemical exergy of a gas mixture, for example, can be evaluated from

$$\bar{e}^{CH} = \sum_{i=1}^{j} y_i \bar{e}_i^{CH} + \bar{R}T_0 \sum_{i=1}^{j} y_i \ln y_i \tag{2.90}$$

where $\bar{e}_i^{CH}$ is the molar chemical exergy of the ith component.

Example 13

Ignoring the kinetic and potential exergies, determine the exergy rate, in kJ/kg, associated with each of the following streams of matter:

(a) Saturated water vapor at 20 bar.
(b) Methane at 5 bar, 25°C.

Let T_0 = 298 K, p_0 = 1.013 bar (1 atm).

Solution. Equation 2.88 reduces to read

$$e = (h - h_0) - T_0(s - s_0) + e^{CH}$$

(a) From Table A.5, h = 2799.5 kJ/kg, s = 6.3409 kJ/kg · K. At T_0 = 298 K (25°C), water would be a liquid; thus with Equations 2.50c and 2.50d, $h_0 \approx 104.9$ kJ/kg, $s_0 \approx 0.3674$ kJ/kg · K. Table 2.12 (model I) gives e^{CH} = 45/18.02 = 2.5 kJ/kg. Then

$$e = (2799.5 - 104.9) - 298(6.3409 - 0.3674) + 2.5$$
$$= 914.5 + 2.5 = 917.0 \text{ kJ/kg}$$

Here the specific exergy is determined predominately by the physical component.

(b) Assuming the ideal gas model for methane, $h - h_0 = 0$. Also, Equation 2.58 reduces to give $s - s_0 = -R\ln p/p_0$. Then, Equation 2.88 reads

$$e = RT_0 \ln p/p_0 + e^{CH}$$

With e^{CH} = 824,350/16.04 = 51,393.4 kJ/kg from Table 2.12 (model I),

$$e = \left(\frac{8.314}{16.04}\frac{\text{kJ}}{\text{kg}\cdot\text{K}}\right)(298\text{ K})\ln\frac{5}{1.013} + 51{,}393.4\frac{\text{kJ}}{\text{kg}}$$
$$= 246.6 + 51{,}393.4$$
$$= 51{,}640 \text{ kJ/kg}$$

Here the specific exergy is determined predominately by the chemical component.

TABLE 2.12 Standard Molar Chemical Exergy, e^{CH} (kJ/kmol), of Various Substances at 298 K and p_0

Substance	Formula	Model I[a]	Model II[b]
Nitrogen	$N_2(g)$	640	720
Oxygen	$O_2(g)$	3,950	3,970
Carbon dioxide	$CO_2(g)$	14,175	19,870
Water	$H_2O(g)$	8,635	9,500
	$H_2O(l)$	45	900
Carbon (graphite)	C(s)	404,590	410,260
Hydrogen	$H_2(g)$	235,250	236,100
Sulfur	S(s)	598,160	609,600
Carbon monoxide	CO(g)	269,410	275,100
Sulfur dioxide	$SO_2(g)$	301,940	313,400
Nitrogen monoxide	NO(g)	88,850	88,900
Nitrogen dioxide	$NO_2(g)$	55,565	55,600
Hydrogen sulfide	$H_2S(g)$	799,890	812,000
Ammonia	$NH_3(g)$	336,685	337,900
Methane	$CH_4(g)$	824,350	831,650
Ethane	$C_2H_6(g)$	1,482,035	1,495,840
Methanol	$CH_3OH(g)$	715,070	722,300
	$CH_3OH(l)$	710,745	718,000
Ethyl alcohol	$C_2H_5OH(g)$	1,348,330	1,363,900
	$C_2H_5OH(l)$	1,342,085	1,357,700

[a] Ahrendts, J. 1977. Die Exergie Chemisch Reaktionsfähiger Systeme, *VDI-Forschungsheft*. VDI-Verlag, Dusseldorf, 579. Also see Reference States, *Energy — The International Journal*, 5: 667–677, 1980. In Model I, p_0 = 1.019 atm. This model attempts to impose a criterion that the reference environment be in equilibrium. The reference substances are determined assuming restricted chemical equilibrium for nitric acid and nitrates and unrestricted thermodynamic equilibrium for all other chemical components of the atmosphere, the oceans, and a portion of the Earth's crust. The chemical composition of the gas phase of this model approximates the composition of the natural atmosphere.

[b] Szargut, J., Morris, D. R., and Steward, F. R. 1988. *Energy Analysis of Thermal, Chemical, and Metallurgical Processes*. Hemisphere, New York. In Model II, p_0 = 1.0 atm. In developing this model a reference substance is selected for each chemical element from among substances that contain the element being considered and that are abundantly present in the natural environment, even though the substances are not in completely mutual stable equilibrium. An underlying rationale for this approach is that substances found abundantly in nature have little economic value. On an overall basis, the chemical composition of the exergy reference environment of Model II is closer than Model I to the composition of the natural environment, but the equilibrium criterion is not always satisfied.

The small difference between p_0 = 1.013 bar and the value of p_0 for model I has been ignored.

Exergetic Efficiency

The exergetic efficiency (second law efficiency, effectiveness, or rational efficiency) provides a true measure of the performance of a system from the thermodynamic viewpoint. To define the exergetic efficiency both a *product* and a *fuel* for the system being analyzed are identified. The product represents the desired result of the system (power, steam, some combination of power and steam, etc.). Accordingly, the definition of the product must be consistent with the purpose of purchasing and using the system.

The fuel represents the resources expended to generate the product and is not necessarily restricted to being an actual fuel such as a natural gas, oil, or coal. Both the product and the fuel are expressed in terms of exergy.

For a control volume at steady state whose exergy rate balance reads

$$\dot{E}_F = \dot{E}_P + \dot{E}_D + \dot{E}_L$$

the exergetic efficiency is

$$\varepsilon = \frac{\dot{E}_P}{\dot{E}_F} = 1 - \frac{\dot{E}_D + \dot{E}_L}{\dot{E}_F} \tag{2.91}$$

where the rates at which the fuel is supplied and the product is generated are $\dot{E}_F$ and $\dot{E}_P$, respectively. $\dot{E}_D$ and $\dot{E}_L$ denote the rates of exergy destruction and exergy loss, respectively. Exergy is destroyed by irreversibilities within the control volume, and exergy is lost from the control volume via stray heat transfer, material streams vented to the surroundings, and so on. The exergetic efficiency shows the percentage of the fuel exergy provided to a control volume that is found in the product exergy. Moreover, the difference between 100% and the value of the exergetic efficiency, expressed as a percent, is the percentage of the fuel exergy wasted in this control volume as exergy destruction and exergy loss.

To apply Equation 2.91, decisions are required concerning what are considered as the fuel and the product. Table 2.13 provides illustrations for several common components. Similar considerations are used to write exergetic efficiencies for systems consisting of several such components, as, for example, a power plant.

Exergetic efficiencies can be used to assess the thermodynamic performance of a component, plant, or industry relative to the performance of *similar* components, plants, or industries. By this means the performance of a gas turbine, for instance, can be gauged relative to the typical present-day performance level of gas turbines. A comparison of exergetic efficiencies for *dissimilar* devices — gas turbines and heat exchangers, for example — is generally not significant, however.

The exergetic efficiency is generally more meaningful, objective, and useful than other efficiencies based on the first or second law of thermodynamics, including the thermal efficiency of a power plant, the isentropic efficiency of a compressor or turbine, and the effectiveness of a heat exchanger. The thermal efficiency of a cogeneration system, for instance, is misleading because it treats both work and heat transfer as having equal thermodynamic value. The isentropic turbine efficiency (Equation 2.95a) does not consider that the working fluid at the outlet of the turbine has a higher temperature (and consequently a higher exergy that may be used in the next component) in the actual process than in the isentropic process. The heat exchanger effectiveness fails, for example, to identify the exergy destruction associated with the pressure drops of the heat exchanger working fluids.

Example 14

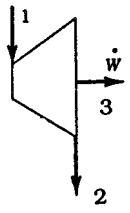
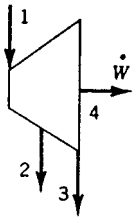
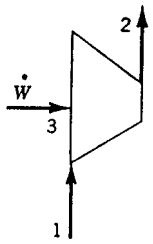
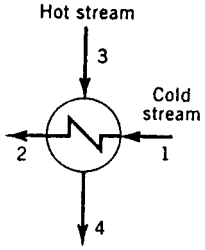
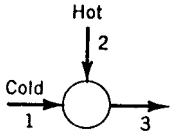
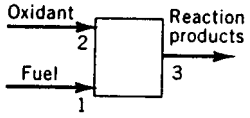
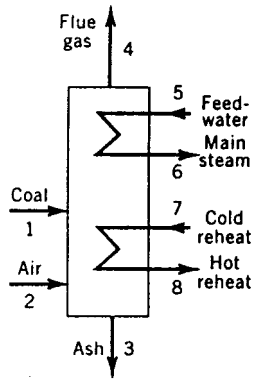
Evaluate the exergetic efficiency of the turbine in part (a) of Example 1 for T_0 = 298 K.

Solution. The exergetic efficiency from Table 2.13 is

$$\varepsilon = \frac{\dot{W}}{\dot{E}_1 - \dot{E}_2} = \frac{\dot{W}}{\dot{m}(e_1 - e_2)}$$

Using Equations 2.88 and 2.89a, and noting that the chemical exergy at 1 and 2 cancels,

TABLE 2.13 The Exergetic Efficiency for Selected Components at Steady State[a]

	Turbine or Expander	Extraction Turbine	Compressor, Pump, or Fan	Heat Exchanger[b]	Mixing Unit	Gasifier or Combustion Chamber	Boiler
Component							
E_P	W	W	$E_2 - E_1$	$E_2 - E_1$	E_3	E_3	$(E_6 - E_5) + (E_8 - E_7)$
E_F	$E_1 - E_2$	$E_1 - E_2 - E_3$	W	$E_3 - E_4$	$E_1 + E_2$	$E_1 + E_2$	$(E_1 + E_2) + (E_3 + E_4)$
ε	$\dfrac{W}{E_1 - E_2}$	$\dfrac{W}{E_1 - E_2 - E_3}$	$\dfrac{E_2 - E_1}{W}$	$\dfrac{E_2 - E_1}{E_3 - E_4}$	$\dfrac{E_3}{E_1 + E_2}$	$\dfrac{E_3}{E_1 + E_2}$	$\dfrac{(E_6 - E_5) + (E_8 - E_7)}{(E_1 + E_2) - (E_3 + E_4)}$

[a] For discussion, see Bejan et al. (1996).

[b] This definition assumes that the purpose of the heat exchanger is to heat the cold stream ($T_1 \geq T_0$). If the purpose of the heat exchanger is to provide cooling ($T_3 \geq T_0$), then the following relations should be used: $E_P = E_4 - E_3$ and $E_F = E_1 - E_2$.

$$\varepsilon = \frac{\dot{W}}{\dot{m}\left[\left(h_1 - h_2\right) - T_0\left(s_1 - s_2\right)\right]}$$

Since $\dot{W} = \dot{m}(h_1 - h_2)$,

$$\varepsilon = \frac{\dot{W}}{\dot{W} + \dot{m}T_0\left(s_2 - s_1\right)}$$

Finally, using data from Example 1 and s_2 = 6.8473 kJ/kg · K,

$$\varepsilon = \frac{30 \text{ MW}}{30 \text{ MW} + \left(\frac{162{,}357}{3600}\frac{\text{kg}}{\text{s}}\right)(298 \text{ K})(6.8473 - 6.6022)\left(\frac{\text{kJ}}{\text{kg} \cdot \text{K}}\right)\left(\frac{1 \text{ MW}}{10^3 \text{ kJ/sec}}\right)}$$

$$= \frac{30 \text{ MW}}{(30 + 3.29)\text{MW}} = 0.9(90\%)$$

Exergy Costing

Since exergy measures the true thermodynamic values of the work, heat, and other interactions between the system and its surroundings as well as the effect of irreversibilities within the system, exergy is a rational basis for assigning costs. This aspect of thermoeconomics is called *exergy costing.*

Referring to Figure 2.16 showing a steam turbine-electric generator at steady state, the total cost to produce the electricity and exiting steam equals the cost of the entering steam plus the cost of owning and operating the device. This is expressed by the *cost rate balance* for the turbine-generator:

$$\dot{C}_e + \dot{C}_2 = \dot{C}_1 + \dot{Z} \tag{2.92a}$$

where $\dot{C}_e$ is the cost rate associated with the electricity, $\dot{C}_1$ and $\dot{C}_2$ are the cost rates associated with the entering steam and exiting steam, respectively, and $\dot{Z}$ accounts for the cost rate associated with owning and operating the system, each *annualized* in $ per year.

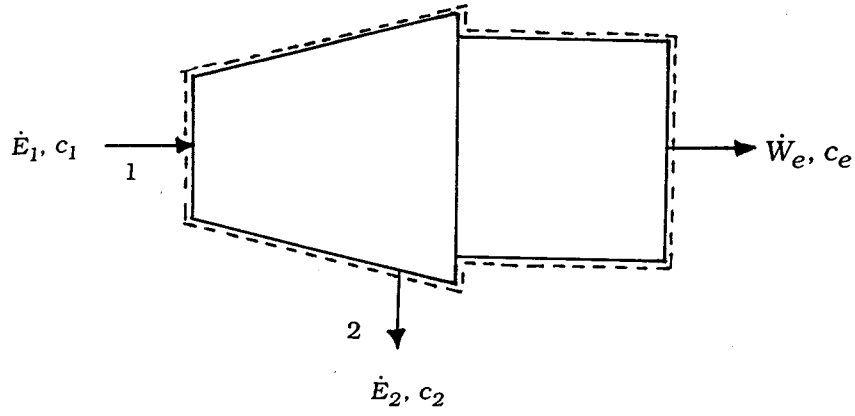


FIGURE 2.16 Steam turbine/electric generator used to discuss exergy costing.

With exergy costing, the cost rates $\dot{C}_1$, $\dot{C}_2$, and $\dot{C}_e$ are evaluated in terms of the associated rate of exergy transfer and a *unit cost.* Equation 2.92a then appears as

$$c_e \dot{W}_e + c_2 \dot{E}_2 = c_1 \dot{E}_1 + \dot{Z} \tag{2.92b}$$

The coefficients c_1, c_2, and c_e in Equation 2.92b denote the *average* cost per unit of exergy for the associated exergy rate. The unit cost c_1 of the entering steam would be obtained from exergy costing applied to the components upstream of the turbine. Assigning the same unit cost to the exiting steam: $c_2 = c_1$ on the basis that the purpose of the turbine-generator is to generate electricity and thus all costs associated with owning and operating the system should be charged to the power, Equation 2.92b becomes

$$c_e \dot{W}_e = c_1 \left(\dot{E}_1 - \dot{E}_2 \right) + \dot{Z} \tag{2.92c}$$

The first term on the right side accounts for the cost of the net exergy used and the second term accounts for cost of the system itself. Introducing the exergetic efficiency from Table 2.13, the unit cost of the electricity is

$$c_e = \frac{c_1}{\varepsilon} + \frac{\dot{Z}}{\dot{W}_e} \tag{2.93}$$

This equation shows, for example, that the unit cost of electricity would increase if the exergetic efficiency were to decrease owing to a deterioration of the turbine with use.

Example 15

A turbine-generator with an exergetic efficiency of 90% develops 7×10^7 kW · hr of electricity annually. The annual cost of owning and operating the system is $\$2.5 \times 10^5$. If the average unit cost of the steam entering the system is \$0.0165 per kW · hr of exergy, evaluate the unit cost of the electricity.

Solution. Substituting values into Equation 2.93,

$$c_e = \frac{\$0.0165/\text{kW} \cdot \text{h}}{0.9} + \frac{\$2.5 \times 10^5/\text{year}}{7 \times 10^7 \text{ kW} \cdot \text{h/year}}$$

$$= 0.0183 + 0.0036 = \$0.0219/\text{kW} \cdot \text{h}$$

2.6 Vapor and Gas Power Cycles

Vapor and gas power systems develop electrical or mechanical power from energy sources of chemical, solar, or nuclear origin. In *vapor* power systems the *working fluid*, normally water, undergoes a phase change from liquid to vapor, and conversely. In *gas* power systems, the working fluid remains a gas throughout, although the composition normally varies owing to the introduction of a fuel and subsequent combustion. The present section introduces vapor and gas power systems. Further discussion is provided in Chapter 8. Refrigeration systems are considered in Chapter 9.

The processes taking place in power systems are sufficiently complicated that idealizations are typically employed to develop tractable thermodynamic models. The *air standard analysis* of gas power systems considered later in the present section is a noteworthy example. Depending on the degree of idealization, such models may provide only qualitative information about the performance of the corresponding real-world systems. Yet such information is frequently useful in gauging how changes in major operating parameters might affect actual performance. Elementary thermodynamic models can also provide simple settings to assess, at least approximately, the advantages and disadvantages of features proposed to improve thermodynamic performance.

Rankine and Brayton Cycles

In their simplest embodiments vapor power and gas turbine power plants are represented conventionally in terms of four components in series, forming, respectively, the Rankine cycle and the Brayton cycle shown schematically in Table 2.14. The thermodynamically ideal counterparts of these cycles are composed of four internally reversible processes in series: two isentropic processes alternated with two constant pressure processes. Table 2.14 provides property diagrams of the actual and corresponding ideal cycles. Each actual cycle is denoted 1-2-3-4-1; the ideal cycle is 1-2s-3-4s-1. For simplicity, pressure drops through the boiler, condenser, and heat exchangers are not shown. Invoking Equation 2.29 for the ideal cycles, the heat added per unit of mass flowing is represented by the area *under* the isobar from state 2s to state 3: area a-2s-3-b-a. The heat rejected is the area *under* the isobar from state 4s to state 1: area a-1-4s-b-a. Enclosed area 1-2s-3-4s-1 represents the net heat added per unit of mass flowing. For any power cycle, the net heat added equals the net work done.

Expressions for the principal energy transfers shown on the schematics of Table 2.14 are provided by Equations 1 to 4 of the table. They are obtained by reducing Equation 2.27a with the assumptions of negligible heat loss and negligible changes in kinetic and potential energy from the inlet to the outlet of each component. All quantities are positive in the directions of the arrows on the figure. Using these expressions, the thermal efficiency is

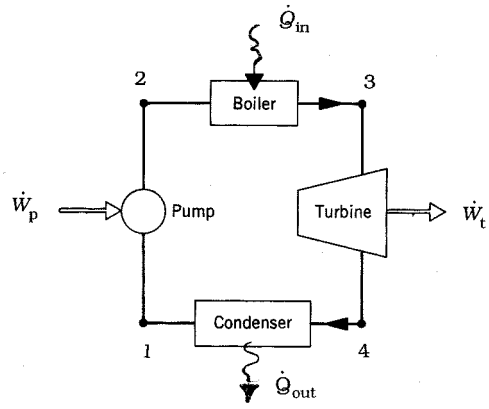
$$\eta = \frac{\left(h_3 - h_4\right) - \left(h_2 - h_1\right)}{h_3 - h_2} \tag{2.94}$$

To obtain the thermal efficiency of the ideal cycle, h_{2s} replaces h_2 and h_{4s} replaces h_4 in Equation 2.94.

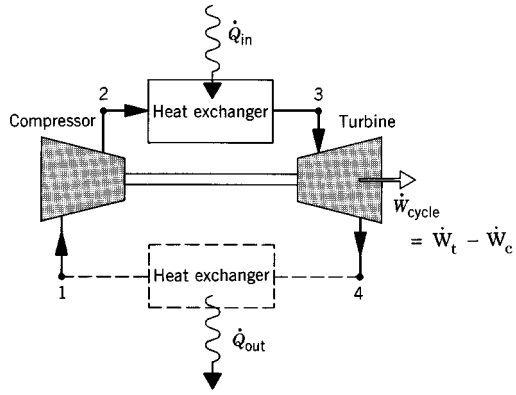
Decisions concerning cycle operating conditions normally recognize that the thermal efficiency tends to increase as the average temperature of heat addition increases and/or the temperature of heat rejection decreases. In the Rankine cycle, a high average temperature of heat addition can be achieved by superheating the vapor prior to entering the turbine, and/or by operating at an elevated steam-generator pressure. In the Brayton cycle an increase in the compressor pressure ratio p_2/p_1 tends to increase the average temperature of heat addition. Owing to materials limitations at elevated temperatures and pressures, the state of the working fluid at the turbine inlet must observe practical limits, however. The turbine inlet temperature of the Brayton cycle, for example, is controlled by providing air far in excess of what is required for combustion. In a Rankine cycle using water as the working fluid, a low temperature of heat rejection is typically achieved by operating the condenser at a pressure below 1 atm. To reduce

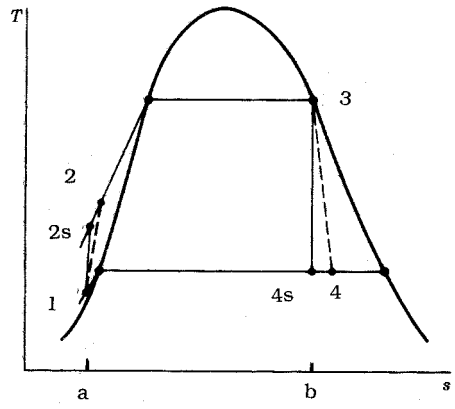
TABLE 2.14 Rankine and Brayton Cycles

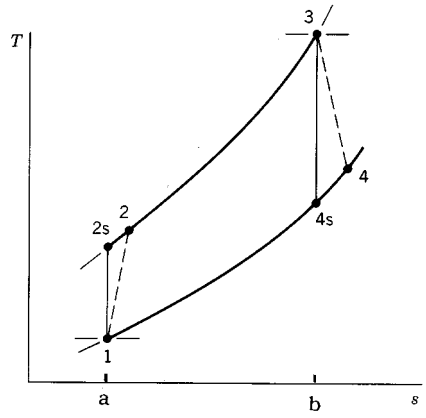
Rankine Cycle



Brayton Cycle







$$\left.\begin{matrix}\dot{W}_p \\ \dot{W}_c\end{matrix}\right\} = \dot{m}(h_2 - h_1) \qquad (>0) \tag{1}$$

$$\dot{Q}_{in} = \dot{m}(h_3 - h_2) \qquad (>0) \tag{2}$$

$$\dot{W}_t = \dot{m}(h_3 - h_4) \qquad (>0) \tag{3}$$

$$\dot{Q}_{out} = \dot{m}(h_1 - h_4) \qquad (>0) \tag{4}$$

erosion and wear by liquid droplets on the blades of the Rankine cycle steam turbine, at least 90% quality should be maintained at the turbine exit: $x_4 > 0.9$.

The *back work ratio*, bwr, is the ratio of the work required by the pump or compressor to the work developed by the turbine:

$$bwr = \frac{h_2 - h_1}{h_3 - h_4} \tag{2.95}$$

As a relatively high specific volume vapor expands through the turbine of the Rankine cycle and a much lower specific volume liquid is pumped, the back work ratio is characteristically quite low in vapor power plants — in many cases on the order of 1 to 2%. In the Brayton cycle, however, both the turbine and compressor handle a relatively high specific volume gas, and the back ratio is much larger, typically 40% or more.

The effect of friction and other irreversibilities for flow-through turbines, compressors, and pumps is commonly accounted for by an appropriate *isentropic efficiency.* The isentropic turbine efficiency is

$$\eta_t = \frac{h_3 - h_4}{h_3 - h_{4s}} \tag{2.95a}$$

The isentropic compressor efficiency is

$$\eta_c = \frac{h_{2s} - h_1}{h_2 - h_1} \tag{2.95b}$$

In the isentropic pump efficiency, η_p, which takes the same form as Equation 2.95b, the numerator is frequently approximated via Equation 2.30c *as* $h_{2s} - h_1 \approx v_1 \Delta p$, where Δp is the pressure rise across the pump.

Simple gas turbine power plants differ from the Brayton cycle model in significant respects. In actual operation, excess air is continuously drawn into the compressor, where it is compressed to a higher pressure; then fuel is introduced and combustion occurs; finally the mixture of combustion products and air expands through the turbine and is subsequently discharged to the surroundings. Accordingly, the low-temperature heat exchanger shown by a dashed line in the Brayton cycle schematic of Table 2.14 is not an actual component, but included only to account formally for the cooling in the surroundings of the hot gas discharged from the turbine.

Another frequently employed idealization used with gas turbine power plants is that of an *air-standard analysis.* An air-standard analysis involves two major assumptions: (1) as shown by the Brayton cycle schematic of Table 2.14, the temperature rise that would be brought about by combustion is effected instead by a heat transfer from an external source; (2) the working fluid throughout the cycle is air, which behaves as an ideal gas. In a *cold* air-standard analysis the specific heat ratio k for air is taken as constant. Equations 1 to 6 of Table 2.7 together with data from Table A.8 apply generally to air-standard analyses. Equations 1′ to 6′ of Table 2.7 apply to cold air-standard analyses, as does the following expression for the turbine power obtained from Table 2.1 (Equation 27c″):

$$\dot{W}_t = \dot{m} \frac{kRT_3}{k-1} \left[1 - \left(p_4 / p_3 \right)^{(k-1)/k} \right] \tag{2.96}$$

(Equation 2.96 also corresponds to Equation 5′ of Table 2.8 when $n = k$.) An expression similar in form can be written for the power required by the compressor.

For the simple Rankine and Brayton cycles of Table 2.14 the results of sample calculations are provided in Table 2.15. The Brayton cycle calculations are on an air-standard analysis basis.

Otto, Diesel, and Dual Cycles

Although most gas turbines are also internal combustion engines, the name is usually reserved to *reciprocating* internal combustion engines of the type commonly used in automobiles, trucks, and buses. Two principal types of reciprocating internal combustion engines are the *spark-ignition* engine and the *compression-ignition* engine. In a spark-ignition engine a mixture of fuel and air is ignited by a spark

TABLE 2.15 Sample Calculations for the Rankine and Brayton Cycles of Table 2.14

Rankine Cycle

Given data: $p_1 = p_4 = 8$ kPa (saturated liquid at 1)
$T_3 = 480°C$ (superheated vapor at 3)
$p_2 = p_3 = 8$ MPa
$\dot{W}_{net} = 100$ MW
Ideal cycle: $\eta_t = \eta_p = 100\%$
Actual cycle: $\eta_t = 85\%$, $\eta_p = 70\%$

Parameter	Ideal Cycle	Actual Cycle
x_4	0.794	0.873
h_2 (kJ/kg)	181.9[a]	185.4
$\dot{m}$ (kg/h)	2.86×10^5	3.38×10^5
η (%)	39.7	33.6
$\dot{Q}_{out}$ (MW)	151.9	197.6
$\dot{E}_{q,out}$ (MW)[b]	8.2	10.7

[a] $h_{2s} \approx h_1 + v_1 \Delta p$
[b] Equation 2.86 with $T_0 = 298$ K, $T_j = T_{sat}$ (8 kPa) = 315 K

Brayton Cycle

Given data: $p_1 = p_4 = 1$ bar
$p_2 = p_3 = 10$ bar
$T_3 = 1400$ K
$\eta_t = \eta_c = 100\%$

Parameter	Air-Standard Analysis	Cold Air-Standard Analysis $k = 1.4$
T_2 (K)	574.1	579.2
T_4 (K)	787.7	725.1
$\dot{W}_{net}/\dot{m}$ (kJ/kg)	427.2	397.5
η (%)	45.7	48.2
bwr	0.396	0.414

plug. In a compression ignition engine air is compressed to a high-enough pressure and temperature that combustion occurs spontaneously when fuel is injected.

In a *four-stroke* internal combustion engine, a piston executes four distinct strokes within a cylinder for every two revolutions of the crankshaft. Figure 2.17 gives *a* pressure-displacement diagram as it might be displayed electronically. With the intake valve open, the piston makes an *intake stroke* to draw a fresh charge into the cylinder. Next, with both valves closed, the piston undergoes a *compression stroke* raising the temperature and pressure of the charge. A combustion process is then initiated, resulting in a high-pressure, high-temperature gas mixture. A *power* stroke follows the compression stroke, during which the gas mixture expands and work is done on the piston. The piston then executes an *exhaust stroke* in which the burned gases are purged from the cylinder through the open exhaust valve. Smaller engines operate on *two-stroke* cycles. In two-stroke engines, the intake, compression, expansion, and

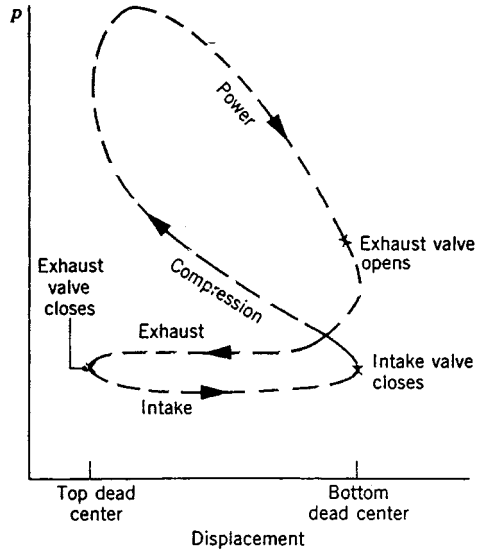


FIGURE 2.17 Pressure-displacement diagram for a reciprocating internal combustion engine.

exhaust operations are accomplished in one revolution of the crankshaft. Although internal combustion engines undergo *mechanical* cycles, the cylinder contents do not execute a *thermodynamic* cycle, since matter is introduced with one composition and is later discharged at a different composition.

A parameter used to describe the performance of reciprocating piston engines is the *mean effective pressure*, or mep. The mean effective pressure is the theoretical constant pressure that, if it acted on the piston during the power stroke, would produce the same *net* work as actually developed in one cycle. That is,

$$\text{mep} = \frac{\text{net work for one cycle}}{\text{displacement volume}} \tag{2.97}$$

where the displacement volume is the volume swept out by the piston as it moves from the top dead center to the bottom dead center. For two engines of equal displacement volume, the one with a higher mean effective pressure would produce the greater net work and, if the engines run at the same speed, greater power.

Detailed studies of the performance of reciprocating internal combustion engines may take into account many features, including the combustion process occurring within the cylinder and the effects of irreversibilities associated with friction and with pressure and temperature gradients. Heat transfer between the gases in the cylinder and the cylinder walls and the work required to charge the cylinder and exhaust the products of combustion also might be considered. Owing to these complexities, accurate modeling of reciprocating internal combustion engines normally involves computer simulation.

To conduct *elementary* thermodynamic analyses of internal combustion engines, considerable simplification is required. A procedure that allows engines to be studied *qualitatively* is to employ an *air-standard analysis* having the following elements: (1) a fixed amount of air modeled as an ideal gas is the system; (2) the combustion process is replaced by a heat transfer from an external source and generally represented in terms of elementary thermodynamic processes; (3) there are no exhaust and intake processes as in an actual engine: the cycle is completed by a constant-volume heat rejection process; (4) all processes are internally reversible.

The processes employed in air-standard analyses of internal combustion engines are selected to represent the events taking place within the engine simply and mimic the appearance of observed

pressure-displacement diagrams. In addition to the constant volume heat rejection noted previously, the compression stroke and at least a portion of the power stroke are conventionally taken as isentropic. The heat addition is normally considered to occur at constant volume, at constant pressure, or at constant volume followed by a constant pressure process, yielding, respectively, the Otto, Diesel, and Dual cycles shown in Table 2.16.

Reducing the closed system energy balance, Equation 2.8, gives the following expressions for heat and work applicable in each case shown in Table 2.16:

$$\frac{W_{12}}{m} = u_1 - u_2 \qquad (< 0)$$

$$\frac{W_{34}}{m} = u_3 - u_4 \qquad (> 0)$$

$$\frac{Q_{41}}{m} = u_1 - u_4 \qquad (< 0)$$

Table 2.16 provides additional expressions for work, heat transfer, and thermal efficiency identified with each case individually. The thermal efficiency, evaluated from Equation 2.9, takes the form

$$\eta = 1 - \frac{|Q_{41}/m|}{Q_A/m}$$

Equations 1 to 6 of Table 2.7 together with data from Table A.8, apply generally to air-standard analyses. In a *cold* air-standard analysis the specific heat ratio k for air is taken as constant. Equations 1′ to 6′ of Table 2.7 apply to cold air-standard analyses, as does Equation 4′ of Table 2.8, with $n = k$ for the isentropic processes of these cycles.

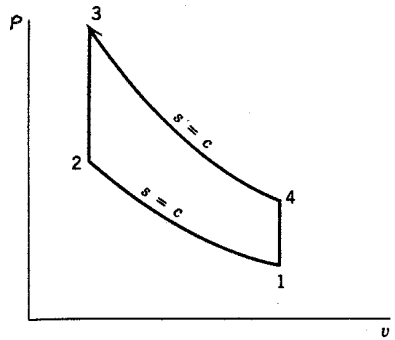
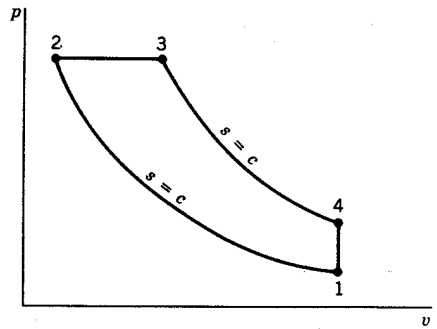
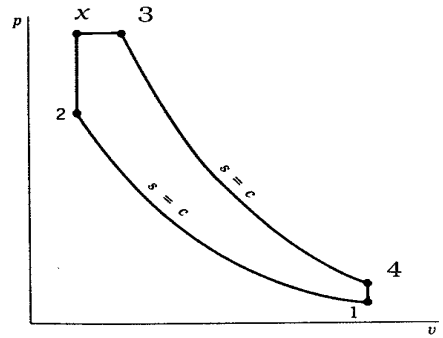
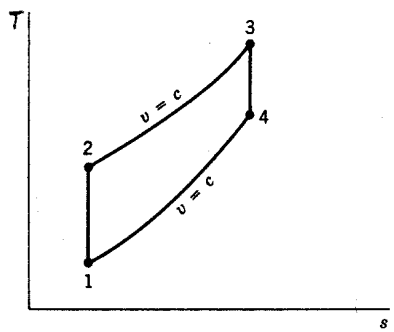
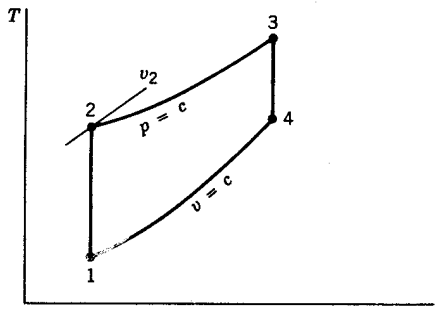
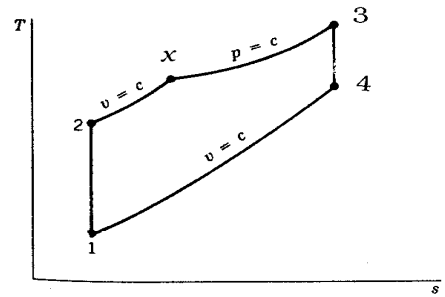
Referring to Table 2.16, the ratio v_1/v_2 is the *compression ratio*, r. For the Diesel cycle, the ratio v_3/v_2 is the *cutoff ratio*, r_c. Figure 2.18 shows the variation of the thermal efficiency with compression ratio for an Otto cycle and Diesel cycles having cutoff ratios of 2 and 3. The curves are determined on a cold air-standard basis with $k = 1.4$ using the following expression:

$$\eta = 1 - \frac{1}{r^{k-1}}\left[\frac{r_c^k - 1}{k(r_c - 1)}\right] \qquad (\text{constant } k) \tag{2.98}$$

where the Otto cycle corresponds to $r_c = 1$.

As all processes are internally reversible, areas on the *p-v* and *T-s* diagrams of Table 2.16 can be interpreted, respectively, as work and heat transfer. Invoking Equation 2.10 and referring to the *p-v* diagrams, the areas under process 3-4 of the Otto cycle, process 2-3-4 of the Diesel cycle, and process x-3-4 of the Dual cycle represent the work done by the gas during the power stroke, per unit of mass. For each cycle, the area under the isentropic process 1-2 represents the work done on the gas during the compression stroke, per unit of mass. The enclosed area of each cycle represents the net work done per unit mass. With Equation 2.15 and referring to the *T-s* diagrams, the areas under process 2-3 of the Otto and Diesel cycles and under process 2-x-3 of the Dual cycle represent the heat added per unit of mass. For each cycle, the area under the process 4-1 represent the heat rejected per unit of mass. The enclosed area of each cycle represents the net heat added, which equals the net work done, each per unit of mass.

TABLE 2.16 Otto, Diesel, and Dual Cycles

(a) Otto Cycle	(b) Diesel Cycle	(c) Dual Cycle
		
		
$\frac{W_{23}}{m} = 0$	$\frac{W_{23}}{m} = p_2(v_3 - v_2)$	$\frac{W_{2x}}{m} = 0, \quad \frac{Q_{2x}}{m} = u_x - u_2$
$\frac{Q_{23}}{m} = u_3 - u_2$	$\frac{Q_{23}}{m} = h_3 - h_2$	$\frac{W_{x3}}{m} = p_3(v_3 - v_2), \quad \frac{Q_{x3}}{m} = h_3 - h_x$
$\eta = 1 - \frac{u_4 - u_1}{u_3 - u_2}$	$\eta = 1 - \frac{u_4 - u_1}{h_3 - h_2}$	$\eta = 1 - \frac{u_4 - u_1}{(u_x - u_2) + (h_3 - h_x)}$

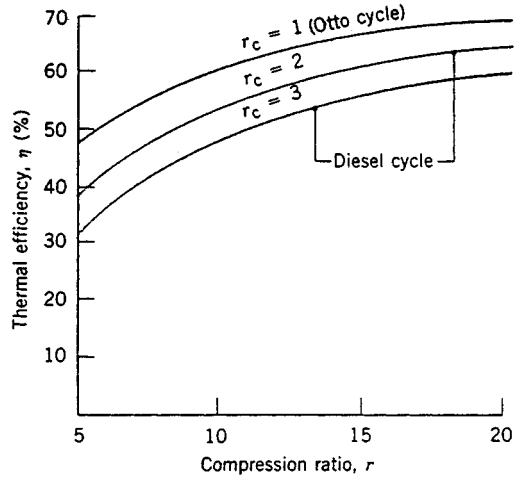


FIGURE 2.18 Thermal efficiency of the cold air-standard Otto and Diesel cycles, $k = 1.4$.

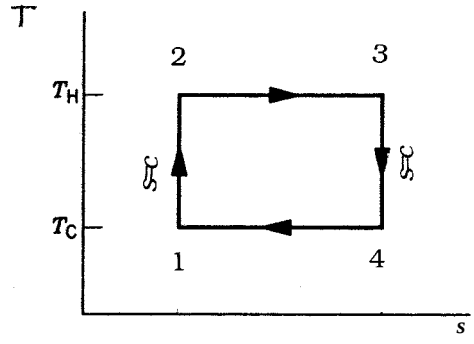
Carnot, Ericsson, and Stirling Cycles

Three thermodynamic cycles that exhibit the Carnot efficiency (Equation 2.12) are the Carnot, Ericsson, and Stirling cycles shown in Figure 2.19. Each case represents a reversible power cycle in which heat is added from an external source at a constant temperature T_H (process 2-3) and rejected to the surroundings at a constant temperature T_C (process 4-1). Carnot cycles can be configured both as vapor power cycles and as cycles executed by a gas in a piston-cylinder assembly (see, e.g., Moran and Shapiro, 1995). Carnot cycles also can be executed in systems where a capacitor is charged and discharged, a paramagnetic substance is magnetized and demagnetized, and in other ways. Regardless of the type of device and the working substance used, the Carnot cycle always has the same four internally reversible processes in series: two isentropic processes alternated with two isothermal processes.

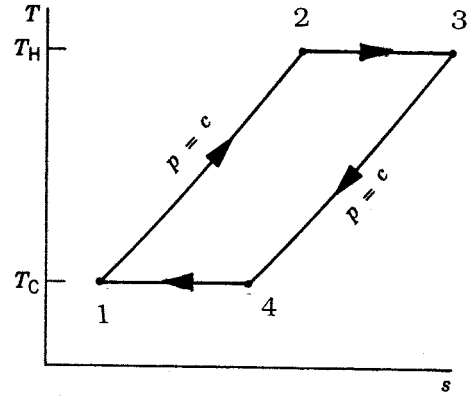
The Ericsson and Stirling cycles also consist of four internally reversible processes in series: heating from state 1 to state 2 (at constant pressure in the Ericsson cycle and at constant volume in the Stirling cycle), isothermal heating from state 2 to state 3 at temperature T_H, cooling from state 3 to state 4 (at constant pressure in the Ericsson cycle and at constant volume in the Stirling cycle), and isothermal cooling from state 4 to state 1 at temperature T_C. An ideal regenerator allows the heat input required for process 1-2 to be obtained from the heat rejected in process 3-4. Accordingly, as in the Carnot cycle all the heat added externally occurs at T_H and all of the heat rejected to the surroundings occurs at T_C.

The Ericsson and Stirling cycles are principally of theoretical interest as examples of cycles that exhibit the same thermal efficiency as the Carnot cycle: Equation 2.12. However, a practical engine of the piston-cylinder type that operates on a closed regenerative cycle having features in common with the Stirling cycle has been under study in recent years. This engine, known as the *Stirling engine*, offers the opportunity for high efficiency together with reduced emissions from combustion products because the combustion takes place externally and not within the cylinder as in internal combustion engines. In the Stirling engine, energy is transferred to the working fluid from products of combustion, which are kept separate. It is an *external* combustion engine.

(A)



(B)



(C)

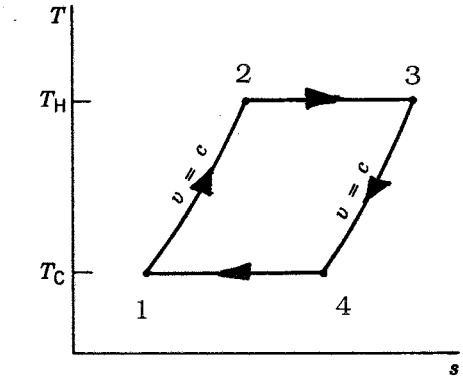


FIGURE 2.19 (A) Carnot, (B) Ericsson, and (C) Stirling cycles.

2.7 Guidelines for Improving Thermodynamic Effectiveness

Thermal design frequently aims at the most effective system from the cost viewpoint. Still, in the cost optimization process, particularly of complex energy systems, it is often expedient to begin by identifying a design that is nearly optimal thermodynamically; such a design can then be used as a point of departure for cost optimization. Presented in this section are guidelines for improving the use of fuels (natural gas, oil, and coal) by reducing sources of thermodynamic inefficiency in thermal systems. Further discussion is provided by Bejan et al. (1996).

To improve thermodynamic effectiveness it is necessary to deal directly with inefficiencies related to exergy destruction and exergy loss. The primary contributors to exergy destruction are chemical reaction, heat transfer, mixing, and friction, including unrestrained expansions of gases and liquids. To deal with them effectively, the principal sources of inefficiency not only should be understood qualitatively, but also determined quantitatively, at least approximately. Design changes to improve effectiveness must be done judiciously, however, for the cost associated with different sources of inefficiency can be different. For example, the unit cost of the electrical or mechanical power required to provide for the exergy destroyed owing to a pressure drop is generally higher than the unit cost of the fuel required for the exergy destruction caused by combustion or heat transfer.

Since chemical reaction is a significant source of thermodynamic inefficiency, it is generally good practice to minimize the use of combustion. In many applications the use of combustion equipment such as boilers is unavoidable, however. In these cases a significant reduction in the combustion irreversibility by conventional means simply cannot be expected, for the major part of the exergy destruction introduced by combustion is an inevitable consequence of incorporating such equipment. Still, the exergy destruction in practical combustion systems can be reduced by minimizing the use of excess air and by preheating the reactants. In most cases only a small part of the exergy destruction in a combustion chamber can be avoided by these means. Consequently, after considering such options for reducing the exergy destruction related to combustion, efforts to improve thermodynamic performance should focus on components of the overall system that are more amenable to betterment by cost-effective conventional measures. In other words, *some exergy destructions and energy losses can be avoided, others cannot. Efforts should be centered on those that can be avoided.*

Nonidealities associated with heat transfer also typically contribute heavily to inefficiency. Accordingly, unnecessary or cost-ineffective heat transfer must be avoided. Additional guidelines follow:

- The higher the temperature T at which a heat transfer occurs in cases where $T > T_0$, where T_0 denotes the temperature of the environment (Section 2.5), the more valuable the heat transfer and, consequently, the greater the need to avoid heat transfer to the ambient, to cooling water, or to a refrigerated stream. Heat transfer across T_0 should be avoided.
- The lower the temperature T at which a heat transfer occurs in cases where $T < T_0$, the more valuable the heat transfer and, consequently, the greater the need to avoid direct heat transfer with the ambient or a heated stream.
- Since exergy destruction associated with heat transfer between streams varies inversely with the temperature level, the lower the temperature level, the greater the need to minimize the stream-to-stream temperature difference.
- Avoid the use of intermediate heat transfer fluids when exchanging energy by heat transfer between two streams

Although irreversibilities related to friction, unrestrained expansion, and mixing are often secondary in importance to those of combustion and heat transfer, they should not be overlooked, and the following guidelines apply:

- Relatively more attention should be paid to the design of the lower temperature stages of turbines and compressors (the last stages of turbines and the first stages of compressors) than to the remaining stages of these devices.

- For turbines, compressors, and motors, consider the most thermodynamically efficient options.
- Minimize the use of throttling; check whether power recovery expanders are a cost-effective alternative for pressure reduction.
- Avoid processes using excessively large thermodynamic driving forces (differences in temperature, pressure, and chemical composition). In particular, minimize the mixing of streams differing significantly in temperature, pressure, or chemical composition.
- The greater the mass rate of flow, the greater the need to use the exergy of the stream effectively.
- The lower the temperature level, the greater the need to minimize friction.

Flowsheeting or *process simulation* software can assist efforts aimed at improving thermodynamic effectiveness by allowing engineers to readily model the behavior of an overall system, or system components, under specified conditions and do the required thermal analysis, sizing, costing, and optimization. Many of the more widely used flowsheeting programs: ASPEN PLUS, PROCESS, and CHEMCAD are of the *sequential-modular* type. SPEEDUP is a popular program of the *equation-solver* type. Since process simulation is a rapidly evolving field, vendors should be contacted for up-to-date information concerning the features of flowsheeting software, including optimization capabilities (if any). As background for further investigation of suitable software, see Biegler (1989) for a survey of the capabilities of 15 software products.

References

Ahrendts, J. 1980. Reference states. *Energy Int. J.* 5: 667–677.

ASHRAE Handbook 1993 Fundamentals. 1993. American Society of Heating, Refrigerating, and Air Conditioning Engineers, Atlanta.

ASME Steam Tables, 6th ed. 1993. ASME Press, Fairfield, NJ.

Bejan, A., Tsatsaronis, G., and Moran, M. 1996. *Thermal Design and Optimization*, John Wiley & Sons, New York.

Biegler, L.T. 1989. Chemical process simulation. *Chem. Eng. Progr.* October: 50–61.

Bird, R.B., Stewart, W.E., and Lightfoot, E.N. 1960. *Transport Phenomena*. John Wiley & Sons, New York.

Bolz, R.E. and Tuve, G.L. (eds.). 1973. *Handbook of Tables for Applied Engineering Science*, 2nd ed. CRC Press, Boca Raton, FL.

Bornakke, C. and Sonntag, R.E. 1996. *Tables of Thermodynamic and Transport Properties*. John Wiley & Sons, New York.

Cooper, H.W. and Goldfrank, J.C. 1967. B-W-R constants and new correlations. *Hydrocarbon Processing*. 46(12): 141–146.

Gray, D.E. (ed.). 1972. *American Institute of Physics Handbook*. McGraw-Hill, New York.

Haar, L. Gallagher, J.S., and Kell, G.S. 1984. *NBS/NRC Steam Tables*. Hemisphere, New York.

Handbook of Chemistry and Physics, annual editions. CRC Press, Boca Raton, FL.

JANAF Thermochemical Tables, 3rd ed. 1986. American Chemical Society and the American Institute of Physics for the National Bureau of Standards.

Jones, J.B. and Dugan, R.E. 1996. *Engineering Thermodynamics*. Prentice-Hall, Englewood Cliffs, NJ.

Keenan, J.H., Keyes, F.G., Hill, P.G., and Moore, J.G. 1969 and 1978. *Steam Tables*. John Wiley & Sons, New York (1969, English Units; 1978, SI Units).

Keenan, J.H., Chao, J., and Kaye, J. 1980 and 1983. *Gas Tables — International Version*, 2nd ed. John Wiley & Sons, New York (1980, English Units; 1983, SI Units).

Knacke, O., Kubaschewski, O., and Hesselmann, K. 1991. *Thermochemical Properties of Inorganic Substances*, 2nd ed. Springer-Verlag, Berlin.

Kotas, T.J. 1995. *The Exergy Method of Thermal Plant Analysis*, Krieger, Melbourne, FL.

Lee, B.I. and Kessler, M.G. 1975. A generalized thermodynamic correlation based on three-parameter corresponding states. *AIChE J.* 21: 510–527.

Liley, P.E. 1987. Thermodynamic properties of substances. In *Marks' Standard Handbook for Mechanical Engineers*, E.A. Avallone and T. Baumeister, (eds.). 9th ed. McGraw-Hill, New York, Sec. 4.2.

Liley, P.E., Reid, R.C., and Buck, E. 1984. Physical and chemical data. In *Perrys' Chemical Engineers, Handbook*, R.H. Perry and D.W. Green, (eds.). 6th ed. McGraw-Hill, New York, Sec. 3.

Moran, M.J. 1989. *Availability Analysis — A Guide to Efficient Energy Use*. ASME Press, New York.

Moran, M.J. and Shapiro, H.N. 1995. *Fundamentals of Engineering Thermodynamics*, 3rd ed. John Wiley & Sons, New York.

Moran, M.J. and Shapiro, H.N. 1996. *IT: Interactive Thermodynamics*. Computer software to accompany *Fundamentals of Engineering Thermodynamics*, 3rd ed. developed by Intellipro Inc., John Wiley & Sons, New York.

Obert, E.F. 1960. *Concepts of Thermodynamics*. McGraw-Hill, New York.

Preston-Thomas, H. 1990. The International Temperature Scale of 1990 (ITS-90). *Metrologia*. 27: 3–10.

Reid, R.C. and Sherwood, T.K. 1966. *The Properties of Gases and Liquids*, 2nd ed. McGraw-Hill, New York.

Reid, R.C., Prausnitz, J.M., and Poling, B.E. 1987. *The Properties of Gases and Liquids, 4th ed.* McGraw-Hill, New York.

Reynolds, W.C. 1979. *Thermodynamic Properties in SI — Graphs, Tables and Computational Equations for 40 Substances*. Department of Mechanical Engineering, Stanford University, Palo Alto, CA.

Stephan, K. 1994. Tables. In *Dubbel Handbook of Mechanical Engineering*, W. Beitz and K.-H. Kuttner, (eds.). Springer-Verlag, London, Sec. C11.

Szargut, J., Morris, D.R., and Steward, F.R. 1988. *Exergy Analysis of Thermal, Chemical and Metallurgical Processes*. Hemisphere, New York.

Van Wylen, G.J., Sonntag, R.E., and Bornakke, C. 1994. *Fundamentals of Classical Thermodynamics*, 4th ed. John Wiley & Sons, New York.

Wark, K. 1983. *Thermodynamics*, 4th ed. McGraw-Hill, New York

Zemansky, M.W. 1972. Thermodynamic symbols, definitions, and equations. In *American Institute of Physics Handbook*, D.E. Gray, (ed.). McGraw-Hill, New York, Sec. 4b.

Kreith, F.; Berger, S.A.; et. al. "Fluid Mechanics"
Mechanical Engineering Handbook
Ed. Frank Kreith
Boca Raton: CRC Press LLC, 1999

Fluid Mechanics*

Frank Kreith
University of Colorado
Stanley A. Berger
University of California, Berkeley
Stuart W. Churchill
University of Pennsylvania
J. Paul Tullis
Utah State University
Frank M. White
University of Rhode Island
Alan T. McDonald
Purdue University
Ajay Kumar
NASA Langley Research Center
John C. Chen
Lehigh University
Thomas F. Irvine, Jr.
State University of New York, Stony Brook
Massimo Capobianchi
State University of New York, Stony Brook
Francis E. Kennedy
Dartmouth College
E. Richard Booser
Consultant, Scotia, NY
Donald F. Wilcock
Tribolock, Inc.
Robert F. Boehm
University of Nevada-Las Vegas
Rolf D. Reitz
University of Wisconsin
Sherif A. Sherif
University of Florida
Bharat Bhushan
The Ohio State University

* Nomenclature for Section 3 appears at end of chapter.

3.1 Fluid Statics

Stanley A. Berger

Equilibrium of a Fluid Element

If the sum of the external forces acting on a fluid element is zero, the fluid will be either at rest or moving as a solid body — in either case, we say the fluid element is in equilibrium. In this section we consider fluids in such an equilibrium state. For fluids in equilibrium the only internal stresses acting will be normal forces, since the shear stresses depend on velocity gradients, and all such gradients, by the definition of equilibrium, are zero. If one then carries out a balance between the normal surface stresses and the body forces, assumed proportional to volume or mass, such as gravity, acting on an elementary prismatic fluid volume, the resulting equilibrium equations, after shrinking the volume to zero, show that the normal stresses at a point are the same in all directions, and since they are known to be negative, this common value is denoted by $-p$, p being the pressure.

Hydrostatic Pressure

If we carry out an equilibrium of forces on an elementary volume element $dxdydz$, the forces being pressures acting on the faces of the element and gravity acting in the $-z$ direction, we obtain

$$\frac{\partial p}{\partial x} = \frac{\partial p}{\partial y} = 0, \quad \text{and} \quad \frac{\partial p}{\partial z} = -\rho g = -\gamma \tag{3.1.1}$$

The first two of these imply that the pressure is the same in all directions at the same vertical height in a gravitational field. The third, where γ is the specific weight, shows that the pressure increases with depth in a gravitational field, the variation depending on $\rho(z)$. For homogeneous fluids, for which ρ = constant, this last equation can be integrated immediately, yielding

$$p_2 - p_1 = -\rho g\left(z_2 - z_1\right) = -\rho g\left(h_2 - h_1\right) \tag{3.1.2}$$

or

$$p_2 + \rho g h_2 = p_1 + \rho g h_1 = \text{constant} \tag{3.1.3}$$

where h denotes the elevation. These are the equations for the hydrostatic pressure distribution.

When applied to problems where a liquid, such as the ocean, lies below the atmosphere, with a constant pressure p_{atm}, h is usually measured from the ocean/atmosphere interface and p at any distance h below this interface differs from p_{atm} by an amount

$$p - p_{atm} = \rho g h \tag{3.1.4}$$

Pressures may be given either as *absolute pressure,* pressure measured relative to absolute vacuum, or *gauge pressure,* pressure measured relative to atmospheric pressure.

Manometry

The hydrostatic pressure variation may be employed to measure pressure differences in terms of heights of liquid columns — such devices are called manometers and are commonly used in wind tunnels and a host of other applications and devices. Consider, for example the U-tube manometer shown in Figure 3.1.1 filled with liquid of specific weight γ, the left leg open to the atmosphere and the right to the region whose pressure p is to be determined. In terms of the quantities shown in the figure, in the left leg

$$p_0 - \rho g h_2 = p_{atm} \tag{3.1.5a}$$

and in the right leg

$$p_0 - \rho g h_1 = p \tag{3.1.5b}$$

the difference being

$$p - p_{atm} = -\rho g\left(h_1 - h_2\right) = -\rho g d = -\gamma d \tag{3.1.6}$$

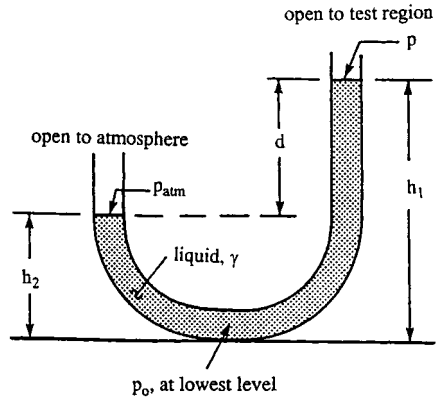


FIGURE 3.1.1 U-tube manometer.

and determining p in terms of the height difference $d = h_1 - h_2$ between the levels of the fluid in the two legs of the manometer.

Hydrostatic Forces on Submerged Objects

We now consider the force acting on a submerged object due to the hydrostatic pressure. This is given by

$$\boldsymbol{F} = \iint p\, d\boldsymbol{A} = \iint p \cdot \boldsymbol{n}\, dA = \iint \rho g h\, d\boldsymbol{A} = p_0 \iint d\boldsymbol{A} \tag{3.1.7}$$

where h is the variable vertical depth of the element $d\boldsymbol{A}$ and p_0 is the pressure at the surface. In turn we consider plane and nonplanar surfaces.

Forces on Plane Surfaces

Consider the planar surface A at an angle θ to a free surface shown in Figure 3.1.2. The force on one side of the planar surface, from Equation (3.1.7), is

$$\boldsymbol{F} = \rho g \boldsymbol{n} \iint_A h\, dA + p_0 A \boldsymbol{n} \tag{3.1.8}$$

but $h = y \sin\theta$, so

$$\iint_A h\, dA = \sin\theta \iint_A y\, dA = y_c A \sin\theta = h_c A \tag{3.1.9}$$

where the subscript c indicates the distance measured to the centroid of the area A. Thus, the total force (on one side) is

$$\boldsymbol{F} = \gamma h_c \boldsymbol{A} + p_0 \boldsymbol{A} \tag{3.1.10}$$

Thus, the magnitude of the force is independent of the angle θ, and is equal to the pressure at the centroid, $\gamma h_c + p_0$, times the area. If we use gauge pressure, the term $p_0 A$ in Equation (3.1.10) can be dropped.

Since p is not evenly distributed over A, but varies with depth, $\boldsymbol{F}$ does not act through the centroid. The point action of $\boldsymbol{F}$, called the *center of pressure*, can be determined by considering moments in Figure 3.1.2. The moment of the hydrostatic force acting on the elementary area dA about the axis perpendicular to the page passing through the point 0 on the free surface is

$$y\, dF = y\left(\gamma\, y \sin\theta\, dA\right) = \gamma\, y^2 \sin\theta\, dA \tag{3.1.11}$$

so if y_{cp} denotes the distance to the center of pressure,

$$y_{cp} F = \gamma \sin\theta \iint y^2\, dA = \gamma \sin\theta\, I_x \tag{3.1.12}$$

where I_x is the moment of inertia of the plane area with respect to the axis formed by the intersection of the plane containing the planar surface and the free surface (say $0x$). Dividing by $F = \gamma h_c A = \gamma\, y_c \sin\theta A$ gives

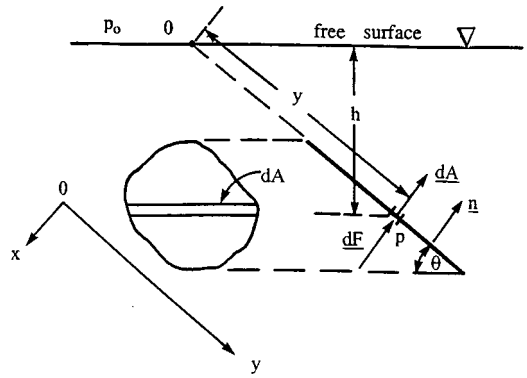


FIGURE 3.1.2 Hydrostatic force on a plane surface.

$$y_{cp} = \frac{I_x}{y_c A} \tag{3.1.13}$$

By using the parallel axis theorem $I_x = I_{xc} + Ay_c^2$, where I_{xc} is the moment of inertia with respect to an axis parallel to $0x$ passing through the centroid, Equation (3.1.13) becomes

$$y_{cp} = y_c + \frac{I_{xc}}{y_c A} \tag{3.1.14}$$

which shows that, in general, the center of pressure lies below the centroid.

Similarly, we find x_{cp} by taking moments about the y axis, specifically

$$x_{cp} F = \gamma \sin\theta \iint xy\, dA = \gamma \sin\theta\, I_{xy} \tag{3.1.15}$$

or

$$x_{cp} = \frac{I_{xy}}{y_c A} \tag{3.1.16}$$

where I_{xy} is the product of inertia with respect to the x and y axes. Again, the parallel axis theorem $I_{xy} = I_{xyc} + Ax_c y_c$, where the subscript c denotes the value at the centroid, allows Equation (3.1.16) to be written

$$x_{cp} = x_c + \frac{I_{xyc}}{y_c A} \tag{3.1.17}$$

This completes the determination of the center of pressure (x_{cp}, y_{cp}). Note that if the submerged area is symmetrical with respect to an axis passing through the centroid and parallel to either the x or y axes that $I_{xyc} = 0$ and $x_{cp} = x_c$; also that as y_c increases, $y_{cp} \to y_c$.

Centroidal moments of inertia and centroidal coordinates for some common areas are shown in Figure 3.1.3.

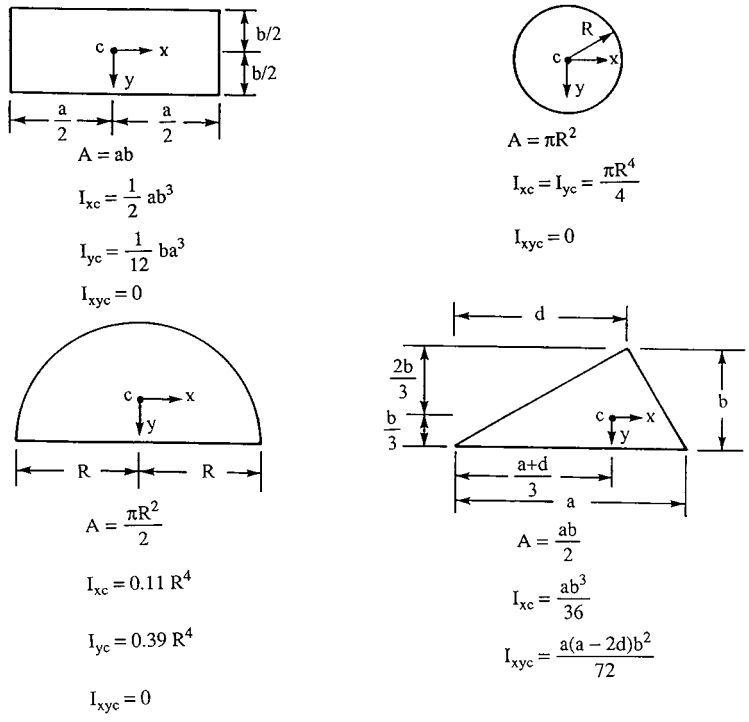


FIGURE 3.1.3 Centroidal moments of inertia and coordinates for some common areas.

Forces on Curved Surfaces

On a curved surface the forces on individual elements of area differ in direction so a simple summation of them is not generally possible, and the most convenient approach to calculating the pressure force on the surface is by separating it into its horizontal and vertical components.

A free-body diagram of the forces acting on the volume of fluid lying above a curved surface together with the conditions of static equilibrium of such a column leads to the results that:

1. The horizontal components of force on a curved submerged surface are equal to the forces exerted on the planar areas formed by the projections of the curved surface onto vertical planes normal to these components, the lines of action of these forces calculated as described earlier for planar surfaces; and
2. The vertical component of force on a curved submerged surface is equal in magnitude to the weight of the entire column of fluid lying above the curved surface, and acts through the center of mass of this volume of fluid.

Since the three components of force, two horizontal and one vertical, calculated as above, need not meet at a single point, there is, in general, no single resultant force. They can, however, be combined into a single force at any arbitrary point of application together with a moment about that point.

Hydrostatic Forces in Layered Fluids

All of the above results which employ the linear hydrostatic variation of pressure are valid only for homogeneous fluids. If the fluid is heterogeneous, consisting of individual layers each of constant density, then the pressure varies linearly with a different slope in each layer and the preceding analyses must be remedied by computing and summing the separate contributions to the forces and moments.

Buoyancy

The same principles used above to compute hydrostatic forces can be used to calculate the net pressure force acting on completely submerged or floating bodies. These laws of buoyancy, the principles of Archimedes, are that:

1. A completely submerged body experiences a vertical upward force equal to the weight of the displaced fluid; and
2. A floating or partially submerged body displaces its own weight in the fluid in which it floats (i.e., the vertical upward force is equal to the body weight).

The line of action of the buoyancy force in both (1) and (2) passes through the centroid of the displaced volume of fluid; this point is called the *center of buoyancy.* (This point need not correspond to the center of mass of the body, which could have nonuniform density. In the above it has been assumed that the displaced fluid has a constant γ. If this is not the case, such as in a layered fluid, the magnitude of the buoyant force is still equal to the weight of the displaced fluid, but the line of action of this force passes through the center of gravity of the displaced volume, not the centroid.)

If a body has a weight exactly equal to that of the volume of fluid it displaces, it is said to be *neutrally buoyant* and will remain at rest at any point where it is immersed in a (homogeneous) fluid.

Stability of Submerged and Floating Bodies

Submerged Body

A body is said to be in stable equilibrium if, when given a slight displacement from the equilibrium position, the forces thereby created tend to restore it back to its original position. The forces acting on a submerged body are the buoyancy force, F_B, acting through the center of buoyancy, denoted by CB, and the weight of the body, W, acting through the center of gravity denoted by CG (see Figure 3.1.4). We see from Figure 3.1.4 that if the CB lies above the CG a rotation from the equilibrium position creates a restoring couple which will rotate the body back to its original position — thus, this is a *stable* equilibrium situation. The reader will readily verify that when the CB lies below the CG, the couple that results from a rotation from the vertical increases the displacement from the equilibrium position — thus, this is an *unstable* equilibrium situation.

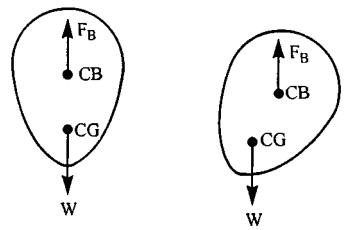


FIGURE 3.1.4 Stability for a submerged body.

Partially Submerged Body

The stability problem is more complicated for floating bodies because as the body rotates the location of the center of buoyancy may change. To determine stability in these problems requires that we determine the location of the *metacenter.* This is done for a symmetric body by tilting the body through a small angle $\Delta\theta$ from its equilibrium position and calculating the new location of the center of buoyancy CB′; the point of intersection of a vertical line drawn upward from CB′ with the line of symmetry of the floating body is the metacenter, denoted by M in Figure 3.1.5, and it is independent of $\Delta\theta$ for small angles. If M lies above the CG of the body, we see from Figure 3.1.5 that rotation of the body leads to

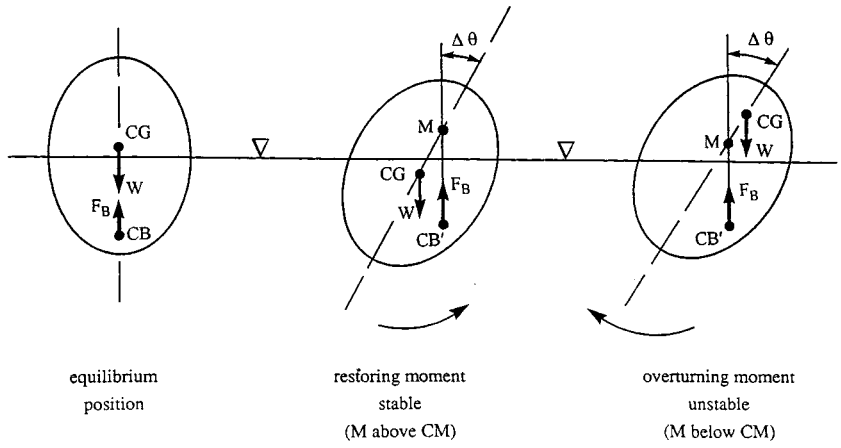


FIGURE 3.1.5 Stability for a partially submerged body.

a restoring couple, whereas *M* lying below the CG leads to a couple which will increase the displacement. Thus, the stability of the equilibrium depends on whether *M* lies above or below the CG. The directed distance from CG to *M* is called the *metacentric height*, so equivalently the equilibrium is stable if this vector is positive and unstable if it is negative; stability increases as the metacentric height increases. For geometrically complex bodies, such as ships, the computation of the metacenter can be quite complicated.

Pressure Variation in Rigid-Body Motion of a Fluid

In rigid-body motion of a fluid all the particles translate and rotate as a whole, there is no relative motion between particles, and hence no viscous stresses since these are proportional to velocity gradients. The equation of motion is then a balance among pressure, gravity, and the fluid acceleration, specifically.

$$\nabla p = \rho(\boldsymbol{g} - \boldsymbol{a}) \tag{3.1.18}$$

where $\boldsymbol{a}$ is the uniform acceleration of the body. Equation (3.1.18) shows that the lines of constant pressure, including a free surface if any, are perpendicular to the direction $\boldsymbol{g} - \boldsymbol{a}$. Two important applications of this are to a fluid in uniform linear translation and rigid-body rotation. While such problems are not, strictly speaking, fluid statics problems, their analysis and the resulting pressure variation results are similar to those for static fluids.

Uniform Linear Acceleration

For a fluid partially filling a large container moving to the right with constant acceleration $\boldsymbol{a} = (a_x, a_y)$ the geometry of Figure 3.1.6 shows that the magnitude of the pressure gradient in the direction $\boldsymbol{n}$ normal to the accelerating free surface, in the direction $\boldsymbol{g} - \boldsymbol{a}$, is

$$\frac{dp}{dn} = \rho\left[a_x^2 + \left(g + a_y\right)^2\right]^{1/2} \tag{3.1.19}$$

and the free surface is oriented at an angle to the horizontal

$$\theta = \tan^{-1}\left(\frac{a_x}{g + a_y}\right) \tag{3.1.20}$$

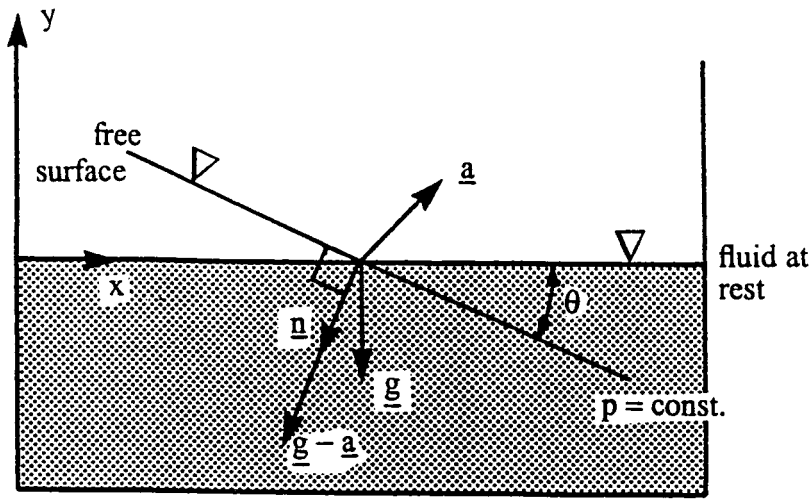


FIGURE 3.1.6 A fluid with a free surface in uniform linear acceleration.

Rigid-Body Rotation

Consider the fluid-filled circular cylinder rotating uniformly with angular velocity $\Omega = \Omega \boldsymbol{e}_r$ (Figure 3.1.7). The only acceleration is the centripetal acceleration $\Omega \times \Omega \times \boldsymbol{r}) = -r\Omega^2 \boldsymbol{e}_r$, so Equation 3.1.18 becomes

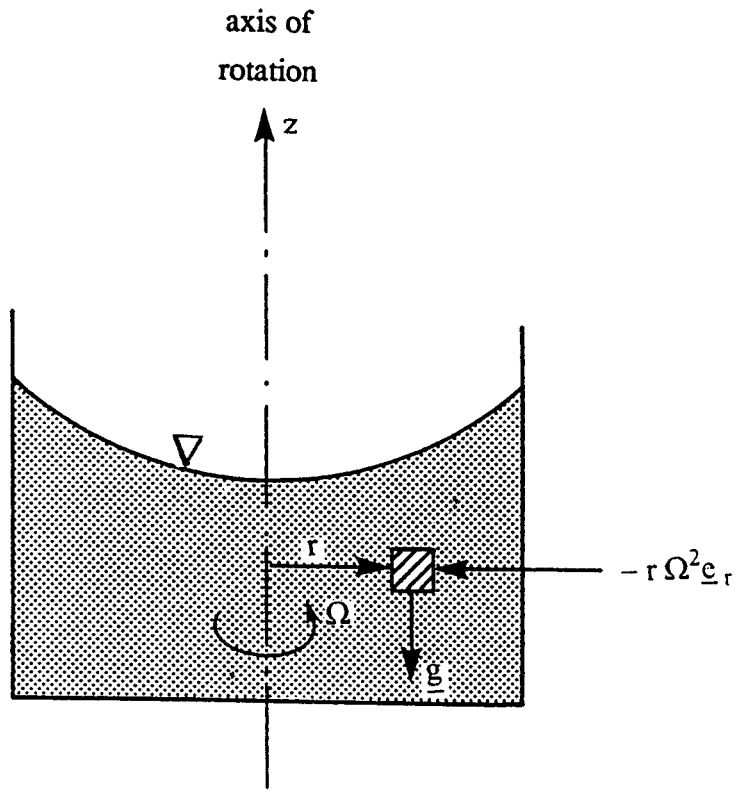


FIGURE 3.1.7 A fluid with a free surface in rigid-body rotation.

$$\nabla p = \frac{\partial p}{\partial r}\boldsymbol{e}_r + \frac{\partial p}{\partial z}\boldsymbol{e}_z = \rho(\boldsymbol{g} - \boldsymbol{a}) = \rho\left(r\Omega^2\boldsymbol{e}_r - g\boldsymbol{e}_z\right) \tag{3.1.21}$$

or

$$\frac{\partial p}{\partial r} = \rho r\Omega^2, \qquad \frac{\partial p}{\partial z} = -\rho g = -\gamma \tag{3.1.22}$$

Integration of these equations leads to

$$p = p_o - \gamma z + \frac{1}{2}\rho r^2\Omega^2 \tag{3.1.23}$$

where p_o is some reference pressure. This result shows that at any fixed r the pressure varies hydrostatically in the vertical direction, while the constant pressure surfaces, including the free surface, are paraboloids of revolution.

Further Information

The reader may find more detail and additional information on the topics in this section in any one of the many excellent introductory texts on fluid mechanics, such as

White, F.M. 1994. *Fluid Mechanics,* 3rd ed., McGraw-Hill, New York.
Munson, B.R., Young, D.F., and Okiishi, T.H. 1994. *Fundamentals of Fluid Mechanics,* 2nd ed., John Wiley & Sons, New York.

3.2 Equations of Motion and Potential Flow

Stanley A. Berger

Integral Relations for a Control Volume

Like most physical conservation laws those governing motion of a fluid apply to material particles or systems of such particles. This so-called Lagrangian viewpoint is generally not as useful in practical fluid flows as an analysis through fixed (or deformable) control volumes — the Eulerian viewpoint. The relationship between these two viewpoints can be deduced from the Reynolds transport theorem, from which we also most readily derive the governing integral and differential equations of motion.

Reynolds Transport Theorem

The *extensive* quantity B, a scalar, vector, or tensor, is defined as any property of the fluid (e.g., momentum, energy) and b as the corresponding value per unit mass (the *intensive* value). The Reynolds transport theorem for a moving and arbitrarily deforming control volume CV, with boundary CS (see Figure 3.2.1), states that

$$\frac{d}{dt}\left(B_{\text{system}}\right) = \frac{d}{dt}\left(\iiint_{\text{CV}} \rho b \, d\upsilon\right) + \iint_{\text{CS}} \rho b\left(\boldsymbol{V}_r \cdot \boldsymbol{n}\right) dA \tag{3.2.1}$$

where B_{system} is the total quantity of B in the system (any mass of fixed identity), $\boldsymbol{n}$ is the outward normal to the CS, $\boldsymbol{V}_r = \boldsymbol{V}(\boldsymbol{r}, t) - \boldsymbol{V}_{\text{CS}}(\boldsymbol{r}, t)$, the velocity of the fluid particle, $\boldsymbol{V}(\boldsymbol{r}, t)$, relative to that of the CS, $\boldsymbol{V}_{\text{CS}}(\boldsymbol{r}, t)$, and d/dt on the left-hand side is the derivative following the fluid particles, i.e., the fluid mass comprising the system. The theorem states that the time rate of change of the total B in the system is equal to the rate of change within the CV plus the net flux of B through the CS. To distinguish between the d/dt which appears on the two sides of Equation (3.2.1) but which have different interpretations, the derivative on the left-hand side, following the system, is denoted by D/Dt and is called the material derivative. This notation is used in what follows. For any function $f(x, y, z, t)$,

$$\frac{Df}{Dt} = \frac{\partial f}{\partial t} + \boldsymbol{V} \cdot \nabla f$$

For a CV fixed with respect to the reference frame, Equation (3.2.1) reduces to

$$\frac{D}{Dt}\left(B_{\text{system}}\right) = \frac{d}{dt}\underset{\substack{\text{CV}\\(\text{fixed})}}{\iiint}\left(\rho b\right) d\upsilon + \iint_{\text{CS}} \rho b\left(\boldsymbol{V} \cdot \boldsymbol{n}\right) dA \tag{3.2.2}$$

(The time derivative operator in the first term on the right-hand side may be moved inside the integral, in which case it is then to be interpreted as the partial derivative $\partial/\partial t$.)

Conservation of Mass

If we apply Equation (3.2.2) for a fixed control volume, with B_{system} the total mass in the system, then since conservation of mass requires that $DB_{\text{system}}/Dt = 0$ there follows, since $b = B_{\text{system}}/m = 1$,

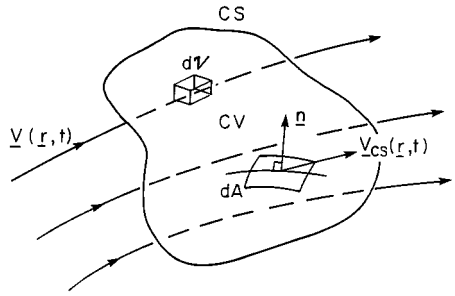


FIGURE 3.2.1 Control volume.

$$\iiint_{\substack{CV\\(fixed)}} \frac{\partial \rho}{\partial t}\, d\upsilon + \iint_{CS} \rho(\boldsymbol{V}\cdot\boldsymbol{n})\, dA = 0 \tag{3.2.3}$$

This is the integral form of the conservation of mass law for a fixed control volume. For a steady flow, Equation (3.2.3) reduces to

$$\iint_{CS} \rho(\boldsymbol{V}\cdot\boldsymbol{n})\, dA = 0 \tag{3.2.4}$$

whether compressible or incompressible. For an incompressible flow, ρ = constant, so

$$\iint_{CS} (\boldsymbol{V}\cdot\boldsymbol{n})\, dA = 0 \tag{3.2.5}$$

whether the flow is steady or unsteady.

Conservation of Momentum

The conservation of (linear) momentum states that

$$F_{total} \equiv \sum (\text{external forces acting on the fluid system}) = \frac{D\boldsymbol{M}}{Dt} \equiv \frac{D}{Dt}\left(\iiint_{system} \rho \boldsymbol{V}\, d\upsilon\right) \tag{3.2.6}$$

where $\boldsymbol{M}$ is the total system momentum. For an arbitrarily moving, deformable control volume it then follows from Equation (3.2.1) with b set to $\boldsymbol{V}$,

$$\boldsymbol{F}_{total} = \frac{d}{dt}\left(\iiint_{CV} \rho \boldsymbol{V}\, d\upsilon\right) + \iint_{CS} \rho \boldsymbol{V}(\boldsymbol{V}_r\cdot\boldsymbol{n})\, dA \tag{3.2.7}$$

This expression is only valid in an inertial coordinate frame. To write the equivalent expression for a noninertial frame we must use the relationship between the acceleration $\boldsymbol{a}_I$ in an inertial frame and the acceleration $\boldsymbol{a}_R$ in a noninertial frame,

$$\boldsymbol{a}_I = \boldsymbol{a}_R + \frac{d^2\boldsymbol{R}}{dt^2} + 2\Omega \times \boldsymbol{V} + \Omega \times (\Omega \times \boldsymbol{r}) + \frac{d\Omega}{dt} \times \boldsymbol{r} \tag{3.2.8}$$

where $\boldsymbol{R}$ is the position vector of the origin of the noninertial frame with respect to that of the inertial frame, Ω is the angular velocity of the noninertial frame, and $\boldsymbol{r}$ and $\boldsymbol{V}$ the position and velocity vectors in the noninertial frame. The third term on the right-hand side of Equation (3.2.8) is the Coriolis acceleration, and the fourth term is the centrifugal acceleration. For a noninertial frame Equation (3.2.7) is then

$$\begin{aligned} \boldsymbol{F}_{\text{total}} - \iiint_{\text{system}} \left[\frac{d^2\boldsymbol{R}}{dt^2} + 2\Omega \times \boldsymbol{V} + \Omega \times (\Omega \times \boldsymbol{r}) + \frac{d\Omega}{dt} \times \boldsymbol{r} \right] \rho \, d\upsilon &= \frac{D}{Dt}\left(\iiint_{\text{system}} \rho \boldsymbol{V} \, d\upsilon \right) \\ = \frac{d}{dt}\left(\iiint_{\text{CV}} \rho \boldsymbol{V} \, d\upsilon \right) + \iint_{\text{CS}} \rho \boldsymbol{V} \cdot (\boldsymbol{V}_r \cdot \boldsymbol{n}) \, dA & \end{aligned} \tag{3.2.9}$$

where the frame acceleration terms of Equation (3.2.8) have been brought to the left-hand side because to an observer in the noninertial frame they act as "apparent" body forces.

For a fixed control volume in an inertial frame for steady flow it follows from the above that

$$\boldsymbol{F}_{\text{total}} = \iint_{\text{CS}} \rho \boldsymbol{V} (\boldsymbol{V} \cdot \boldsymbol{n}) \, dA \tag{3.2.10}$$

This expression is the basis of many control volume analyses for fluid flow problems.

The cross product of $\boldsymbol{r}$, the position vector with respect to a convenient origin, with the momentum Equation (3.2.6) written for an elementary particle of mass dm, noting that $(d\boldsymbol{r}/dt) \times \boldsymbol{V} = 0$, leads to the integral moment of momentum equation

$$\sum \boldsymbol{M} - \boldsymbol{M}_I = \frac{D}{Dt} \iiint_{\text{system}} \rho (\boldsymbol{r} \times \boldsymbol{V}) \, d\upsilon \tag{3.2.11}$$

where $\Sigma \boldsymbol{M}$ is the sum of the moments of all the external forces acting on the system about the origin of $\boldsymbol{r}$, and $\boldsymbol{M}_I$ is the moment of the apparent body forces (see Equation (3.2.9)). The right-hand side can be written for a control volume using the appropriate form of the Reynolds transport theorem.

Conservation of Energy

The conservation of energy law follows from the first law of thermodynamics for a moving system

$$\dot{Q} - \dot{W} = \frac{D}{Dt}\left(\iiint_{\text{system}} \rho e \, d\upsilon \right) \tag{3.2.12}$$

where $\dot{Q}$ is the rate at which heat is added to the system, $\dot{W}$ the rate at which the system works on its surroundings, and e is the total energy per unit mass. For a particle of mass dm the contributions to the specific energy e are the internal energy u, the kinetic energy $V^2/2$, and the potential energy, which in the case of gravity, the only body force we shall consider, is gz, where z is the vertical displacement opposite to the direction of gravity. (We assume no energy transfer owing to chemical reaction as well

as no magnetic or electric fields.) For a fixed control volume it then follows from Equation (3.2.2) [with $b = e = u + (V^2/2) + gz$] that

$$\dot{Q} - \dot{W} = \frac{d}{dt}\left(\iiint_{CV} \rho\left(u + \frac{1}{2}V^2 + gz\right) d\upsilon\right) + \iint_{CS} \rho\left(u + \frac{1}{2}V^2 + gz\right)(\boldsymbol{V}\cdot\boldsymbol{n})\, dA \tag{3.2.13}$$

Problem

An incompressible fluid flows through a pump at a volumetric flow rate $\dot{Q}$. The (head) loss between sections 1 and 2 (see Figure 3.2.2) is equal to $\beta\rho V_1^2/2$ (V is the average velocity at the section). Calculate the power that must be delivered by the pump to the fluid to produce a given increase in pressure, $\Delta p = p_2 - p_1$.

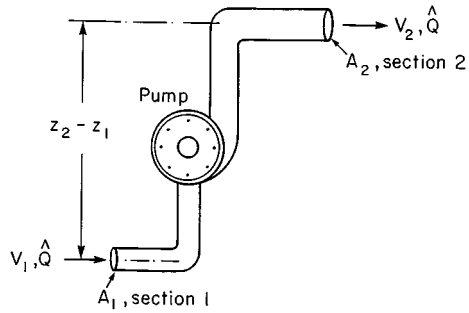


FIGURE 3.2.2 Pump producing pressure increase.

Solution: The principal equation needed is the energy Equation (3.2.13). The term $\dot{W}$, the rate at which the system does work on its surroundings, for such problems has the form

$$\dot{W} = -\dot{W}_{shaft} + \iint_{CS} p\boldsymbol{V}\cdot\boldsymbol{n}\, dA \tag{P.3.2.1}$$

where $\dot{W}_{shaft}$ represents the work done on the fluid by a moving shaft, such as by turbines, propellers, fans, etc., and the second term on the right side represents the rate of working by the normal stress, the pressure, at the boundary. For a steady flow in a control volume coincident with the physical system boundaries and bounded at its ends by sections 1 and 2, Equation (3.2.13) reduces to ($u \equiv 0$),

$$\dot{Q} + \dot{W}_{shaft} - \iint_{CS} p\boldsymbol{V}\cdot\boldsymbol{n}\, dA = \iint_{CS}\left(\frac{1}{2}\rho V^2 + \gamma z\right)(\boldsymbol{V}\cdot\boldsymbol{n})\, dA \tag{P.3.2.2}$$

Using average quantities at sections 1 and 2, and the continuity Equation (3.2.5), which reduces in this case to

$$V_1 A_1 = V_2 A_2 = \dot{Q} \tag{P.3.2.3}$$

We can write Equation (P.3.2.2) as

$$\dot{Q} + \dot{W}_{shaft} - (p_2 - p_1)\dot{Q} = \left[\frac{1}{2}\rho\left(V_2^2 - V_1^2\right) + \gamma(z_2 - z_1)\right]\dot{Q} \tag{P.3.2.4}$$

$\dot{Q}$, the rate at which heat is added to the system, is here equal to $-\beta\rho V_1^2/2$, the head loss between sections 1 and 2. Equation (P.3.2.4) then can be rewritten

$$\dot{W}_{\text{shaft}} = \beta\rho\frac{V_1^2}{2} + (\Delta p)\dot{Q} + \frac{1}{2}\rho\left(V_2^2 - V_1^2\right)\dot{Q} + \gamma\left(z_2 - z_1\right)\dot{Q}$$

or, in terms of the given quantities,

$$\dot{W}_{\text{shaft}} = \frac{\beta\rho\dot{Q}^2}{A_1^2} + (\Delta p)\dot{Q} + \frac{1}{2}\rho\frac{\dot{Q}^3}{A_2^2}\left(1 - \frac{A_2^2}{A_1^2}\right) + \gamma\left(z_2 - z_1\right)\dot{Q} \qquad \text{(P.3.2.5)}$$

Thus, for example, if the fluid is water ($\rho \approx 1000$ kg/m^3, $\gamma = 9.8$ kN/m^3), $\dot{Q} = 0.5$ m^3/sec, the heat loss is $0.2\rho V_1^2/2$, and $\Delta p = p_2 - p_1 = 2 \times 10^5$N/m^2 = 200 kPa, $A_1 = 0.1$ m^2 = $A_2/2$, $(z_2 - z_1) = 2$ m, we find, using Equation (P.3.2.5)

$$\dot{W}_{\text{shaft}} = \frac{0.2(1000)(0.5)^2}{(0.1)^2} + \left(2\times10^5\right)(0.5) + \frac{1}{2}(1000)\frac{(0.5)^3}{(0.2)^2}(1-4) + \left(9.8\times10^3\right)(2)(0.5)$$

$$= 5{,}000 + 10{,}000 - 4{,}688 + 9{,}800 = 20{,}112 \text{ Nm/sec}$$

$$= 20{,}112 \text{ W} = \frac{20{,}112}{745.7}\text{hp} = 27 \text{ hp}$$

Differential Relations for Fluid Motion

In the previous section the conservation laws were derived in integral form. These forms are useful in calculating, generally using a control volume analysis, gross features of a flow. Such analyses usually require some *a priori* knowledge or assumptions about the flow. In any case, an approach based on integral conservation laws cannot be used to determine the point-by-point variation of the dependent variables, such as velocity, pressure, temperature, etc. To do this requires the use of the differential forms of the conservation laws, which are presented below.

Mass Conservation–Continuity Equation

Applying Gauss's theorem (the divergence theorem) to Equation (3.2.3) we obtain

$$\iiint_{\substack{\text{CV}\\(\text{fixed})}}\left[\frac{\partial\rho}{\partial t} + \nabla\cdot(\rho V)\right]d\upsilon = 0 \qquad (3.2.14)$$

which, because the control volume is arbitrary, immediately yields

$$\frac{\partial\rho}{\partial t} + \nabla\cdot(\rho\boldsymbol{V}) = 0 \qquad (3.2.15)$$

This can also be written as

$$\frac{D\rho}{Dt} + \rho\nabla\cdot\boldsymbol{V} = 0 \qquad (3.2.16)$$

using the fact that

$$\frac{D\rho}{Dt} = \frac{\partial \rho}{\partial t} + \boldsymbol{V} \cdot \nabla \rho \tag{3.2.17}$$

Special cases:

1. Steady flow $[(\partial/\partial t)(\ ) \equiv 0]$

$$\nabla \cdot (\rho \boldsymbol{V}) = 0 \tag{3.2.18}$$

2. Incompressible flow $(D\rho/Dt \equiv 0)$

$$\nabla \cdot \boldsymbol{V} = 0 \tag{3.2.19}$$

Momentum Conservation

We note first, as a consequence of mass conservation for a system, that the right-hand side of Equation (3.2.6) can be written as

$$\frac{D}{Dt}\left(\iiint_{\text{system}} \rho \boldsymbol{V}\, d\upsilon\right) \equiv \iiint_{\text{system}} \rho \frac{D\boldsymbol{V}}{Dt}\, d\upsilon \tag{3.2.20}$$

The total force acting on the system which appears on the left-hand side of Equation (3.2.6) is the sum of body forces $\boldsymbol{F}_b$ and surface forces $\boldsymbol{F}_s$. The body forces are often given as forces per unit mass (e.g., gravity), and so can be written

$$\boldsymbol{F}_b = \iiint_{\text{system}} \rho \boldsymbol{f}\, d\upsilon \tag{3.2.21}$$

The surface forces are represented in terms of the second-order stress tensor* $\underline{\underline{\sigma}} = \{\sigma_{ij}\}$, where σ_{ij} is defined as the force per unit area in the i direction on a planar element whose normal lies in the j direction. From elementary angular momentum considerations for an infinitesimal volume it can be shown that σ_{ij} is a symmetric tensor, and therefore has only six independent components. The total surface force exerted on the system by its surroundings is then

$$\boldsymbol{F}_s = \iint_{\substack{\text{system}\\ \text{surface}}} \underline{\underline{\sigma}} \cdot \boldsymbol{n}\, dA, \quad \text{with } i-\text{component } F_{s_i} = \iint \sigma_{ij} n_j\, dA \tag{3.2.22}$$

The integral momentum conservation law Equation (3.2.6) can then be written

$$\iiint_{\text{system}} \rho \frac{D\boldsymbol{V}}{Dt}\, d\upsilon = \iiint_{\text{system}} \rho \boldsymbol{f}\, d\upsilon + \iint_{\substack{\text{system}\\ \text{surface}}} \underline{\underline{\sigma}} \cdot \boldsymbol{n}\, dA \tag{3.2.23}$$

* We shall assume the reader is familiar with elementary Cartesian tensor analysis and the associated subscript notation and conventions. The reader for whom this is not true should skip the details and concentrate on the final principal results and equations given at the ends of the next few subsections.

The application of the divergence theorem to the last term on the right-side of Equation (3.2.23) leads to

$$\iiint_{\text{system}} \rho \frac{DV}{Dt}\, d\upsilon = \iiint_{\text{system}} \rho f\, d\upsilon + \iiint_{\text{system}} \nabla \cdot \underline{\underline{\sigma}}\, d\upsilon \tag{3.2.24}$$

where $\nabla \cdot \underline{\underline{\sigma}} \equiv \{\partial \sigma_{ij}/x_j\}$. Since Equation (3.2.24) holds for any material volume, it follows that

$$\rho \frac{DV}{Dt} = \rho \boldsymbol{f} + \nabla \cdot \underline{\underline{\sigma}} \tag{3.2.25}$$

(With the decomposition of $\boldsymbol{F}_{\text{total}}$ above, Equation (3.2.10) can be written

$$\iiint_{CV} \rho f\, d\upsilon + \iint_{CS} \underline{\underline{\sigma}} \cdot \boldsymbol{n}\, dA = \iint_{CS} \rho \boldsymbol{V}(\boldsymbol{V} \cdot \boldsymbol{n})\, dA \tag{3.2.26}$$

If ρ is uniform and $\boldsymbol{f}$ is a conservative body force, i.e., $\boldsymbol{f} = -\nabla\Psi$, where Ψ is the force potential, then Equation (3.2.26), after application of the divergence theorem to the body force term, can be written

$$\iint_{CS} \left(-\rho\Psi \boldsymbol{n} + \underline{\underline{\sigma}} \cdot \boldsymbol{n}\right) dA = \iint_{CS} \rho \boldsymbol{V}(\boldsymbol{V} \cdot \boldsymbol{n})\, dA \tag{3.2.27}$$

It is in this form, involving only integrals over the surface of the control volume, that the integral form of the momentum equation is used in control volume analyses, particularly in the case when the body force term is absent.)

Analysis of Rate of Deformation

The principal aim of the following two subsections is to derive a relationship between the stress and the rate of strain to be used in the momentum Equation (3.2.25). The reader less familiar with tensor notation may skip these sections, apart from noting some of the terms and quantities defined therein, and proceed directly to Equations (3.2.38) or (3.2.39).

The relative motion of two neighboring points P and Q, separated by a distance η, can be written (using $\boldsymbol{u}$ for the local velocity)

$$\boldsymbol{u}(Q) = \boldsymbol{u}(P) + (\nabla \boldsymbol{u})\eta$$

or, equivalently, writing $\nabla\boldsymbol{u}$ as the sum of antisymmetric and symmetric tensors,

$$\boldsymbol{u}(Q) = \boldsymbol{u}(P) + \frac{1}{2}\left((\nabla \boldsymbol{u}) - (\nabla \boldsymbol{u})^*\right)\eta + \frac{1}{2}\left((\nabla \boldsymbol{u}) + (\nabla \boldsymbol{u})^*\right)\eta \tag{3.2.28}$$

where $\nabla\boldsymbol{u} = \{\partial u_i/\partial x_j\}$, and the superscript * denotes transpose, so $(\nabla\boldsymbol{u})^* = \{\partial u_j/\partial x_i\}$. The second term on the right-hand side of Equation (3.2.28) can be rewritten in terms of the *vorticity*, $\nabla \times \boldsymbol{u}$, so Equation (3.2.28) becomes

$$\boldsymbol{u}(Q) = \boldsymbol{u}(P) + \frac{1}{2}(\nabla \times \boldsymbol{u}) \times \eta + \frac{1}{2}\left((\nabla \boldsymbol{u}) + (\nabla \boldsymbol{u})^*\right)\eta \tag{3.2.29}$$

which shows that the local rate of deformation consists of a rigid-body translation, a rigid-body rotation with angular velocity $^1/_2\,(\nabla \times \boldsymbol{u})$, and a velocity or rate of deformation. The coefficient of η in the last term in Equation (3.2.29) is defined as the rate-of-strain tensor and is denoted by $\underline{\underline{e}}$, in subscript form

$$e_{ij} = \frac{1}{2}\left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i}\right) \tag{3.2.30}$$

From $\underline{\underline{e}}$ we can define a rate-of-strain central quadric, along the principal axes of which the deforming motion consists of a pure extension or compression.

Relationship Between Forces and Rate of Deformation

We are now in a position to determine the required relationship between the stress tensor $\underline{\underline{\sigma}}$ and the rate of deformation. Assuming that in a static fluid the stress reduces to a (negative) hydrostatic or thermodynamic pressure, equal in all directions, we can write

$$\underline{\underline{\sigma}} = -p\underline{\underline{I}} + \underline{\underline{\tau}} \quad \text{or} \quad \underline{\underline{\sigma}}_{ij} = -p\delta_{ij} + \tau_{ij} \tag{3.2.31}$$

where $\underline{\underline{\tau}}$ is the viscous part of the total stress and is called the deviatoric stress tensor, $\underline{\underline{I}}$ is the identity tensor, and δ_{ij} is the corresponding Kronecker delta ($\delta_{ij} = 0$ if $i \neq j$; $\delta_{ij} = 1$ if $i = j$). We make further assumptions that (1) the fluid exhibits no preferred directions; (2) the stress is independent of any previous history of distortion; and (3) that the stress depends only on the local thermodynamic state and the kinematic state of the immediate neighborhood. Precisely, we assume that $\underline{\underline{\tau}}$ is linearly proportional to the first spatial derivatives of $\boldsymbol{u}$, the coefficient of proportionality depending only on the local thermodynamic state. These assumptions and the relations below which follow from them are appropriate for a Newtonian fluid. Most common fluids, such as air and water under most conditions, are Newtonian, but there are many other fluids, including many which arise in industrial applications, which exhibit so-called non-Newtonian properties. The study of such non-Newtonian fluids, such as viscoelastic fluids, is the subject of the field of rheology.

With the Newtonian fluid assumptions above, and the symmetry of $\underline{\underline{\tau}}$ which follows from the symmetry of $\underline{\underline{\sigma}}$, one can show that the viscous part $\underline{\underline{\tau}}$ of the total stress can be written as

$$\underline{\underline{\tau}} = \lambda(\nabla \cdot \boldsymbol{u})\underline{\underline{I}} + 2\mu\underline{\underline{e}} \tag{3.2.32}$$

so the total stress for a Newtonian fluid is

$$\underline{\underline{\sigma}} = -p\underline{\underline{I}} + \lambda(\nabla \cdot \boldsymbol{u})\underline{\underline{I}} + 2\mu\underline{\underline{e}} \tag{3.2.33}$$

or, in subscript notation

$$\sigma_{ij} = -p\delta_{ij} + \lambda\left(\frac{\partial u_k}{\partial x_k}\right)\delta_{ij} + \mu\left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i}\right) \tag{3.2.34}$$

(the Einstein summation convention is assumed here, namely, that a repeated subscript, such as in the second term on the right-hand side above, is summed over; note also that $\nabla \cdot \boldsymbol{u} = \partial u_k/\partial x_k = e_{kk}$.) The coefficient λ is called the "second viscosity" and μ the "absolute viscosity," or more commonly the "dynamic viscosity," or simply the "viscosity." For a Newtonian fluid λ and μ depend only on local thermodynamic state, primarily on the temperature.

We note, from Equation (3.2.34), that whereas in a fluid at rest the pressure is an isotropic normal stress, this is not the case for a moving fluid, since in general $\sigma_{11} \neq \sigma_{22} \neq \sigma_{33}$. To have an analogous quantity to p for a moving fluid we define the pressure in a moving fluid as the negative mean normal stress, denoted, say, by $\bar{p}$

$$\bar{p} = -\frac{1}{3}\sigma_{ii} \tag{3.2.35}$$

(σ_{ii} is the trace of $\underline{\underline{\sigma}}$ and an invariant of $\underline{\underline{\sigma}}$, independent of the orientation of the axes). From Equation (3.2.34)

$$\bar{p} = -\frac{1}{3}\sigma_{ii} = p - \left(\lambda + \frac{2}{3}\mu\right)\nabla \cdot \boldsymbol{u} \tag{3.2.36}$$

For an incompressible fluid $\nabla \cdot \boldsymbol{u} = 0$ and hence $\bar{p} \equiv p$. The quantity $(\lambda + {}^2/_3\mu)$ is called the bulk viscosity. If one assumes that the deviatoric stress tensor τ_{ij} makes no contribution to the mean normal stress, it follows that $\lambda + {}^2/_3\mu = 0$, so again $\bar{p} = p$. This condition, $\lambda = -{}^2/_3\mu$, is called the Stokes assumption or hypothesis. If neither the incompressibility nor the Stokes assumptions are made, the difference between $\bar{p}$ and p is usually still negligibly small because $(\lambda + {}^2/_3\mu)\,\nabla \cdot \boldsymbol{u} << p$ in most fluid flow problems. If the Stokes hypothesis is made, as is often the case in fluid mechanics, Equation (3.2.34) becomes

$$\sigma_{ij} = -p\delta_{ij} + 2\mu\left(e_{ij} - \frac{1}{3}e_{kk}\delta_{ij}\right) \tag{3.2.37}$$

The Navier–Stokes Equations

Substitution of Equation (3.2.33) into (3.2.25), since $\nabla \cdot (\phi \underline{\underline{I}}) = \nabla\phi$, for any scalar function ϕ, yields (replacing $\boldsymbol{u}$ in Equation (3.2.33) by $\boldsymbol{V}$)

$$\rho\frac{D\boldsymbol{V}}{Dt} = \rho\boldsymbol{f} - \nabla p + \nabla\left(\lambda\nabla \cdot \boldsymbol{V}\right) + \nabla \cdot \left(2\mu\underline{\underline{e}}\right) \tag{3.2.38}$$

These equations are the Navier–Stokes equations (although the name is as often given to the full set of governing conservation equations). With the Stokes assumption ($\lambda = -{}^2/_3\mu$), Equation (3.2.38) becomes

$$\rho\frac{D\boldsymbol{V}}{Dt} = \rho\boldsymbol{f} - \nabla p + \nabla \cdot \left[2\mu\left(\underline{\underline{e}} - \frac{1}{3}e_{kk}\underline{\underline{I}}\right)\right] \tag{3.2.39}$$

If the Eulerian frame is not an inertial frame, then one must use the transformation to an inertial frame either using Equation (3.2.8) or the "apparent" body force formulation, Equation (3.2.9).

Energy Conservation — The Mechanical and Thermal Energy Equations

In deriving the differential form of the energy equation we begin by assuming that heat enters or leaves the material or control volume by heat conduction across the boundaries, the heat flux per unit area being $\boldsymbol{q}$. It then follows that

$$\dot{Q} = -\iint \boldsymbol{q} \cdot \boldsymbol{n}\, dA = -\iiint \nabla \cdot \boldsymbol{q}\, d\upsilon \tag{3.2.40}$$

The work-rate term $\dot{W}$ can be decomposed into the rate of work done against body forces, given by

$$-\iiint \rho \boldsymbol{f} \cdot \boldsymbol{V}\, d\upsilon \tag{3.2.41}$$

and the rate of work done against surface stresses, given by

$$-\iint_{\substack{\text{system}\\ \text{surface}}} \boldsymbol{V} \cdot \left(\underline{\underline{\sigma}}\boldsymbol{n}\right) dA \tag{3.2.42}$$

Substitution of these expressions for $\dot{Q}$ and $\dot{W}$ into Equation (3.2.12), use of the divergence theorem, and conservation of mass lead to

$$\rho \frac{D}{Dt}\left(u + \frac{1}{2}V^2\right) = -\nabla \cdot \boldsymbol{q} + \rho \boldsymbol{f} \cdot \boldsymbol{V} + \nabla \cdot \left(\boldsymbol{V}\underline{\underline{\sigma}}\right) \tag{3.2.43}$$

(note that a potential energy term is no longer included in e, the total specific energy, as it is accounted for by the body force rate-of-working term $\rho \boldsymbol{f} \cdot \boldsymbol{V}$).

Equation (3.2.43) is the total energy equation showing how the energy changes as a result of working by the body and surface forces and heat transfer. It is often useful to have a purely thermal energy equation. This is obtained by subtracting from Equation (3.2.43) the dot product of $\boldsymbol{V}$ with the momentum Equation (3.2.25), after expanding the last term in Equation (3.2.43), resulting in

$$\rho \frac{Du}{Dt} = \frac{\partial V_i}{\partial x_j}\sigma_{ij} - \nabla \cdot \boldsymbol{q} \tag{3.2.44}$$

With $\sigma_{ij} = -p\delta_{ij} + \tau_{ij}$, and the use of the continuity equation in the form of Equation (3.2.16), the first term on the right-hand side of Equation (3.2.44) may be written

$$\frac{\partial V_i}{\partial x_j}\sigma_{ij} = -\rho \frac{D\left(\frac{p}{\rho}\right)}{Dt} + \frac{Dp}{Dt} + \Phi \tag{3.2.45}$$

where Φ is the rate of dissipation of mechanical energy per unit mass due to viscosity, and is given by

$$\Phi \equiv \frac{\partial V_i}{\partial x_j}\tau_{ij} = 2\mu\left(e_{ij}e_{ij} - \frac{1}{3}e_{kk}^2\right) = 2\mu\left(e_{ij} - \frac{1}{3}e_{kk}\delta_{ij}\right)^2 \tag{3.2.46}$$

With the introduction of Equation (3.2.45), Equation (3.2.44) becomes

$$\rho \frac{De}{Dt} = -p\nabla \cdot \boldsymbol{V} + \Phi - \nabla \cdot \boldsymbol{q} \tag{3.2.47}$$

or

$$\rho \frac{Dh}{Dt} = \frac{Dp}{Dt} + \Phi - \nabla \cdot \boldsymbol{q} \tag{3.2.48}$$

where $h = e + (p/\rho)$ is the specific enthalpy. Unlike the other terms on the right-hand side of Equation (3.2.47), which can be negative or positive, Φ is always nonnegative and represents the increase in internal energy (or enthalpy) owing to irreversible degradation of mechanical energy. Finally, from elementary thermodynamic considerations

$$\frac{Dh}{Dt} = T\frac{DS}{Dt} + \frac{1}{\rho}\frac{Dp}{Dt}$$

where S is the entropy, so Equation (3.2.48) can be written

$$\rho T\frac{DS}{Dt} = \Phi - \nabla\cdot\boldsymbol{q} \tag{3.2.49}$$

If the heat conduction is assumed to obey the Fourier heat conduction law, so $\boldsymbol{q} = -k\nabla T$, where k is the thermal conductivity, then in all of the above equations

$$-\nabla\cdot\boldsymbol{q} = \nabla\cdot(k\nabla T) = k\nabla^2 T \tag{3.2.50}$$

the last of these equalities holding only if k = constant.

In the event the thermodynamic quantities vary little, the coefficients of the constitutive relations for $\underline{\underline{\sigma}}$ and $\boldsymbol{q}$ may be taken to be constant and the above equations simplified accordingly.

We note also that if the flow is incompressible, then the mass conservation, or continuity, equation simplifies to

$$\nabla\cdot\boldsymbol{V} = 0 \tag{3.2.51}$$

and the momentum Equation (3.2.38) to

$$\rho\frac{D\boldsymbol{V}}{Dt} = \rho\boldsymbol{f} - \nabla p + \mu\nabla^2\boldsymbol{V} \tag{3.2.52}$$

where ∇^2 is the Laplacian operator. The small temperature changes, compatible with the incompressibility assumption, are then determined, for a perfect gas with constant k and specific heats, by the energy equation rewritten for the temperature, in the form

$$\rho c_v\frac{DT}{Dt} = k\nabla^2 T + \Phi \tag{3.2.53}$$

Boundary Conditions

The appropriate boundary conditions to be applied at the boundary of a fluid in contact with another medium depends on the nature of this other medium — solid, liquid, or gas. We discuss a few of the more important cases here in turn:

1. *At a solid surface:* $\boldsymbol{V}$ and T are continuous. Contained in this boundary condition is the "no-slip" condition, namely, that the tangential velocity of the fluid in contact with the boundary of the solid is equal to that of the boundary. For an inviscid fluid the no-slip condition does not apply, and only the normal component of velocity is continuous. If the wall is permeable, the tangential velocity is continuous and the normal velocity is arbitrary; the temperature boundary condition for this case depends on the nature of the injection or suction at the wall.

2. *At a liquid/gas interface:* For such cases the appropriate boundary conditions depend on what can be assumed about the gas the liquid is in contact with. In the classical liquid free-surface problem, the gas, generally atmospheric air, can be ignored and the necessary boundary conditions are that (a) the normal velocity in the liquid at the interface is equal to the normal velocity of the interface and (b) the pressure in the liquid at the interface exceeds the atmospheric pressure by an amount equal to

$$\Delta p = p_{\text{liquid}} - p_{\text{atm}} = \sigma\left(\frac{1}{R_1} + \frac{1}{R_2}\right) \tag{3.2.54}$$

where R_1 and R_2 are the radii of curvature of the intercepts of the interface by two orthogonal planes containing the vertical axis. If the gas is a vapor which undergoes nonnegligible interaction and exchanges with the liquid in contact with it, the boundary conditions are more complex. Then, in addition to the above conditions on normal velocity and pressure, the shear stress (momentum flux) and heat flux must be continuous as well.

For interfaces in general the boundary conditions are derived from continuity conditions for each "transportable" quantity, namely continuity of the appropriate intensity across the interface and continuity of the normal component of the flux vector. Fluid momentum and heat are two such transportable quantities, the associated intensities are velocity and temperature, and the associated flux vectors are stress and heat flux. (The reader should be aware of circumstances where these simple criteria do not apply, for example, the velocity slip and temperature jump for a rarefied gas in contact with a solid surface.)

Vorticity in Incompressible Flow

With μ = constant, ρ = constant, and $\boldsymbol{f} = -\boldsymbol{g} = -g\boldsymbol{k}$ the momentum equation reduces to the form (see Equation (3.2.52))

$$\rho\frac{D\boldsymbol{V}}{Dt} = -\nabla p - \rho g\boldsymbol{k} + \mu\nabla^2\boldsymbol{V} \tag{3.2.55}$$

With the vector identities

$$(\boldsymbol{V}\cdot\nabla)\boldsymbol{V} = \nabla\left(\frac{V^2}{2}\right) - \boldsymbol{V}\times(\nabla\times\boldsymbol{V}) \tag{3.2.56}$$

and

$$\nabla^2\boldsymbol{V} = \nabla(\nabla\cdot\boldsymbol{V}) - \nabla\times(\nabla\times\boldsymbol{V}) \tag{3.2.57}$$

and defining the *vorticity*

$$\boldsymbol{\zeta} \equiv \nabla\times\boldsymbol{V} \tag{3.2.58}$$

Equation (3.2.55) can be written, noting that for incompressible flow $\nabla\cdot\boldsymbol{V} = 0$,

$$\rho\frac{\partial\boldsymbol{V}}{\partial t} + \nabla\left(p + \frac{1}{2}\rho V^2 + \rho gz\right) = \rho\boldsymbol{V}\times\boldsymbol{\zeta} - \mu\nabla\times\boldsymbol{\zeta} \tag{3.2.59}$$

The flow is said to be *irrotational* if

$$\zeta \equiv \nabla \times \boldsymbol{V} = 0 \tag{3.2.60}$$

from which it follows that a *velocity potential* Φ can be defined

$$\boldsymbol{V} = \nabla\Phi \tag{3.2.61}$$

Setting $\zeta = 0$ in Equation (3.2.59), using Equation (3.2.61), and then integrating with respect to all the spatial variables, leads to

$$\rho\frac{\partial\Phi}{\partial t} + \left(p + \frac{1}{2}\rho V^2 + \rho g z\right) = F(t) \tag{3.2.62}$$

(the arbitrary function $F(t)$ introduced by the integration can either be absorbed in Φ, or is determined by the boundary conditions). Equation (3.2.62) is the unsteady *Bernoulli equation* for irrotational, incompressible flow. (Irrotational flows are always potential flows, even if the flow is compressible. Because the viscous term in Equation (3.2.59) vanishes identically for $\zeta = 0$, it would appear that the above Bernoulli equation is valid even for viscous flow. Potential solutions of hydrodynamics are in fact exact solutions of the full Navier–Stokes equations. Such solutions, however, are not valid near solid boundaries or bodies because the no-slip condition generates vorticity and causes nonzero ζ; the potential flow solution is invalid in all those parts of the flow field that have been "contaminated" by the spread of the vorticity by convection and diffusion. See below.)

The curl of Equation (3.2.59), noting that the curl of any gradient is zero, leads to

$$\rho\frac{\partial\zeta}{\partial t} = \rho\nabla \times (\boldsymbol{V} \times \zeta) - \mu\nabla \times \nabla \times \zeta \tag{3.2.63}$$

but

$$\begin{aligned}\nabla^2\zeta &= \nabla(\nabla \cdot \zeta) - \nabla \times \nabla \times \zeta \\ &= -\nabla \times \nabla \times \zeta\end{aligned} \tag{3.2.64}$$

since div curl ( ) $\equiv 0$, and therefore also

$$\nabla \times (\boldsymbol{V} \times \zeta) \equiv \zeta(\nabla\boldsymbol{V}) + \boldsymbol{V}\nabla \cdot \zeta - \boldsymbol{V}\nabla\zeta - \zeta\nabla \cdot \boldsymbol{V} \tag{3.2.65}$$

$$= \zeta(\nabla\boldsymbol{V}) - \boldsymbol{V}\nabla\zeta \tag{3.2.66}$$

Equation (3.2.63) can then be written

$$\frac{D\zeta}{Dt} = (\zeta \cdot \nabla)\boldsymbol{V} + \nu\nabla^2\zeta \tag{3.2.67}$$

where $\nu = \mu/\rho$ is the kinematic viscosity. Equation (3.2.67) is the vorticity equation for incompressible flow. The first term on the right, an inviscid term, increases the vorticity by vortex stretching. In inviscid, two-dimensional flow both terms on the right-hand side of Equation (3.2.67) vanish, and the equation reduces to $D\zeta/Dt = 0$, from which it follows that the vorticity of a fluid particle remains constant as it moves. This is Helmholtz's theorem. As a consequence it also follows that if $\zeta = 0$ initially, $\zeta \equiv 0$ always;

i.e., initially irrotational flows remain irrotational (for inviscid flow). Similarly, it can be proved that $D\Gamma/Dt = 0$; i.e., the circulation around a material closed circuit remains constant, which is Kelvin's theorem.

If $\nu \neq 0$, Equation (3.2.67) shows that the vorticity generated, say, at solid boundaries, diffuses and stretches as it is convected.

We also note that for steady flow the Bernoulli equation reduces to

$$p + \frac{1}{2}\rho V^2 + \rho g z = \text{constant} \tag{3.2.68}$$

valid for steady, irrotational, incompressible flow.

Stream Function

For two-dimensional flows the continuity equation, e.g., for plane, incompressible flows ($\boldsymbol{V} = (u, \nu)$)

$$\frac{\partial u}{\partial x} + \frac{\partial \nu}{\partial y} = 0 \tag{3.2.69}$$

can be identically satisfied by introducing a stream function ψ, defined by

$$u = \frac{\partial \psi}{\partial y}, \qquad \nu = -\frac{\partial \psi}{\partial x} \tag{3.2.70}$$

Physically ψ is a measure of the flow between streamlines. (Stream functions can be similarly defined to satisfy identically the continuity equations for incompressible cylindrical and spherical axisymmetric flows; and for these flows, as well as the above planar flow, also when they are compressible, but only then if they are steady.) Continuing with the planar case, we note that in such flows there is only a single nonzero component of vorticity, given by

$$\boldsymbol{\zeta} = \left(0, 0, \zeta_z\right) = \left(0, 0, \frac{\partial \nu}{\partial x} - \frac{\partial u}{\partial y}\right) \tag{3.2.71}$$

With Equation (3.2.70)

$$\zeta_z = -\frac{\partial^2 \psi}{\partial x^2} - \frac{\partial^2 \psi}{\partial y^2} = -\nabla^2 \psi \tag{3.2.72}$$

For this two-dimensional flow Equation (3.2.67) reduces to

$$\frac{\partial \zeta_z}{\partial t} + u\frac{\partial \zeta_z}{\partial x} + \nu\frac{\partial \zeta_z}{\partial y} = \nu\left(\frac{\partial^2 \zeta_z}{\partial x^2} + \frac{\partial^2 \zeta_z}{\partial y^2}\right) \tag{3.2.73}$$

Substitution of Equation (3.2.72) into Equation (3.2.73) yields an equation for the stream function

$$\frac{\partial\left(\nabla^2 \psi\right)}{\partial t} + \frac{\partial \psi}{\partial y}\frac{\partial\left(\nabla^2 \psi\right)}{\partial x} - \frac{\partial \psi}{\partial x}\frac{\partial}{\partial y}\left(\nabla^2 \psi\right) = \nu \nabla^4 \psi \tag{3.2.74}$$

where $\nabla^4 = \nabla^2(\nabla^2)$. For uniform flow past a solid body, for example, this equation for Ψ would be solved subject to the boundary conditions:

$$\begin{aligned} &\frac{\partial \psi}{\partial x} = 0, \quad \frac{\partial \psi}{\partial y} = V_\infty \quad \text{at infinity} \\ &\frac{\partial \psi}{\partial x} = 0, \quad \frac{\partial \psi}{\partial y} = 0 \quad \text{at the body (no-slip)} \end{aligned} \tag{3.2.75}$$

For the special case of irrotational flow it follows immediately from Equations (3.2.70) and (3.2.71) with $\zeta_z = 0$, that ψ satisfies the Laplace equation

$$\nabla^2 \psi = \frac{\partial^2 \psi}{\partial x^2} + \frac{\partial^2 \psi}{\partial y^2} = 0 \tag{3.2.76}$$

Inviscid Irrotational Flow: Potential Flow

For irrotational flows we have already noted that a velocity potential Φ can be defined such that $\boldsymbol{V} = \nabla\Phi$. If the flow is also incompressible, so $\nabla \cdot \boldsymbol{V} = 0$, it then follows that

$$\nabla \cdot (\nabla \Phi) = \nabla^2 \Phi = 0 \tag{3.2.77}$$

so Φ satisfies Laplace's equation. (Note that unlike the stream function ψ, which can only be defined for two-dimensional flows, the above considerations for Φ apply to flow in two and three dimensions. On the other hand, the existence of ψ does not require the flow to be irrotational, whereas the existence of Φ does.)

Since Equation (3.2.77) with appropriate conditions on $\boldsymbol{V}$ at boundaries of the flow completely determines the velocity field, and the momentum equation has played no role in this determination, we see that inviscid irrotational flow — *potential theory* — is a purely kinematic theory. The momentum equation only enters after Φ is known in order to calculate the pressure field consistent with the velocity field $\boldsymbol{V} = \nabla\Phi$.

For both two- and three-dimensional flows the determination of Φ makes use of the powerful techniques of potential theory, well developed in the mathematical literature. For two-dimensional planar flows the techniques of complex variable theory are available, since Φ may be considered as either the real or imaginary part of an analytic function (the same being true for ψ, since for such two-dimensional flows Φ and ψ are conjugate variables.)

Because the Laplace equation, obeyed by both Φ and ψ, is linear, complex flows may be built up from the superposition of simple flows; this property of inviscid irrotational flows underlies nearly all solution techniques in this area of fluid mechanics.

Problem

A two-dimensional inviscid irrotational flow has the velocity potential

$$\Phi = x^2 - y^2 \tag{P.3.2.6}$$

What two-dimensional potential flow does this represent?

Solution. It follows from Equations (3.2.61) and (3.2.70) that for two-dimensional flows, in general

$$u = \frac{\partial \Phi}{\partial x} = \frac{\partial \psi}{\partial y}, \quad v = \frac{\partial \Phi}{\partial y} = -\frac{\partial \psi}{\partial x} \tag{P.3.2.7}$$

It follows from using Equation (P.3.2.6) that

$$u = \frac{\partial \psi}{\partial y} = 2x, \quad v = -\frac{\partial \psi}{\partial x} = -2y \tag{P.3.2.8}$$

Integration of Equation (P.3.2.8) yields

$$\psi = 2xy \tag{P.3.2.9}$$

The streamlines, ψ = constant, and equipotential lines, Φ = constant, both families of hyperbolas and each family the orthogonal trajectory of the other, are shown in Figure 3.2.3. Because the x and y axes are streamlines, Equations (P.3.2.6) and (P.3.2.9) represent the inviscid irrotational flow in a right-angle corner. By symmetry, they also represent the planar flow in the upper half-plane directed toward a stagnation point at $x = y = 0$ (see Figure 3.2.4). In polar coordinates (r, θ), with corresponding velocity components (u_r, u_θ), this flow is represented by

$$\Phi = r^2 \cos 2\theta, \quad \psi = r^2 \sin 2\theta \tag{P.3.2.10}$$

with

$$\begin{aligned} u_r &= \frac{\partial \Phi}{\partial r} = \frac{1}{r}\frac{\partial \psi}{\partial \theta} = 2r\cos 2\theta \\ u_\theta &= \frac{1}{r}\frac{\partial \Phi}{\partial \theta} = -\frac{\partial \psi}{\partial r} = -2r\sin 2\theta \end{aligned} \tag{P.3.2.11}$$

For two-dimensional planar potential flows we may also use complex variables, writing the complex potential $f(z) = \Phi + i\psi$ as a function of the complex variable $z = x + iy$, where the complex velocity is given by $f'(z) = w(z) = u - iv$. For the flow above

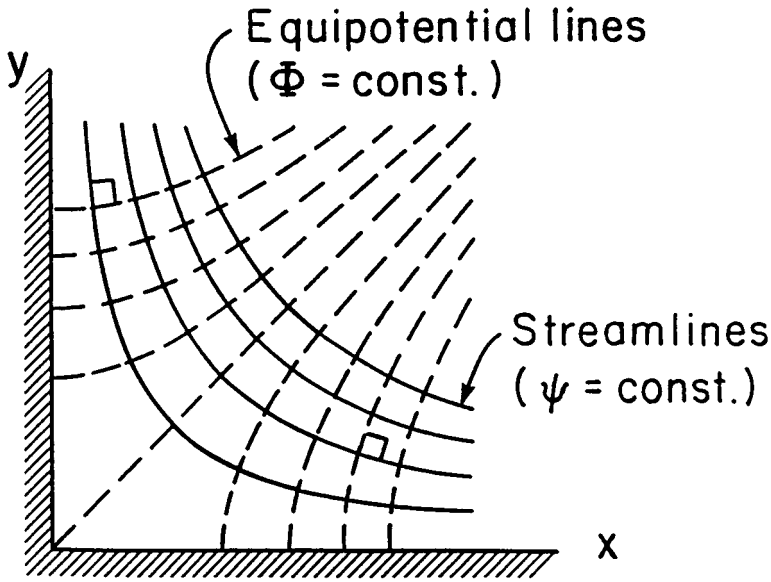


FIGURE 3.2.3 Potential flow in a 90° corner.

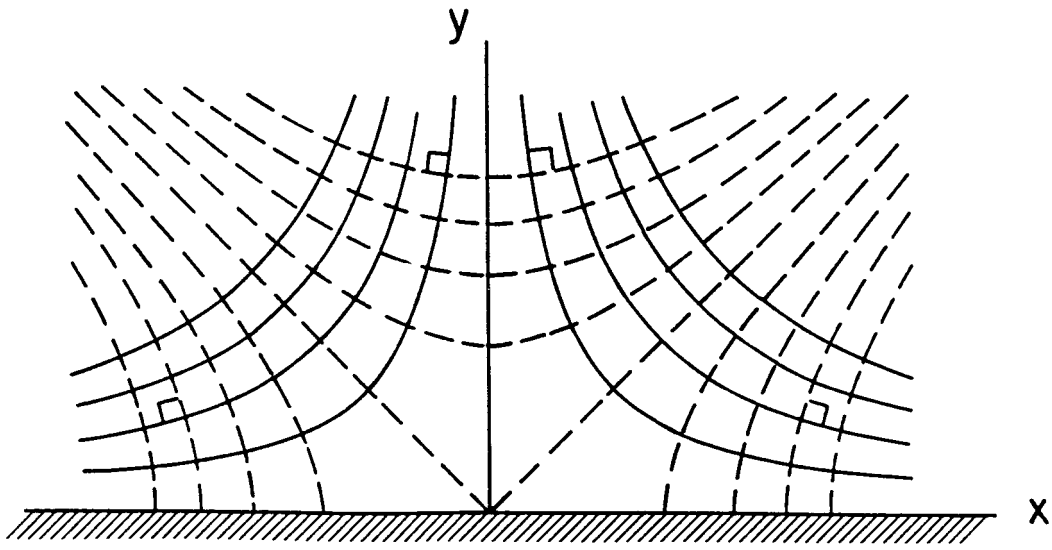


FIGURE 3.2.4 Potential flow impinging against a flat (180°) wall (plane stagnation-point flow).

$$f(z) = z^2 \tag{P.3.2.12}$$

Expressions such as Equation (P.3.2.12), where the right-hand side is an analytic function of z, may also be regarded as a conformal mapping, which makes available as an aid in solving two-dimensional potential problems all the tools of this branch of mathematics.

Further Information

More detail and additional information on the topics in this section may be found in more advanced books on fluid dynamics, such as

Batchelor, G.K. 1967. *An Introduction to Fluid Dynamics,* Cambridge University Press, Cambridge, England.

Warsi, Z.U.A. 1993. *Fluid Dynamics. Theoretical and Computational Approaches,* CRC Press, Boca Raton, FL.

Sherman, F.S. 1990. *Viscous Flow*, McGraw-Hill, New York.

Panton, R.L. 1984. *Incompressible Flow,* John Wiley & Sons, New York.

3.3 Similitude: Dimensional Analysis and Data Correlation

Stuart W. Churchill

Dimensional Analysis

Similitude refers to the formulation of a description for physical behavior that is general and independent of the individual dimensions, physical properties, forces, etc. In this subsection the treatment of similitude is restricted to *dimensional analysis*; for a more general treatment see Zlokarnik (1991). The full power and utility of dimensional analysis is often underestimated and underutilized by engineers. This technique may be applied to a complete mathematical model or to a simple listing of the variables that define the behavior. Only the latter application is described here. For a description of the application of dimensional analysis to a mathematical model see Hellums and Churchill (1964).

General Principles

Dimensional analysis is based on the principle that all additive or equated terms of a complete relationship between the variables must have the same net dimensions. The analysis starts with the preparation of a list of the individual dimensional variables (dependent, independent, and parametric) that are presumed to define the behavior of interest. The performance of dimensional analysis in this context is reasonably simple and straightforward; the principal difficulty and uncertainty arise from the identification of the variables to be included or excluded. If one or more important variables are inadvertently omitted, the reduced description achieved by dimensional analysis will be incomplete and inadequate as a guide for the correlation of a full range of experimental data or theoretical values. The familiar band of plotted values in many graphical correlations is more often a consequence of the omission of one or more variables than of inaccurate measurements. If, on the other hand, one or more irrelevant or unimportant variables are included in the listing, the consequently reduced description achieved by dimensional analysis will result in one or more unessential dimensionless groupings. Such excessive dimensionless groupings are generally less troublesome than missing ones because the redundancy will ordinarily be revealed by the process of correlation. Excessive groups may, however, suggest unnecessary experimental work or computations, or result in misleading correlations. For example, real experimental scatter may inadvertently and incorrectly be correlated in all or in part with the variance of the excessive grouping.

In consideration of the inherent uncertainty in selecting the appropriate variables for dimensional analysis, it is recommended that this process be interpreted as a *speculative* and subject to correction of the basis of experimental data or other information. Speculation may also be utilized as a formal technique to identify the effect of eliminating a variable or of combining two or more. The latter aspect of speculation, which may be applied either to the original listing of dimensional variables or to the resulting set of dimensionless groups, is often of great utility in identifying possible limiting behavior or dimensionless groups of marginal significance. The systematic speculative elimination of all but the most certain variables, one at a time, followed by regrouping, is recommended as a general practice. The additional effort as compared with the original dimensional analysis is minimal, but the possible return is very high. A general discussion of this process may be found in Churchill (1981).

The minimum number of independent dimensionless groups i required to describe the fundamental and parametric behavior is (Buckingham, 1914)

$$i = n - m \tag{3.3.1}$$

where n is the number of variables and m is the number of fundamental dimensions such as mass M, length L, time θ, and temperature T that are introduced by the variables. The inclusion of redundant dimensions such as force F and energy E that may be expressed in terms of mass, length, time, and

temperature is at the expense of added complexity and is to be avoided. (Of course, mass could be replaced by force or temperature by energy as alternative fundamental dimensions.) In some rare cases i is actually greater than $n - m$. Then

$$i = n - k \tag{3.3.2}$$

where k is the maximum number of the chosen variables that cannot be combined to form a dimensionless group. Determination of the minimum number of dimensionless groups is helpful if the groups are to be chosen by inspection, but is unessential if the algebraic procedure described below is utilized to determine the groups themselves since the number is then obvious from the final result.

The *particular* minimal set of dimensionless groups is arbitrary in the sense that two or more of the groups may be multiplied together to any positive, negative, or fractional power as long as the number of independent groups is unchanged. For example, if the result of a dimensional analysis is

$$\phi\left\{XY^{1/2}, Z/Y^2, Z\right\} = 0 \tag{3.3.3}$$

where X, Y, and Z are independent dimensionless groups, an equally valid expression is

$$\phi\left\{X, Y, Z\right\} = 0 \tag{3.3.4}$$

Dimensional analysis itself does not provide any insight as to the best choice of equivalent dimensionless groupings, such as between those of Equations (3.3.3) and (3.3.4). However, isolation of each of the variables that are presumed to be the most important in a separate group may be convenient in terms of interpretation and correlation. Another possible criterion in choosing between alternative groupings may be the relative invariance of a particular one. The functional relationship provided by Equation (3.3.3) may equally well be expressed as

$$X = \phi\left\{Y, Z\right\} \tag{3.3.5}$$

where X is implied to be the dependent grouping and Y and Z to be independent or parametric groupings.

Three primary methods of determining a minimal set of dimensionless variables are (1) by inspection; (2) by combination of the residual variables, one at a time, with the set of chosen variables that cannot be combined to obtain a dimensionless group; and (3) by an algebraic procedure. These methods are illustrated in the examples that follow.

Example 3.3.1: Fully Developed Flow of Water Through a Smooth Round Pipe

Choice of Variables. The shear stress τ_w on the wall of the pipe may be postulated to be a function of the density ρ and the dynamic viscosity μ of the water, the inside diameter D of the pipe, and the space-mean of the time-mean velocity u_m. The limitation to fully developed flow is equivalent to a postulate of independence from distance x in the direction of flow, and the specification of a smooth pipe is equivalent to the postulate of independence from the roughness e of the wall. The choice of τ_w rather than the pressure drop per unit length $-dP/dx$ avoids the need to include the acceleration due to gravity g and the elevation z as variables. The choice of u_m rather than the volumetric rate of flow V, the mass rate of flow w, or the mass rate of flow per unit area G is arbitrary but has some important consequences as noted below. The postulated dependence may be expressed functionally as $\phi\{\tau_w, \rho, \mu, D, u_m\} = 0$ or $\tau_w = \phi\{\rho, \mu, D, u_m\}$.

Tabulation. Next prepare a tabular listing of the variables and their dimensions:

	τ_w	ρ	μ	D	u_m
M	1	1	1	0	0
L	–1	–3	–1	1	1
θ	–2	0	–1	0	–1
T	0	0	0	0	0

Minimal Number of Groups. The number of postulated variables is 5. Since the temperature does not occur as a dimension for any of the variables, the number of fundamental dimensions is 3. From Equation (3.3.1), the minimal number of dimensionless groups is 5 – 3 = 2. From inspection of the above tabulation, a dimensionless group cannot be formed from as many as three variables such as D, μ, and ρ. Hence, Equation (3.3.2) also indicates that $i = 5 - 3 = 2$.

Method of Inspection. By inspection of the tabulation or by trial and error it is evident that only two independent dimensionless groups may be formed. One such set is

$$\phi\left\{\frac{\tau_w}{\rho u_m^2}, \frac{D u_m \rho}{\mu}\right\} = 0$$

Method of Combination. The residual variables τ_w and μ may be combined in turn with the noncombining variables ρ, D, and u_m to obtain two groups such as those above.

Algebraic Method. The algebraic method makes formal use of the postulate that the functional relationship between the variables may in general be represented by a power series. In this example such a power series may be expressed as

$$\tau_w = \sum_{i=1}^{N} A_i \rho^{a_i} \mu^{b_i} D^{c_i} u_m^{d_i}$$

where the coefficients A_i are dimensionless. Each additive term on the right-hand side of this expression must have the same net dimensions as τ_w. Hence, for the purposes of dimensional analysis, only the first term need be considered and the indexes may be dropped. The resulting highly restricted expression is $\tau_w = A\rho^a \mu^b D^c u_m^d$. Substituting the dimensions for the variables gives

$$\frac{M}{L\theta^2} = A\left(\frac{M}{L^3}\right)^a \left(\frac{M}{L\theta}\right)^b L^c \left(\frac{L}{\theta}\right)^d$$

Equating the sum of the exponents of M, L, and θ on the right-hand side of the above expression with those of the left-hand side produces the following three simultaneous linear algebraic equations: $1 = a + b$; $-1 = -3a - b + c + d$; and $-2 = -b - d$, which may be solved for a, c, and d in terms of b to obtain $a = 1 - b$, $c = -b$, and $d = 2 - b$. Substitution then gives $\tau_w = A\rho^{1-b}\mu^b D^{-b} u_m^{2-b}$ which may be regrouped as

$$\frac{\tau_w}{\rho u_m^2} = A\left(\frac{\mu}{D u_m \rho}\right)^b$$

Since this expression is only the first term of a power series, it should *not* be interpreted to imply that $\tau_w/\rho u_m^2$ is necessarily proportional to some power at $\mu/Du_m\rho$ but instead only the equivalent of the expression derived by the method of inspection. The inference of a power dependence between the

dimensionless groups is the most common and serious error in the use of the algebraic method of dimensional analysis.

Speculative Reductions. Eliminating ρ as a variable on speculative grounds to

$$\phi\left\{\frac{\tau_w D}{\mu u_m}\right\} = 0$$

or its exact equivalent:

$$\frac{\tau_w D}{\mu u_m} = A$$

The latter expression with $A = 8$ is actually the exact solution for the laminar regime ($Du_m\rho/\mu < 1800$). A relationship that does not include ρ may alternatively be derived directly from the solution by the method of inspection as follows. First, ρ is eliminated from one group, say $\tau_w/\rho u_m^2$, by multiplying it with $Du_m\rho/\mu$ to obtain

$$\phi\left\{\frac{\tau_w D}{\mu u_m}, \frac{Du_m\rho}{\mu}\right\} = 0$$

The remaining group containing ρ is now simply dropped. Had the original expression been composed of three independent groups each containing ρ, that variable would have to be eliminated from two of them before dropping the third one.

The relationships that are obtained by the speculative elimination of μ, D, and u_m, one at a time, do not appear to have any range of physical validity. Furthermore, if w or G had been chosen as the independent variable rather than u_m, the limited relationship for the laminar regime would not have been obtained by the elimination of ρ.

Alternative Forms. The solution may also be expressed in an infinity of other forms such as

$$\phi\left\{\frac{\tau_w D^2\rho}{\mu^2}, \frac{Du_m\rho}{\mu}\right\} = 0$$

If τ_w is considered to be the principal dependent variable and u_m the principal independent variable, this latter form is preferable in that these two quantities do not then appear in the same grouping. On the other hand, if D is considered to be the principal independent variable, the original formulation is preferable. The variance of $\tau_w/\rho u_m^2$ is less than that of $\tau_w D/\mu u_m$ and $\tau_w D^2\rho/\mu^2$ in the turbulent regime while that of $\tau_w D/\mu u_m$ is zero in the laminar regime. Such considerations may be important in devising convenient graphical correlations.

Alternative Notations. The several solutions above are more commonly expressed as

$$\phi\left\{\frac{f}{2}, \mathrm{Re}\right\} = 0$$

$$\phi\left\{\frac{f\,\mathrm{Re}}{2}, \mathrm{Re}\right\} = 0$$

or

$$\phi\left\{\frac{f\mathrm{Re}^2}{2},\mathrm{Re}\right\}=0$$

where $f = 2\,\tau_w/\rho u_m^2$ is the *Fanning friction factor* and $Re = Du_m\rho/\mu$ is the *Reynolds number.*

The more detailed forms, however, are to be preferred for purposes of interpretation or correlation because of the explicit appearance of the individual, physically measurable variables.

Addition of a Variable. The above results may readily be extended to incorporate the roughness e of the pipe as a variable. If two variables have the same dimensions, they will always appear as a dimensionless group in the form of a ratio, in this case e appears most simply as e/D. Thus, the solution becomes

$$\phi\left\{\frac{\tau_w}{\rho u_m^2},\frac{Du_m\rho}{\mu},\frac{e}{D}\right\}=0$$

Surprisingly, as contrasted with the solution for a smooth pipe, the speculative elimination of μ and hence of the group $Du_m\rho/\mu$ now results in a valid asymptote for $Du_m\rho/\mu \to \infty$ and all finite values of e/D, namely,

$$\phi\left\{\frac{\tau_w}{\rho u_m^2},\frac{e}{D}\right\}=0$$

Example 3.3.2: Fully Developed Forced Convection in Fully Developed Flow in a Round Tube

It may be postulated for this process that $h = \phi\{D, u_m, \rho, \mu, k, c_p\}$, where here h is the local heat transfer coefficient, and c_p and k are the specific heat capacity and thermal conductivity, respectively, of the fluid. The corresponding tabulation is

	h	D	u_m	ρ	μ	k	c_p
M	1	0	0	1	1	1	0
L	0	1	1	–3	–1	1	2
θ	–3	0	–1	0	–1	–3	–2
T	–1	0	0	0	0	–1	–1

The number of variables is 7 and the number of independent dimensions is 4, as is the number of variables such as D, u_m, ρ, and k that cannot be combined to obtain a dimensionless group. Hence, the minimal number of dimensionless groups is 7 – 4 = 3. The following acceptable set of dimensionless groups may be derived by any of the procedures illustrated in Example 1:

$$\frac{hD}{k}=\phi\left\{\frac{Du_m\rho}{\mu},\frac{c_p\mu}{k}\right\}$$

Speculative elimination of μ results in

$$\frac{hD}{k}=\phi\left\{\frac{Du_m\rho c_p}{k}\right\}$$

which has often erroneously been inferred to be a valid asymptote for $c_p\mu/k \to 0$. Speculative elimination of D, u_m, ρ, k, and c_p individually also does not appear to result in expressions with any physical validity. However, eliminating c_p and ρ or u_m gives a valid result for the laminar regime, namely,

$$\frac{hD}{k} = A$$

The general solutions for flow and convection in a smooth pipe may be combined to obtain

$$\frac{hD}{k} = \phi\left\{\frac{\tau_w D^2 \rho}{\mu^2}, \frac{c_p \mu}{k}\right\}$$

which would have been obtained directly had u_m been replaced by τ_w in the original tabulation. This latter expression proves to be superior in terms of speculative reductions. Eliminating D results in

$$\frac{h\mu}{k(\tau_w \rho)^{1/2}} = \phi\left\{\frac{c_p \mu}{k}\right\}$$

which may be expressed in the more conventional form of

$$\mathrm{Nu} = \mathrm{Re}\left(\frac{f}{2}\right)^{1/2} \phi\{\mathrm{Pr}\}$$

where Nu = hD/k is the *Nusselt number* and Pr = $c_p\mu/k$ is the *Prandtl number*. This result appears to be a valid asymptote for Re $\to \infty$ and a good approximation for even moderate values (>5000) for large values of Pr. Elimination of μ as well as D results in

$$\frac{h}{c_p(\tau_w \rho)^{1/2}} = A$$

or

$$\mathrm{Nu} = A\,\mathrm{Re}\,\mathrm{Pr}\left(\frac{f}{2}\right)^{1/2}$$

which appears to be an approximate asymptote for Re $\to \infty$ and Pr $\to 0$. Elimination of both c_p and ρ again yields the appropriate result for laminar flow, indicating that ρ rather than u_m is the meaningful variable to eliminate in this respect.

The numerical value of the coefficient A in the several expressions above depends on the mode of heating, a true variable, but one from which the purely functional expressions are independent. If j_w the heat flux density at the wall, and $T_w - T_m$, the temperature difference between the wall and the bulk of the fluid, were introduced as variables in place of $h \equiv j_w/(T_w - T_m)$, another group such as $c_p(T_w - T_m)(D\rho/\mu)^2$ or $\rho c_p(T_w - T_m)/\tau_w$ or $c_p(T_w - T_m)/u_m^2$, which represents the effect of viscous dissipation, would be obtained. This effect is usually but not always negligible. (See Chapter 4.)

Example 3.3.3: Free Convection from a Vertical Isothermal Plate

The behavior for this process may be postulated to be represented by

$$h = \phi\{g, \beta, T_w - T_\infty, x, \mu, \rho, c_p, k\}$$

where g is the acceleration due to gravity, β is the volumetric coefficient of expansion with temperature, T_∞ is the unperturbed temperature of the fluid, and x is the vertical distance along the plate. The corresponding tabulation is

	h	g	β	$T_w - T_\infty$	x	μ	ρ	c_p	k
M	1	0	0	0	0	1	1	0	1
L	0	1	0	0	1	–1	–3	2	1
θ	–3	–2	0	0	0	–1	0	–2	–3
T	–1	0	–1	1	0	0	0	–1	1

The minimal number of dimensionless groups indicated by both methods is 9 – 4 = 5. A satisfactory set of dimensionless groups, as found by any of the methods illustrated in Example 1 is

$$\frac{hx}{k} = \phi\left\{\frac{\rho^2 g x^3}{\mu^2}, \frac{c_p\mu}{k}, \beta\left(T_w - T_\infty\right), c_p\left(T_w - T_\infty\right)\left(\frac{\rho x}{\mu}\right)^2\right\}$$

It may be reasoned that the buoyant force which generates the convective motion must be proportional to $\rho g\beta(T_w - T_\infty)$, thus, g in the first term on the right-hand side must be multiplied by $\beta(T_w - T_\infty)$, resulting in

$$\frac{hx}{k} = \phi\left\{\frac{\rho^2 g\beta\left(T_w - T_\infty\right)x^3}{\mu^2}, \frac{c_p\mu}{k}, \beta\left(T_w - T_\infty\right), c_p\left(T_w - T_\infty\right)\left(\frac{\rho x}{\mu}\right)^2\right\}$$

The effect of expansion other than on the buoyancy is now represented by $\beta(T_w - T_\infty)$, and the effect of viscous dissipation by $c_p(T_w - T_\infty)(\rho x/\mu)^2$. Both effects are negligible for all practical circumstances. Hence, this expression may be reduced to

$$\frac{hx}{k} = \phi\left\{\frac{\rho^2 g\beta\left(T_w - T_\infty\right)x^3}{\mu^2}, \frac{c_p\mu}{k}\right\}$$

or

$$\mathrm{Nu}_x = \phi\left\{\mathrm{Gr}_x, \mathrm{Pr}\right\}$$

where $\mathrm{Nu}_x = hx/k$ and $\mathrm{Gr}_x = \rho^2 g\beta(T_w - T_\infty)x^3/\mu^2$ is the *Grashof number.*

Elimination of x speculatively now results in

$$\frac{hx}{k} = \left(\frac{\rho^2 g\beta\left(T_w - T_\infty\right)x^3}{\mu^2}\right)^{1/3} \phi\left\{\mathrm{Pr}\right\}$$

or

$$\mathrm{Nu}_x = \mathrm{Gr}_x^{1/3}\phi\left\{\mathrm{Pr}\right\}$$

This expression appears to be a valid asymptote for $\mathrm{Gr}_x \to \infty$ and a good approximation for the entire turbulent regime. Eliminating μ speculatively rather than x results in

$$\frac{hx}{k} = \phi\left\{\frac{\rho^2 c_p^2 g\beta\left(T_w - T_\infty\right)x^3}{k^2}\right\}$$

or

$$\mathrm{Nu}_x = \phi\left\{\mathrm{Gr}_x \mathrm{Pr}^2\right\}$$

The latter expression appears to be a valid asymptote for $\mathrm{Pr} \to 0$ for all Gr_x, that is, for both the laminar and the turbulent regimes. The development of a valid asymptote for large values of Pr requires more subtle reasoning. First $c_p\mu/k$ is rewritten as $\mu/\rho\alpha$ where $\alpha = k/\rho c_p$. Then ρ is eliminated speculatively except as it occurs in $\rho g\beta(T_w - T_\infty)$ and $k/\rho c_p$. The result is

$$\frac{hx}{k} = \phi\left\{\frac{c_p \rho^2 g\beta\left(T_w - T_\infty\right)x^3}{\mu k}\right\}$$

or

$$\mathrm{Nu}_x = \phi\left\{\mathrm{Ra}_x\right\}$$

where

$$\mathrm{Ra}_x = \frac{c_p \rho^2 g\beta\left(T_w - T_\infty\right)x^3}{\mu k} = \mathrm{Gr}_x \mathrm{Pr}$$

is the *Rayleigh number*. The expression appears to be a valid asymptote for $\mathrm{Pr} \to \infty$ and a reasonable approximation for even moderate values of Pr for all Gr_x, that is, for both the laminar and the turbulent regimes.

Eliminating x speculatively from the above expressions for small and large values of Pr results in

$$\mathrm{Nu}_x = A\left(\mathrm{Gr}_x \mathrm{Pr}^2\right)^{1/3} = A\left(\mathrm{Ra}_x \mathrm{Pr}\right)^{1/3}$$

and

$$\mathrm{Nu}_x = B\left(\mathrm{Gr}_x \mathrm{Pr}\right)^{1/3} = B\left(\mathrm{Ra}_x\right)^{1/3}$$

The former appears to be a valid asymptote for $\mathrm{Pr} \to 0$ and $\mathrm{Gr}_x \to \infty$ and a reasonable approximation for very small values of Pr in the turbulent regime, while the latter is well confirmed as a valid asymptote for $\mathrm{Pr} \to \infty$ and $\mathrm{Gr}_x \to \infty$ and as a good approximation for moderate and large values of Pr over the entire turbulent regime. The expressions in terms of Gr_x are somewhat more complicated than those in terms of Ra_x, but are to be preferred since Gr_x is known to characterize the transition from laminar to turbulent motion in natural convection just as Re_D does in forced flow in a channel. The power of speculation combined with dimensional analysis is well demonstrated by this example in which valid asymptotes are thereby attained for several regimes.

Correlation of Experimental Data and Theoretical Values

Correlations of experimental data are generally developed in terms of dimensionless groups rather than in terms of the separate dimensional variables in the interests of compactness and in the hope of greater generality. For example, a complete set of graphical correlations for the heat transfer coefficient h of Example 3.3.2 above in terms of each of the six individual independent variables and physical properties might approach book length, whereas the dimensionless groupings both imply that a single plot with one parameter should be sufficient. Furthermore, the reduced expression for the turbulent regime implies that a plot of Nu/Re $f^{1/2}$ vs. Pr should demonstrate only a slight parametric dependence on Re or Re $f^{1/2}$. Of course, the availability of a separate correlation for f as a function of Re is implied.

Theoretical values, that is, ones obtained by numerical solution of a mathematical model in terms of either dimensional variables or dimensionless groups, are presumably free from imprecision. Even so, because of their discrete form, the construction of a correlation or correlations for such values may be essential for the same reasons as for experimental data.

Graphical correlations have the merit of revealing general trends, of providing a basis for evaluation of the choice of coordinates, and most of all of displaying visually the scatter of the individual experimental values about a curve representing a correlation or their behavior on the mean. (As mentioned in the previous subsection, the omission of a variable may give the false impression of experimental error in such a plot.) On the other hand, correlating equations are far more convenient as an input to a computer than is a graphical correlation. These two formats thus have distinct and complementary roles; both should generally be utilized. The merits and demerits of various graphical forms of correlations are discussed in detail by Churchill (1979), while the use of logarithmic and arithmetic coordinates, the effects of the appearance of a variable in both coordinates, and the effects of the distribution of error between the dependent and independent variable are further illustrated by Wilkie (1985).

Churchill and Usagi (1972; 1974) proposed general usage of the following expression for the formulation of correlating equations:

$$y^n\{x\} = y_0^n\{x\} + y_\infty^n\{x\} \tag{3.3.6}$$

where $y_o\{x\}$ and $y_\infty\{x\}$ denote asymptotes for small and large values of x, respectively, and n is an arbitrary exponent. For convenience and simplicity, Equation (3.3.6) may be rearranged in either of the following two forms:

$$\left(Y(x)\right)^n = 1 + Z^n\{x\} \tag{3.3.7}$$

or

$$\left(\frac{Y\{x\}}{Z\{x\}}\right)^n = 1 + \frac{1}{Z^n\{x\}} \tag{3.3.8}$$

where $Y\{x\} \equiv y\{x\}/y_o\{x\}$ and $Z\{x\} \equiv y_\infty\{x\}/y_o\{x\}$. Equations (3.3.6), (3.3.7), and (3.3.9) are hereafter denoted collectively as the CUE (Churchill–Usagi equation). The principle merits of the CUE as a canonical expression for correlation are its simple form, generality, and minimal degree of explicit empiricism, namely, only that of the exponent n, since the asymptotes $y_o\{x\}$ and $y_\infty\{x\}$ are ordinarily known in advance from theoretical considerations or well-established correlations. Furthermore, as will be shown, the CUE is quite insensitive to the numerical value of n. Although the CUE is itself very simple in form, it is remarkably successful in representing closely very complex behavior, even including the dependence on secondary variables and parameters, by virtue of the introduction of such dependencies through $y_o\{x\}$ and $y_\infty\{x\}$. In the rare instances in which such dependencies are not represented in the asymptotes, n may be correlated as a function of the secondary variables and/or parameters. Although

the CUE usually produces very close representations, it is empirical and not exact. In a few instances, numerical values of n have been derived or rationalized on theoretical grounds, but even then some degree of approximation is involved. Furthermore, the construction of a correlating expression in terms of the CUE is subject to the following severe limitations:

1. The asymptotes $y_o\{x\}$ and $y_\infty\{x\}$ must intersect once and only once;
2. The asymptotes $y_o\{x\}$ and $y_\infty\{x\}$ must be free of singularities. Even though a singularity occurs beyond the asserted range of the asymptote, it will persist and disrupt the prediction of the CUE, which is intended to encompass all values of the independent variable x; and
3. The asymptotes must both be upper or lower bounds.

In order to avoid or counter these limitations it may be necessary to modify or replace the asymptotes with others. Examples of this process are provided below. A different choice for the dependent variable may be an option in this respect. The suitable asymptotes for use in Equation (3.3.6) may not exist in the literature and therefore may need to be devised or constructed. See, for example, Churchill (1988b) for guidance in this respect. Integrals and derivatives of the CUE are generally awkward and inaccurate, and may include singularities not present or troublesome in the CUE itself. It is almost always preferable to develop a separate correlating equation for such quantities using derivatives or integrals of $y_o\{x\}$ and $y_\infty\{x\}$, simplified or modified as appropriate.

The Evaluation of *n*

Equation (3.3.6) may be rearranged as

$$n = \frac{\ln\left\{1 + \left(\dfrac{y_\infty\{x\}}{y_0\{x\}}\right)^n\right\}}{\ln\left\{\dfrac{y\{x\}}{y_0\{x\}}\right\}} \tag{3.3.9}$$

and solved for n by iteration for any known value of $y\{x\}$, presuming that $y_o\{x\}$ and $y_\infty\{x\}$ are known. If $y\{x^*\}$ is known, where x^* represents the value of x at the point of intersection of the asymptotes, that is, for $y_o\{x\} = y_\infty\{x\}$, Equation (3.3.9) reduces to

$$n = \frac{\ln\{2\}}{\ln\left\{\dfrac{y\{x^*\}}{y_0\{x^*\}}\right\}} \tag{3.3.10}$$

and iterative determination of n is unnecessary.

A graphical and visual method of evaluation of n is illustrated in Figure 3.3.1 in which $Y\{Z\}$ is plotted vs. Z for $0 \le Z \le 1$ and $Y\{Z\}/Z$ vs. $1/Z$ for $0 \le 1/Z \le 1$ in arithmetic coordinates with n as a parameter. Values of $y\{x\}$ may be plotted in this form and the best overall value of n selected visually (as illustrated in Figure 3.3.2). A logarithmic plot of $Y\{Z\}$ vs. Z would have less sensitivity relative to the dependence on n. (See, for example, Figure 1 of Churchill and Usagi, 1972.) Figure 3.3.1 explains in part the success of the CUE. Although y and x may both vary from 0 to ∞, the composite variables plotted in Figure 3.3.1 are highly constrained in that the compound independent variables Z and $1/Z$ vary only between 0 and 1, while for $n \ge 1$, the compound dependent variables $Y\{Z\}$ and $Y\{Z\}/Z$ vary only from 1 to 2.

Because of the relative insensitivity of the CUE to the numerical value of n, an integer or a ratio of two small integers may be chosen in the interest of simplicity and without significant loss of accuracy. For example, the maximum variance in Y (for $0 \le Z \le 1$) occurs at $Z = 1$ and increases only $100(2^{1/20} - 1) = 3.5\%$ if n is decreased from 5 to 4. If $y_o\{x\}$ and $y_\infty\{x\}$ are both lower bounds, n will be positive,

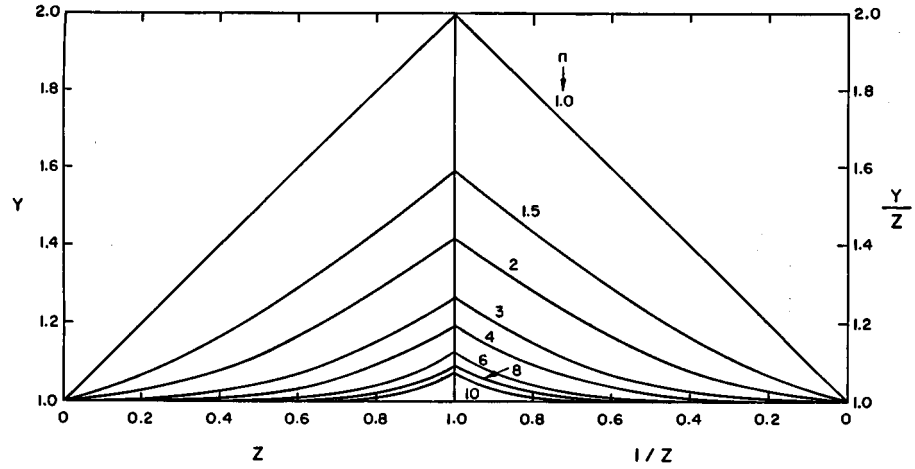


FIGURE 3.3.1 Arithmetic, split-coordinate plot of Equation 3.3.10. (From Churchill, S.W. and Usagi, R. *AIChE J.* 18(6), 1123, 1972. With permission from the American Institute of Chemical Engineers.)

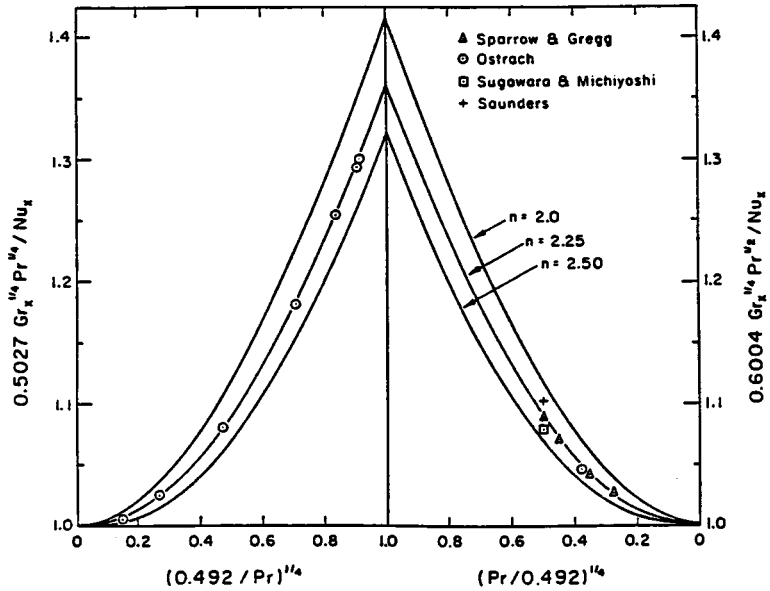


FIGURE 3.3.2 Arithmetic, split-coordinate plot of computed values and experimental data for laminar free convection from an isothermal vertical plate. (From Churchill, S.W. and Usagi, R. *AIChE J.* 18(6), 1124, 1972. With permission from the American Institute of Chemical Engineers.)

and if they are both upper bounds, n will be negative. To avoid extending Figure 3.3.1 for negative values of n, $1/y\{x\}$ may simply be interpreted as the dependent variable.

Intermediate Regimes

Equations (3.3.6), (3.3.7), and (3.3.8) imply a slow, smooth transition between $y_o\{x\}$ and $y_\infty\{x\}$ and, moreover, one that is symmetrical with respect to $x^*(Z = 1)$. Many physical systems demonstrate instead a relatively abrupt transition, as for example from laminar to turbulent flow in a channel or along a flat plate. The CUE may be applied serially as follows to represent such behavior if an expression $y_i\{x\}$ is

postulated for the intermediate regime. First, the transition from the initial to the intermediate regime is represented by

$$y_1^n = y_0^n + y_i^n \tag{3.3.11}$$

Then the transition from this combined regime to the final regime by

$$y^m = y_1^m + y_\infty^m = \left(y_0^n + y_i^n\right)^{m/n} + y_\infty^m \tag{3.3.12}$$

Here, and throughout the balance of this subsection, in the interests of simplicity and clarity, the functional dependence of all the terms on x is implied rather written out explicitly. If y_o is a lower bound and y_i is implied to be one, y_1 and y_∞ must be upper bounds. Hence, n will then be positive and m negative. If y_o and y_i are upper bounds, y_1 and y_∞ must be lower bounds; then n will be negative and m positive. The reverse formulation starting with y_∞ and y_1 leads by the same procedure to

$$y^n = y_0^n + \left(y_i^m + y_\infty^m\right)^{n/m} \tag{3.3.13}$$

If the intersections of y_i with y_o and y_∞ are widely separated with respect to x, essentially the same pair of values for n and m will be determined for Equations (3.3.12) and (3.3.13), and the two representations for y will not differ significantly. On the other hand, if these intersections are close in terms of x, the pair of values of m and n may differ significantly and one representation may be quite superior to the other. In some instances a singularity in y_o or y_∞ may be tolerable in either Equation (3.3.12) or (3.3.13) because it is overwhelmed by the other terms. Equations (3.3.12) and (3.3.13) have one hidden flaw. For $x \to 0$, Equation (3.3.12) reduces to

$$y \to y_0\left[1 + \left(\frac{y_\infty}{y_0}\right)^m\right]^{1/m} \tag{3.3.14}$$

If y_o is a lower bound, m is necessarily negative, and values of y less than y_o are predicted. If y_o/y_∞ is sufficiently small or if m is sufficiently large in magnitude, this discrepancy may be tolerable. If not, the following alternative expression may be formulated, again starting from Equation (3.3.11):

$$\left(y^n - y_0^n\right)^m = y_i^{nm} + \left(y_\infty^n - y_0^n\right)^m \tag{3.3.15}$$

Equation (3.3.15) is free from the flaw identified by means of Equation (3.3.14) and invokes no additional empiricism, but a singularity may occur at $y_\infty = y_o$, depending on the juxtapositions of y_o, y_i, and y_∞. Similar anomalies occur for Equation (3.3.13) and the corresponding analog of Equation (3.3.14), as well as for behavior for which $n < 0$ and $m > 0$. The preferable form among these four is best chosen by trying each of them.

One other problem with the application of the CUE for a separate transitional regime is the formulation of an expression for $y_i\{x\}$, which is ordinarily not known from theoretical considerations. Illustrations of the empirical determination of such expressions for particular cases may be found in Churchill and Usagi (1974), Churchill and Churchill (1975), and Churchill (1976; 1977), as well as in Example 3.3.5 below.

Example 3.3.4: The Pressure Gradient in Flow through a Packed Bed of Spheres

The pressure gradient at asymptotically low rates of flow (the creeping regime) can be represented by the Kozeny–Carman equation, $\Phi = 150\ \mathrm{Re}_p$, and at asymptotically high rates of flow (the inertial regime)

by the Burke–Plummer equation, $\Phi = 1.75\ (\mathrm{Re}_p)^2$, where $\Phi = \rho\varepsilon^2 d_p(-dP_f/dx)\mu^2(1-\varepsilon)$, $\mathrm{Re}_p = d_p u_o \rho/\mu(1-\varepsilon)$, d_p = diameter of spherical particles, m, ε = void fraction of bed of spheres, dP_f/dx = dynamic pressure gradient (due to friction), Pa/m, and u_o = superficial velocity (in absence of the spheres), m/sec. For the origin of these two asymptotic expressions see Churchill (1988a). They both have a theoretical structure, but the numerical coefficients of 150 and 1.75 are basically empirical. These equations are both lower bounds and have one intersection. Experimental data are plotted in Figure 3.3.3, which has the form of Figure 3.3.1 with $Y = \Phi/150\ \mathrm{Re}_p$, $Y/Z = \Phi/(1.75\ \mathrm{Re}_p)^2$ and $Z = 1.75\ \mathrm{Re}_p^2/150\ \mathrm{Re}_p = \mathrm{Re}_p/85.7$. A value of $n = 1$ is seen to represent these data reasonably well on the mean, resulting in

$$\Phi = 150\mathrm{Re}_p + 1.75\left(\mathrm{Re}_p\right)^2$$

which was originally proposed as a correlating equation by Ergun (1952) on the conjecture that the volumetric fraction of the bed in "turbulent" flow is proportional to Re_p. The success of this expression in conventional coordinates is shown in Figure 3.3.4. The scatter, which is quite evident in the arithmetic split coordinates of Figure 3.3.3, is strongly suppressed in a visual sense in the logarithmic coordinates of Figure 3.3.4.

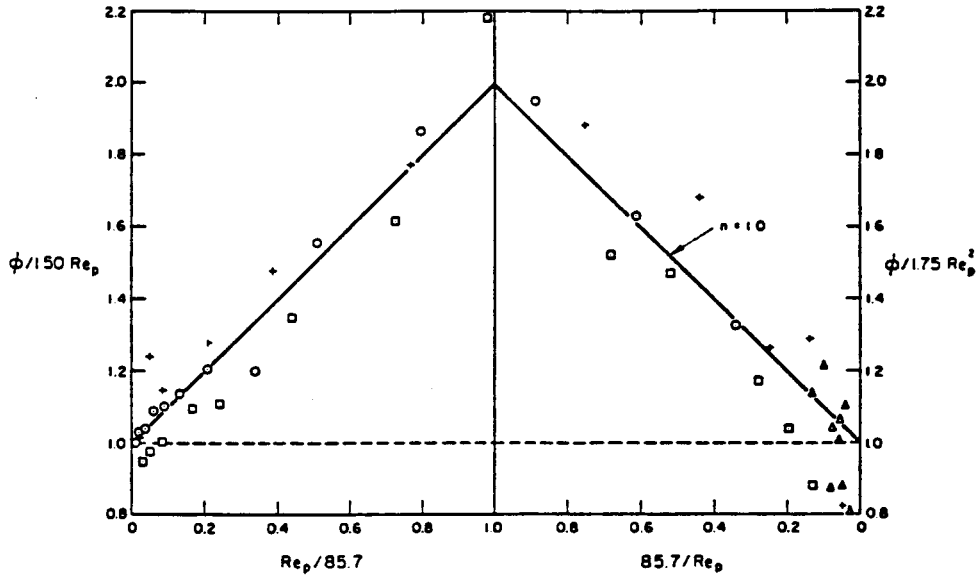


FIGURE 3.3.3 Arithmetic, split-coordinate plot of experimental data for the pressure drop in flow through a packed bed of spheres. (From Churchill, S.W. and Usagi, R. *AIChE J.* 18(6), 1123, 1972. With permission from the American Institute of Chemical Engineers.)

Example 3.3.5: The Friction Factor for Commercial Pipes for All Conditions

The serial application of the CUE is illustrated here by the construction of a correlating equation for both smooth and rough pipes in the turbulent regime followed by combination of that expression with ones for the laminar and transitional regimes.

The Turbulent Regime. The Fanning friction factor, f_F, for turbulent flow in a smooth round pipe for asymptotically large rates of flow (say $\mathrm{Re}_D > 5000$) may be represented closely by the empirical expression:

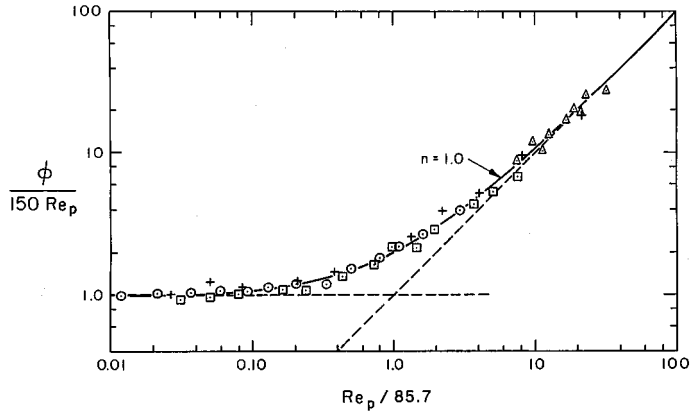


FIGURE 3.3.4 Logarithmic correlation of experimental data for the pressure drop in flow through a packed bed of spheres. (From Churchill, S.W. and Usagi, R. *AIChE J.* 18(6), 1123, 1972. With permission from the American Institute of Chemical Engineers.)

$$\left(\frac{2}{f_F}\right)^{1/2} = 0.256 + 2.5\ln\left\{\left(\frac{f_F}{2}\right)^{1/2} \mathrm{Re}_D\right\}$$

A corresponding empirical representation for naturally rough pipe is

$$\left(\frac{2}{f_F}\right)^{1/2} = 3.26 + 2.5\ln\left\{\frac{D}{e}\right\}$$

Direct combination of these two expressions in the form of the CUE does not produce a satisfactory correlating equation, but their combination in the following rearranged forms:

$$e^{(1/2.5)(2/f_F)^{1/2}} = 1.108\left(\frac{f_F}{2}\right)^{1/2} \mathrm{Re}_D$$

and

$$e^{(1/2.5)(2/f_F)^{1/2}} = 3.68\left(\frac{D}{e}\right)$$

with $n = -1$ results in, after the reverse rearrangement,

$$\left(\frac{2}{f_F}\right)^{1/2} = 0.256 + 2.5\ln\left\{\frac{\left(\frac{f_F}{2}\right)^{1/2} \mathrm{Re}_D}{1 + 0.3012\left(\frac{e}{D}\right)\left(\frac{f_F}{2}\right)^{1/2} \mathrm{Re}_D}\right\}$$

The exact equivalent of this expression in structure but with the slightly modified numerical coefficients of 0.300, 2.46, and 0.304 was postulated by Colebrook (1938–1939) to represent his own experimental data. The coefficients of the expression given here are presumed to be more accurate, but the difference in the predictions of f_F with the two sets of coefficients is within the band of uncertainty of the

experimental data. The turbulent regime of the "friction-factor" plot in most current textbooks and handbooks is simply a graphical representation of the Colebrook equation. Experimental values are not included in such plots since e, the effective roughness of commercial pipes, is simply a correlating factor that forces agreement with the Colebrook equation. Values of e for various types of pipe in various services are usually provided in an accompanying table, that thereby constitutes an integral part of the correlation.

The Laminar Region. The Fanning friction factor in the laminar regime of a round pipe ($\mathrm{Re}_d < 1800$) is represented exactly by the following theoretical expression known as Poiseuille's law: $f_F = 16/\mathrm{Re}_D$. This equation may be rearranged as follows for convenience in combination with that for turbulent flow:

$$\left(\frac{2}{f_F}\right)^{1/2} = \frac{\mathrm{Re}_D\left(f_F/2\right)^{1/2}}{8}$$

The Transitional Regime. Experimental data as well as semitheoretical computed values for the limiting behavior in the transition may be represented closely by $(f_F/2) = (\mathrm{Re}_D/37500)^2$. This expression may be rewritten, in terms of $(2/f_F)^{1/2}$ and $\mathrm{Re}_D(f_F/2)^{1/2}$, as follows:

$$\left(\frac{f_F}{2}\right)^{1/2} = \left(\frac{37500}{\mathrm{Re}_D\left(f_F/2\right)^{1/2}}\right)^{1/2}$$

Overall Correlation. The following correlating equation for all $\mathrm{Re}_D(f_F/2)^{1/2}$ and e/D may now be constructed by the combination of the expressions for the turbulent and transition regimes in the form of the CUE with $n = 8$, and then that expression and that for the laminar regime with $n = -12$, both components being chosen on the basis of experimental data and predicted values for the full regime of transition:

$$\left(\frac{2}{f_F}\right)^{1/2} = \left[\left(\frac{8}{\mathrm{Re}_D\left(f_F/2\right)^{1/2}}\right)^{12} + \left(\left[\frac{37500}{\mathrm{Re}_D\left(f_F/2\right)^{1/2}}\right]^4 + \left|2.5\ln\left\{\frac{1.108\,\mathrm{Re}_D\left(f_F/2\right)^{1/2}}{1+0.3012\left(\frac{e}{a}\right)\mathrm{Re}_D\left(f_F/2\right)^{1/2}}\right\}\right|^8\right)^{-3/2}\right]^{-1/12}$$

The absolute value signs are only included for aesthetic reasons; the negative values of the logarithmic term for very small values of $\mathrm{Re}_D(f_F/2)^{1/2}$ do not affect the numerical value of $(2/f_F)^{1/2}$ in the regime in which they occur. This overall expression appears to have a complicated structure, but it may readily be recognized to reduce to its component parts when the corresponding term is large with respect to the other two. It is insensitive to the numerical values of the two arbitrary exponents. For example, doubling their values would have almost no effect on the predictions of $(f_F/2)^{1/2}$. The principal uncertainty is associated with the expression for the transition regime, but the overall effect of the corresponding term is very small. The uncertainties associated with this correlating equation are common to most graphical correlations and algebraic expressions for the friction factor, and are presumed to be fairly limited in magnitude and to be associated primarily with the postulated value of e. Although the overall expression is explicit in $\mathrm{Re}_D(f_F/2)^{1/2}$ rather than Re_D, the latter quantity may readily be obtained simply by multiplying the postulated value of $\mathrm{Re}_D(f_F/2)^{1/2}$ by the computed values of $(2/f_F)^{1/2}$.

References

Buckingham, E. 1914. On physically similar systems; illustrations of the use of dimensional equations. *Phys. Rev., Ser. 2,* 4(4):345–375.

Churchill, S.W. 1976. A comprehensive correlating equation for forced convection from plates. *AIChE J.* 22(2):264–268.

Churchill, S.W. 1977. Comprehensive correlating equation for heat, mass and momentum transfer in fully developed flow in smooth tubes. *Ind. Eng. Chem. Fundam.* 16(1):109–116.

Churchill, S.W. 1979. *The Interpretation and Use of Rate Data. The Rate Process Concept,* rev. printing, Hemisphere Publishing Corp., Washington, D.C.

Churchill, S.W. 1981. The use of speculation and analysis in the development of correlations. *Chem. Eng. Commun.* 9:19–38.

Churchill, S.W. 1988a. Flow through porous media, Chapter 19 in *Laminar Flows. The Practical Use of Theory,* pp. 501–538, Butterworths, Boston.

Churchill, S.W. 1988b. Derivation, selection, evaluation and use of asymptotes. *Chem. Eng. Technol.* 11:63–72.

Churchill, S.W. and Churchill, R.U. 1975. A general model for the effective viscosity of pseudoplastic and dilatant fluids. *Rheol. Acta.* 14:404–409.

Churchill, S.W. and Usagi, R. 1972. A general expression for the correlation of rates of transfer and other phenomena. *AIChE J.* 18(6):1121–1128.

Churchill, S.W. and Usagi, R. 1974. A standardized procedure for the production of correlations in the form of a common empirical equation. *Ind. Eng. Chem. Fundam.* 13(1):39–44.

Colebrook, C.R. 1938–1939. Turbulent flow in pipes with particular reference to the transition region between the smooth and rough pipe laws. *J. Inst. Civ. Eng.* 11(5024):133–156.

Ergun, S. 1952. Fluid flow through packed beds. *Chem. Eng. Prog.* 48(2):81–96.

Hellums, J.D. and Churchill, S.W. 1964. Simplifications of the mathematical description of boundary and initial value problems. *AIChE J.* 10(1):110–114.

Wilkie, D. 1985. The correlation of engineering data reconsidered. *Int. J. Heat Fluid Flow.* 8(2):99–103.

Zlokarnik, M. 1991. *Dimensional Analysis and Scale-Up in Chemical Engineering*. Springer-Verlag, Berlin.

3.4 Hydraulics of Pipe Systems

J. Paul Tullis

Basic Computations

Equations

Solving fluid flow problems involves the application of one or more of the three basic equations: continuity, momentum, and energy. These three basic tools are developed from the law of conservation of mass, Newton's second law of motion, and the first law of thermodynamics.

The simplest form of the continuity equation is for one-dimensional incompressible steady flow in a conduit. Applying continuity between any two sections gives

$$A_1V_1 = A_2V_2 = Q \tag{3.4.1}$$

For a variable density the equation can be written

$$\rho_1 A_1 V_1 = \rho_2 A_2 V_2 = \dot{m} \tag{3.4.2}$$

in which A is the cross-sectional area of the pipe, V is the mean velocity at that same location, Q is the flow rate, ρ is the fluid density, and $\dot{m}$ is the mass flow rate. The equations are valid for any rigid conduit as long as there is no addition or loss of liquid between the sections.

For steady state pipe flow, the momentum equation relates the net force in a given direction (F_x) acting on a control volume (a section of the fluid inside the pipe), to the net momentum flux through the control volume.

$$F_x = \rho_2 A_2 V_2 V_{2x} - \rho_1 A_1 V_1 V_{1x} \tag{3.4.3}$$

For incompressible flow this equation can be reduced to

$$F_x = \rho Q\left(V_{2x} - V_{1x}\right) \tag{3.4.4.}$$

These equations can easily be applied to a three-dimensional flow problem by adding equations in the y and z directions.

A general form of the energy equation (see Chapter 2) applicable to incompressible pipe or duct flow

$$\frac{P_1}{\gamma} + Z_1 + \frac{V_1^2}{2g} = \frac{P_2}{\gamma} + Z_2 + \frac{V_2^2}{2g} - H_p + H_t + H_f \tag{3.4.5}$$

The units are energy per unit weight of liquid: ft · lb/lb or N · m/N which reduce to ft or m. The first three terms are pressure head (P/γ), elevation head (Z) (above some datum), and velocity head ($V^2/2g$). The last three terms on the right side of the equation are the total dynamic head added by a pump (H_p) or removed by a turbine (H_t) and the friction plus minor head losses (H_f). The sum of the first three terms in Equation 3.4.5 is defined as the total head, and the sum of the pressure and elevation heads is referred to as the piezometric head.

The purpose of this section is to determine the pressure changes resulting from incompressible flow in pipe systems. Since pipes of circular cross sections are most common in engineering application, the analysis in this section will be performed for circular geometry. However, the results can be generalized for a pipe of noncircular geometry by substituting for the diameter D in any of the equations, the hydraulic diameter, D_h, defined as

$$D_h = 4 \times \frac{\text{the cross sectional area}}{\text{the wetted perimeter}}$$

The analysis in this section can also be applied to gases and vapors, provided the Mach number in the duct does not exceed 0.3. For greater values of the Mach number, the compressibility effect becomes significant and the reader is referred to Section 3.7 on compressible flow.

Fluid Friction

The calculation of friction loss in pipes and ducts depends on whether the flow is laminar or turbulent. The Reynolds number is the ratio of inertia forces to viscous forces and is a convenient parameter for predicting if a flow condition will be laminar or turbulent. It is defined as

$$\mathrm{Re}_D = \frac{\rho VD}{\mu} = \frac{VD}{\nu} \tag{3.4.6}$$

in which V is the mean flow velocity, D diameter, ρ fluid density, μ dynamic viscosity, and ν kinematic viscosity.

Friction loss (H_f) depends on pipe diameter (d), length (L), roughness (e), fluid density (ρ) or specific weight (γ), viscosity (ν), and flow velocity (V). Dimensional analysis can be used to provide a functional relationship between the friction loss H_f, pipe dimensions, fluid properties, and flow parameters. The resulting equation is called the Darcy–Weisbach equation:

$$H_f = \frac{fLV^2}{2gd} = \frac{fLQ^2}{1.23gD^5} \tag{3.4.7}$$

The friction factor f is a measure of pipe roughness. It has been evaluated experimentally for numerous pipes. The data wee used to create the Moody friction factor chart shown as Figure 3.4.1. For Re < 2000, the flow in a pipe will be laminar and f is only a function of Re_D. It can be calculated by

$$f = \frac{64}{\mathrm{Re}_D} \tag{3.4.8}$$

At Reynolds numbers between about 2000 and 4000 the flow is unstable as a result of the onset of turbulence (critical zone in Figure 3.4.1). In this range, friction loss calculations are difficult because it is impossible to determine a unique value of f. For Re > 4000 the flow becomes turbulent and f is a function of both Re and relative pipe roughness (e/d). At high Re, f eventually depends only on e/d; defining the region referred to as fully turbulent flow. This is the region in Figure 3.4.1 where the lines for different e/d become horizontal (e is the equivalent roughness height and d pipe diameter). The Re_D at which this occurs depends on the pipe roughness. Laminar flow in pipes is unusual. For example, for water flowing in a 0.3-m-diameter pipe, the velocity would have to be below 0.02 m/sec for laminar flow to exist. Therefore, most practical pipe flow problems are in the turbulent region.

Using the Moody chart in Figure 3.4.1 to get f requires that Re and e/d be known. Calculating Re is direct if the water temperature, velocity, and pipe diameter are known. The problem is obtaining a good value for e. Typical values of e are listed in Figure 3.4.1. These values should be considered as guides only and not used if more–exact values can be obtained from the pipe supplier.

Since roughness may vary with time due to buildup of solid deposits or organic growths, f is also time dependent. Manufacturing tolerances also cause variations in the pipe diameter and surface roughness. Because of these factors, the friction factor for any pipe can only be approximated. A designer is required to use good engineering judgment in selecting a design value for f so that proper allowance is made for these uncertainties.

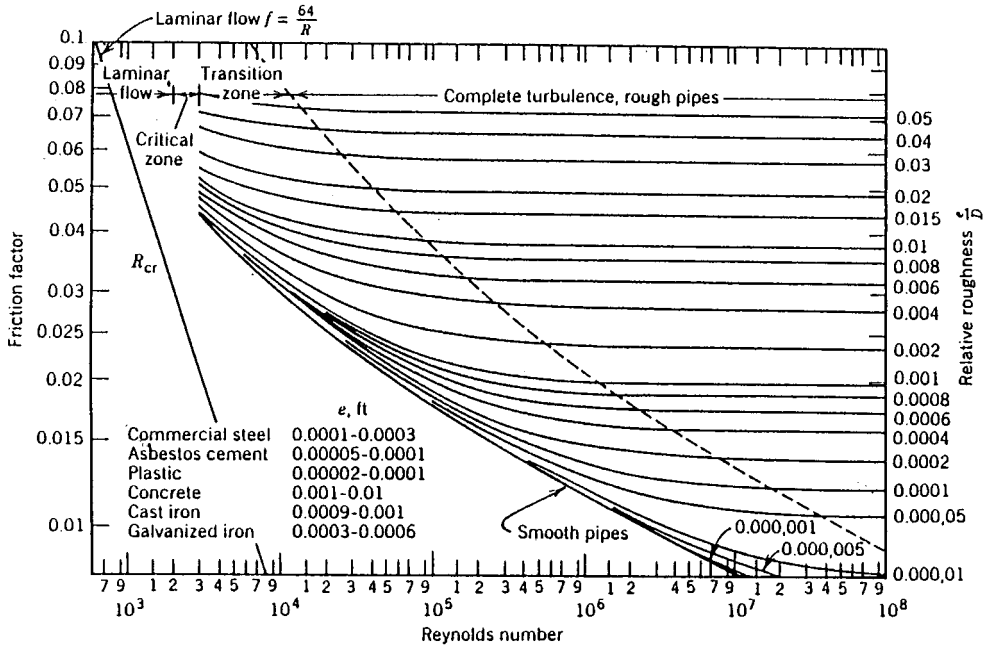


FIGURE 3.4.1 The Moody diagram.

For noncircular pipes, the only change in the friction loss equation is the use of an equivalent diameter — based on the hydraulic radius (R), i.e., $d = 4R$ — in place of the circular pipe diameter d. R is the ratio of the flow area to the wetter perimeter.

Wood (1966) developed equations which can be used in place of the Moody diagram to estimate f for $\mathrm{Re} > 10^4$ and $10^{-5} < k < 0.04$ ($k = e/d$).

$$f = a + b\mathrm{Re}^{-c} \tag{3.4.9}$$

$$a = 0.094k^{0.225} + 0.53k, \quad b = 88k^{0.44}, \quad c = 1.62k^{0.134}$$

The practical problem is still obtaining a reliable value for e. It cannot be directly measured but must be determined from friction loss tests of the pipe.

An exact solution using the Darcy–Weisbach equation can require a trial-and-error solution because of the dependency of f on Re if either the flow or pipe diameter are not known. A typical approach to solving this problem is to estimate a reasonable fluid velocity to calculate Re and obtain f from the Moody chart or Equation (3.4.9). Next, calculate a new velocity and repeat until the solution converges. Converging on a solution is greatly simplified with programmable calculators and a variety of software available for computer.

For long gravity flow pipelines, the starting point in selecting the pipe diameter is to determine the smallest pipe that can pass the required flow without friction loss exceeding the available head. For pumped systems, the selection must be based on an economic analysis that compares the pipe cost with the cost of building and operating the pumping plant.

Local Losses

Flow through valves, orifices, elbows, transitions, etc. causes flow separation which results in the generation and dissipation of turbulent eddies. For short systems containing many bends, valves, tees,

etc. local or minor losses can exceed friction losses. The head loss h_l associated with the dissipation caused by a minor loss is proportional to the velocity head and can be accounted for as a minor or local loss using the following equation.

$$h_l = K_l \frac{Q^2}{\left(2gA_m^2\right)} \tag{3.4.10}$$

in which K_l is the minor loss coefficient and A_m is the area of the pipe at the inlet to the local loss. The loss coefficient K_l is analogous to fL/d in Equation 3.4.7.

The summation of all friction and local losses in a pipe system can be expressed as

$$H_f = h_f + h_l \tag{3.4.11}$$

$$H_f = \left[\sum\left(\frac{fL}{2gdA_p^2}\right) + \sum\left(\frac{K_l}{2gA_m^2}\right)\right]Q^2 = CQ^2 \tag{3.4.12}$$

in which

$$C = \sum\left(\frac{fL}{2gdA_p^2}\right) + \sum\left(\frac{K_l}{2gA_m^2}\right) \tag{3.4.13}$$

It is important to use the correct pipe diameter for each pipe section and local loss.

In the past some have expressed the local losses as an equivalent pipe length: $L/d = K_l/f$. It simply represents the length of pipe that produces the same head loss as the local or minor loss. This is a simple, but not a completely accurate method of including local losses. The problem with this approach is that since the friction coefficient varies from pipe to pipe, the equivalent length will not have a unique value. When local losses are truly minor, this problem becomes academic because the error only influences losses which make up a small percentage of the total. For cases where accurate evaluation of all losses is important, it is recommended that the minor loss coefficients K_l be used rather than an equivalent length.

The challenging part of making minor loss calculations is obtaining reliable values of K_l. The final results cannot be any more accurate than the input data. If the pipe is long, the friction losses may be large compared with the minor losses and approximate values of K_l will be sufficient. However, for short systems with many pipe fittings, the local losses can represent a significant portion of the total system losses, and they should be accurately determined. Numerous factors influence K_l. For example, for elbows, K_l is influenced by the shape of the conduit (rectangular vs. circular), by the radius of the bend, the bend angle, the Reynolds number, and the length of the outlet pipe. For dividing or combining tees or Y-branches, the percent division of flow and the change in pipe diameter must also be included when estimating K_l. One factor which is important for systems where local losses are significant is the interaction between components placed close together. Depending on the type, orientation, and spacing of the components, the total loss coefficient may be greater or less than the simple sum of the individual K_l values.

Comparing the magnitude of $\Sigma(fL/2gA^2)$ to $\Sigma(K_l/2gA_m^2)$ will determine how much care should be given to the selection of the K_l values. Typical values of K_l are listed in Table 3.4.1 (Tullis, 1989). When more comprehensive information on loss coefficients is needed, the reader is referred to Miller (1990).

TABLE 3.4.1 Minor Loss Coefficients

Item	K_l Typical Value	Typical Range
Pipe inlets		
Inward projecting pipe	0.78	0.5–0.9
Sharp corner-flush	0.50	—
Slightly rounded	0.20	0.04–0.5
Bell mouth	0.04	0.03–0.1
Expansions[a]	$(1 - A_1/A_2)^2$ (based on V_1)	
Contractions[b]	$(1/C_c - 1)^2$ (based on V_2)	

A_2/A_1	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
C_c	0.624	0.632	0.643	0.659	0.681	0.712	0.755	0.813	0.892

Bends[c]		
Short radius, $r/d = 1$		
90	—	0.24
45	—	0.1
30	—	0.06
Long radius, $r/d = 1.5$		
90	—	0.19
45	—	0.09
30	—	0.06
Mitered (one miter)		
90	1.1	—
60	0.50	0.40–0.59
45	0.3	0.35–0.44
30	0.15	0.11–0.19
Tees	[c]	—
Diffusers	[c]	—
Valves		
Check valve	0.8	0.5–1.5
Swing check	1.0	0.29–2.2
Tilt disk	1.2	0.27–2.62
Lift	4.6	0.85–9.1
Double door	1.32	1.0–1.8
Full-open gate	0.15	0.1–0.3
Full-open butterfly	0.2	0.2–0.6
Full-open globe	4.0	3–10

[a] See Streeter and Wylie, 1975, p. 304.
[b] See Streeter and Wylie, 1975, p. 305.
[c] See Miller, 1990.
[d] See Kalsi Engineering and Tullis Engineering Consultants, 1993.

Pipe Design

Pipe Materials

Materials commonly used for pressure pipe transporting liquids are ductile iron, concrete, steel, fiberglass, PVC, and polyolefin. Specifications have been developed by national committees for each of these pipe materials. The specifications discuss external loads, internal design pressure, available sizes, quality of materials, installation practices, and information regarding linings. Standards are available from the following organizations:

American Water Works Association (AWWA)
American Society for Testing and Materials (ASTM)
American National Standards Institute (ANSI)
Canadian Standards Association (CSA)

Federal Specifications (FED)
Plastic Pipe Institute (PPI)

In addition, manuals and other standards have been published by various manufacturers and manufacturer's associations. All of these specifications and standards should be used to guide the selection of pipe material. ASCE (1992) contains a description of each of these pipe materials and a list of the specifications for the various organizations which apply to each material. It also discusses the various pipe-lining materials available for corrosion protection.

For air- and low-pressure liquid applications one can use unreinforced concrete, corrugated steel, smooth sheet metal, spiral rib (sheet metal), and HDPE (high-density polyethylene) pipe. The choice of a material for a given application depends on pipe size, pressure requirements, resistance to collapse from internal vacuums, external loads, resistance to internal and external corrosion, ease of handling and installing, useful life, and economics.

Pressure Class Guidelines

Procedures for selecting the pressure class of pipe vary with the type of pipe material. Guidelines for different types of materials are available from AWWA, ASTM, ANSI, CSA, FED, PPI and from the pipe manufacturers. These specifications should be obtained and studied for the pipe materials being considered.

The primary factors governing the selection of a pipe pressure class are (1) the maximum steady state operating pressure, (2) surge and transient pressures, (3) external earth loads and live loads, (4) variation of pipe properties with temperature or long-time loading effects, and (5) damage that could result from handling, shipping, and installing or reduction in strength due to chemical attack or other aging factors. The influence of the first three items can be quantified, but the last two are very subjective and are generally accounted for with a safety factor which is the ratio of the burst pressure to the rated pressure.

There is no standard procedure on how large the safety factor should be or on how the safety factor should be applied. Some may feel that it is large enough to account for all of the uncertainties. Past failures of pipelines designed using this assumption prove that it is not always a reliable approach. The procedure recommended by the author is to select a pipe pressure class based on the internal design pressure (IDP) defined as

$$\text{IDP} = \left(P_{\max} + P_s\right)\text{SF} \tag{3.4.14}$$

in which $P_{\max}$ is the maximum steady state operating pressure, P_s is the surge or water hammer pressure, and SF is the safety factor applied to take care of the unknowns (items 3 to 5) just enumerated. A safety factor between 3 and 4 is typical.

The maximum steady state operating pressure ($P_{\max}$) in a gravity flow system is usually the difference between the maximum reservoir elevation and the lowest elevation of the pipe. For a pumped system it is usually the pump shutoff head calculated based on the lowest elevation of the pipe.

Surge and transient pressures depend on the specific pipe system design and operation. Accurately determining P_s requires analyzing the system using modern computer techniques. The most commonly used method is the "Method of Characteristics" (Tullis, 1989; Wylie and Streeter, 1993). Some of the design standards give general guidelines to predict P_s that can be used if a detailed transient analysis is not made. However, transients are complex enough that simple "rules of thumb" are seldom accurate enough. Transients are discussed again in a later subsection.

Selection of wall thickness for larger pipes is often more dependent on collapse pressure and handling loads than it is on burst pressure. A thin-wall, large-diameter pipe may be adequate for resisting relatively high internal pressures but may collapse under negative internal pressure or, if the pipe is buried, the soil and groundwater pressure plus live loads may be sufficient to cause collapse even if the pressure inside the pipe is positive.

External Loads

There are situations where the external load is the controlling factor determining if the pipe will collapse. The magnitude of the external load depends on the diameter of the pipe, the pipe material, the ovality (out of roundness) of the pipe cross section, the trench width, the depth of cover, the specific weight of the soil, the degree of soil saturation, the type of backfill material, the method used to backfill, the degree of compaction, and live loads. The earth load increases with width and depth of the trench, and the live load reduces with depth of cover. The cumulative effect of all these sources of external loading requires considerable study and analysis.

There are no simple guidelines for evaluating external pipe loads. Because of the complexity of this analysis, the default is to assume that the safety factor is adequate to account for external loads as well as the other factors already mentioned. One should not allow the safety factor to replace engineering judgment and calculations. One option to partially compensate for the lack of a detailed analysis is to use a higher-pressure class of pipe in areas where there will be large live loads or where the earth loading is unusually high. One should consider the cost of a pipe failure caused by external loads compared with the cost of using a thicker pipe or the cost of performing a detailed analysis. Those interested in the details of performing calculations of earth loading should be Spranger and Handy, 1973.

Limiting Velocities

There are concerns about upper and lower velocity limits. If the velocity is too low, problems may develop due to settling of suspended solids and air being trapped at high points and along the crown of the pipe. The safe lower velocity limit to avoid collecting air and sediment depends on the amount and type of sediment and on the pipe diameter and pipe profile. Velocities greater than about 1 m/sec (3 ft/sec) are usually sufficient to move trapped air to air release valves and keep the sediment in suspension.

Problems associated with high velocities are (1) erosion of the pipe wall or liner (especially if coarse suspended sediment is present), (2) cavitation at control valves and other restrictions, (3) increased pumping costs, (4) removal of air at air release valves, (5) increased operator size and concern about valve shaft failures due to excessive flow torques, and (6) an increased risk of hydraulic transients. Each of these should be considered before making the final pipe diameter selection. A typical upper velocity for many applications if 6 m/sec (20 ft/sec). However, with proper pipe design and analysis (of the preceding six conditions), plus proper valve selection, much higher velocities can be tolerated. A typical upper velocity limit for standard pipes and valves is about 6 m/sec (20 ft/sec). However, with proper design and analysis, much higher velocities can be tolerated.

Valve Selection

Valves serve a variety of functions. Some function as isolation or block valves that are either full open or closed. Control valves are used to regulate flow or pressure and must operate over a wide range of valve openings. Check valves prevent reverse flow, and air valves release air during initial filling and air that is collected during operation and admit air when the pipe is drained.

Control Valves

For many flow control applications it is desirable to select a valve that has linear control characteristics. This means that if you close the valve 10%, the flow reduces about 10%. Unfortunately, this is seldom possible since the ability of a valve to control flow depends as much on the system as it does on the design of the valve. The same valve that operates linearly in one system may not in another.

Selecting the proper flow control valve should consider the following criteria:

1. The valve should not produce excessive pressure drop when full open.
2. The valve should control over at least 50% of its movement.
3. At maximum flow, the operating torque must not exceed the capacity of the operator or valve shaft and connections.
4. The valve should not be subjected to excessive cavitation.

5. Pressure transients should not exceed the safe limits of the system.
6. Some valves should not be operated at very small openings. Other valves should be operated near full open.

Controllability. To demonstrate the relationship between a valve and system, consider a butterfly valve that will be used to control the flow between two reservoirs with an elevation difference of ΔZ. System A is a short pipe (0.3 m dia., 100 m long, ΔZ = 10 m) where pipe friction is small $fL/2gdA_p^2 = 46.9$, and System B is a long pipe (0.3 m dia., 10,000 m long, ΔZ = 200 m) with high friction $fL/2gdA_p^2 =$ 4690. Initially, assume that the same butterfly valve will be used in both pipes and it will be the same size as the pipe. The flow can be calculated using the energy equation (Equation 3.4.5) and the system loss equation (Equation (3.4.12):

$$Q = \sqrt{\frac{\Delta Z}{\left[\sum\left(\frac{fL}{2gdA_p^2}\right) + \sum\left(\frac{K_l}{2gA_m^2}\right)\right]}} \tag{3.4.15}$$

For the valve, assume that the K_l full open is 0.2 and at 50% open it is 9.0. Correspondingly, $K_l/2gA_m^2$ = 1.905 and 85.7. For System A, the flow with the valve full open will be 0.453 m^3/sec and at 50% open 0.275 m^3/sec, a reduction of 39%. Repeating these calculations over the full range of valve openings would show that the flow for System A reduces almost linearly as the valve closes.

For System B, the flow with the valve full open will be 0.206 m^3/sec and at 50% open 0.205 m^3/sec, a reduction of less than 1%. For System B the valve does not control until the valve loss, expressed by $K_l/2gA_m^2$ becomes a significant part of the friction term (4690). The same valve in System B will not start to control the flow until it has closed more than 50%. A line-size butterfly valve is obviously not a good choice for a control valve in System B. One solution to this problem is to use a smaller valve. If the butterfly valve installed in System B was half the pipe diameter, it would control the flow over more of the stroke of the valve.

The range of opening over which the valve controls the flow also has a significant effect on the safe closure time for control valves. Transient pressures are created when there is a sudden change in the flow. Most valve operators close the valve at a constant speed. If the valve does not control until it is more than 50% closed, over half of the closing time is wasted and the effective valve closure time is less than half the total closing time. This will increase the magnitude of the transients that will be generated.

Torque. To be sure that the valve shaft, connections, and operator are properly sized, the maximum torque or thrust must be known. If the maximum force exceeds operator capacity, it will not be able to open and close the valve under extreme flow conditions. If the shaft and connectors are underdesigned, the valve may fail and slam shut causing a severe transient.

For quarter-turn valves, the force required to operate a valve consists of seating, bearing, and packing friction, hydrodynamic (flow) forces, and inertial forces. These forces are best determined experimentally. A key step in applying experimental torque information is the determination of the flow condition creating maximum torque. This requires that the system be analyzed for all possible operating conditions and valve openings. For a given size and type of valve, the flow torque depends on the torque coefficient (which is dependent on the specific valve design) and the pressure drop which, in turn, depends on the flow. In short systems where there is little friction loss and high velocities, a quarter-turn valve will see maximum torques at large openings where the flow is high. In long systems with high reservoir heads and smaller velocities, the same valve will see maximum torque at small openings where the pressure drop is high.

One situation where it is easy to overlook the condition causing maximum torque is with parallel pumps. Each pump normally will have a discharge control valve. The maximum system flow occurs with all three pumps operating. However, the flow and the torque for any of the pump discharge valves

is maximum for only one pump operating. One specific example (Tullis, 1989) showed that the torque on a butterfly valve was three times higher when one pump was operating compared with three pumps operating in parallel.

Cavitation. Cavitation is frequently an important consideration in selection and operation of control valves. It is necessary to determine if cavitation will exist, evaluate its intensity, and estimate its effect on the system and environment. Cavitation can cause noise, vibration, and erosion damage and can decrease performance. The analysis should consider the full range of operation. Some valves cavitate worst at small openings and others cavitate heavily near full open. It depends on both the system and the valve design. If cavitation is ignored in the design and selection of the valves, repairs and replacement of the valves may be necessary. Information for making a complete cavitation analysis is beyond the scope of this section. Detailed information on the process to design for cavitation is contained in Tullis (1989; 1993).

The first step in a cavitation analysis is selecting the acceptable level of cavitation. Experimental data are available for four limits: incipient (light, intermittent noise), critical (light, steady noise), incipient damage (pitting damage begins), and choking (very heavy damage and performance drops off). Limited cavitation data are available for each of these limits (Tullis, 1989; 1993). Choosing a cavitation limit depends on several factors related to the operating requirements, expected life, location of the device, details of the design, and economics. For long-term operation of a control valve in a system where noise can be tolerated, the valve should never operate beyond incipient damage. In systems where noise is objectionable, critical cavitation would be a better operating limit.

Using a choking cavitation as a design limit is often misused. It is generally appropriate as a design limit for valves that only operate for short periods of time, such as a pressure relief valve. The intensity of cavitation and the corresponding noise vibration and erosion damage at the valve are at their maximum just before a valve chokes. If the valve operates beyond choking (sometimes referred to as supercavitation), the collapse of the vapor cavities occurs remote from the valve. Little damage is likely to occur at the valve, but farther downstream serious vibration and material erosion problems can occur.

If the cavitation analysis indicates that the valve, orifice, or other device will be operating at a cavitation level greater than can be tolerated, various techniques can be used to limit the level of cavitation. One is to select a different type of valve. Recent developments in valve design have produced a new generation of valves that are more resistant to cavitation. Most of them operate on the principle of dropping the pressure in stages. They usually have multiple paths with numerous sharp turns or orifices in series. Two limitations to these valves are that they often have high pressure drops (even when full open), and they are only usable with clean fluids.

A similar approach is to place multiple conventional valves in series or a valve in series with orifice plates. Proper spacing of valves and orifices placed in series is important. The spacing between valves depends upon the type. For butterfly valves it is necessary to have between five and eight diameters of pipe between valves to prevent flutter of the leaf of the downstream valve and to obtain the normal pressure drop at each valve. For globe, cone, and other types of valves, it is possible to bolt them flange to flange and have satisfactory operation.

For some applications another way to suppress cavitation is to use a free-discharge valve that is vented so cavitation cannot occur. There are valves specifically designed for this application. Some conventional valves can also be used for free discharge, if they can be adequately vented.

Cavitation damage can be suppressed by plating critical areas of the pipe and valve with cavitation-resistant materials. Based on tests using a magnetostriction device, data show that there is a wide variation in the resistance of the various types of material. Limited testing has been done on the erosion resistance of different materials and coating to cavitation in flowing systems. The available data show that there is less variation in the damage resistance of materials in actual flowing systems. However, experience has shown the plating parts of the valve with the right material will extend valve life.

Injecting air to suppress cavitation is a technique which has been used for many years with varying degrees of success. The most common mistake is placing the air injection port in the wrong location so

the air does not get to the cavitation zone. If an adequate amount of air is injected into the proper region, the noise, vibrations, and erosion damage can be significantly reduced. The air provides a cushioning effect reducing the noise, vibration, and erosion damage. If the system can tolerate some air being injected, aeration is usually the cheapest and best remedy for cavitation.

Transients. Transient pressures can occur during filling and flushing air from the line, while operating valves, and when starting or stopping pumps. If adequate design provisions and operational procedures are not established, the transient pressure can easily exceed the safe operating pressure of the pipe. A system should be analyzed to determine the type and magnitudes of possible hydraulic transients. The basic cause is rapid changes in velocity. The larger the incremental velocity change and the faster that change takes place, the greater will be the transient pressure. If the piping system is not designed to withstand the high transient pressures, or if controls are not included to limit the pressure, rupture of the pipe or damage to equipment can result.

All pipelines experience transients. Whether or not the transient creates operational problems or pipe failure depends upon its magnitude and the ability of the pipes and mechanical equipment to tolerate high pressures without damage. For example, an unreinforced concrete pipeline may have a transient pressure head allowance of only a meter above its operating pressure before damage can occur. For such situations even slow closing of control valves or minor interruptions of flow due to any cause may create sufficient transient pressures to rupture the pipeline. In contrast, steel and plastic pipes can take relatively high transient pressures without failure.

Transients caused by slow velocity changes, such as the rise and fall of the water level in a surge tank, are called surges. Surge analysis, or "rigid column theory" involves mathematical or numerical solution of simple ordinary differential equations. The compressibility of the fluid and the elasticity of the conduit are ignored, and the entire column of fluid is assumed to move as a rigid body.

When changes in velocity occur rapidly, both the compressibility of the liquid and the elasticity of the pipe must be included in the analysis. This procedure is often called "elastic" or "waterhammer" analysis and involves tracking acoustic pressure waves through the pipe. The solution requires solving partial differential equations.

An equation predicting the head rise ΔH caused by a sudden change of velocity $\Delta V = V_2 - V_1$ can be derived by applying the unsteady momentum equation to a control volume of a section of the pipe where the change of flow is occurring. Consider a partial valve closure which instantly reduces the velocity by an amount ΔV. Reduction of the velocity can only be accomplished by an increase in the pressure upstream from the valve. This creates a pressure wave of magnitude ΔH which travels up the pipe at the acoustic velocity **a**. The increased pressure compresses the liquid and expands the pipe. The transient head rise due to an incremental change in velocity is

$$\Delta H = -\mathbf{a}\,\Delta V/g, \quad \text{for} \quad \mathbf{a} \gg \Delta V \tag{3.4.16}$$

This equation is easy to use for multiple incremental changes of velocity as long as the first wave has not been reflected back to the point of origin.

The derivation of Equation (3.4.16) is based on an assumption of an instant velocity change. For a valve closing at the end of the pipe, instant closure actually refers to a finite time. It is the longest time that a valve can be closed and still cause a pressure rise equal to that of an instant closure. Normally, it is equal to $2L/a$ sec (which is the time required for the first pressure wave to travel to and from the other end of the pipe of length L); the head rise at the valve will be the same as if the valve were closed instantly. The $2L/a$ time is therefore often the instant closure time.

For a valve at the end of a long pipeline, the instant closure time can be considerably greater than $2L/a$. This is because when the friction loss coefficient fL/d is much greater than the loss coefficient for the valve K_l, the valve can be closed a long way before the flow changes. This dead time must be added to the $2L/a$ time to identify the actual instant closure time. To avoid the maximum potential transient pressure rise, the valve must be closed much slower than the instant closure time.

Computational techniques for estimating transient pressures caused by unsteady flow in pipelines are too complex to be done with simple hand calculations. The solution involves solving partial differential equations based on the equations of motion and continuity. These equations are normally solved by the method of characteristics. This technique transforms the equations into total differential equations. After integration, the equations can be solved numerically by finite differences. This analysis provides equations that can be used to predict the flow and head at any interior pipe section at any time (Tullis, 1989; Wiley and Streeter, 1993).

To complete the analysis, equations describing the boundary conditions are required. Typical boundary conditions are the connection of a pipe to a reservoir, a valve, changes in pipe diameter or material, pipe junctions, etc. Friction loss is included in the development of the basic equations and minor losses are handled as boundary conditions. The analysis properly models friction and the propagation and reflections of the pressure wave. It can also be used for surge calculations.

It is recommended that every pipe system should have at least a cursory transient analysis performed to identify the possibility of serious transients and decide whether or not a detailed analysis is necessary. If an analysis indicates that transients are a problem, the types of solutions available to the engineer include

1. Increasing the closing time of control valves.
2. Using a smaller valve to provide better control.
3. Designing special facilities for filling, flushing, and removing air from pipelines.
4. Increasing the pressure class of the pipeline.
5. Limiting the flow velocity.
6. Using pressure relief valves, surge tanks, air chambers, etc.

Check Valves

Selecting the wrong type or size of check valve can result in poor performance, severe transients, and frequent repairs (Kalsi, 1993). Proper check valve selection requires understanding the characteristics of the various types of check valves and analyzing how they will function as a part of the system in which they will be installed. A check valve that operates satisfactorily in one system may be totally inadequate in another. Each valve type has unique characteristics that give it advantages or disadvantages compared with the others. The characteristics of check valves that describe their hydraulic performance and which should be considered in the selection process include

1. Opening characteristics, i.e., velocity vs. disk position data.
2. Velocity required to fully open and firmly backseat the disk.
3. Pressure drop at maximum flow.
4. Stability of the disk at partial openings.
5. Sensitivity of disk flutter to upstream disturbances.
6. Speed of valve closure compared with the rate of flow reversal of the system.

Disk stability varies with flow rate, disk position, and upstream disturbances and is an important factor in determining the useful life of a check valve. For most applications it is preferable to size the check valve so that the disk is fully open and firmly backseated at normal flow rates. One of the worst design errors is to oversize a check valve that is located just downstream from a disturbance such as a pump, elbow, or control valve. If the disk does not fully open, it will be subjected to severe motion that will accelerate wear. To avoid this problem, it may be necessary to select a check valve that is smaller than the pipe size.

The transient pressure rise generated at check valve closure is another important consideration. The pressure rise is a function of how fast the valve disk closes compared with how fast the flow in the system reverses. The speed that the flow in a system reverses depends on the system. In systems where rapid flow reversals occur, the disk can slam shut causing a pressure transient (Thorley, 1989).

The closing speed of a valve is determined by the mass of the disk, the forces closing the disk, and the distance of travel from full open to closed. Fast closing valves have the following properties: the disk (including all moving parts) is lightweight, closure is assisted by springs, and the full stroke of the disk is short. Swing check valves are the slowest-closing valves because they violate all three of these criteria; i.e., they have heavy disks, no springs, and long disk travel. The nozzle check valve is one of the fastest-closing valves because the closing element is light, is spring loaded, and has a short stroke. The silent, duo, double door, and lift check valves with springs are similar to nozzle valves in their closing times, mainly because of the closing force of the spring.

Systems where rapid flow reversals occur include parallel pumps, where one pump is stopped while the others are still operating, and systems that have air chambers or surge tanks close to the check valve. For these systems there is a high-energy source downstream from the check valve to cause the flow to quickly reverse. As the disk nears its seat, it starts to restrict the reverse flow. This builds up the pressure, accelerates the disk, and slams it into the seat. Results of laboratory experiments, field tests, and computer simulations show that dramatic reductions in the transient pressures at disk closure can be achieved by replacing a slow-closing swing check valve with a fast-acting check valve. For example, in a system containing parallel pumps where the transient was generated by stopping one of the pumps, the peak transient pressure was reduced from 745 to 76 kPa when a swing check was replaced with a nozzle check valve. Such a change improved performance and significantly reduced maintenance.

Air Valves

There are three types of automatic air valves: (1) air/vacuum valves, (2) air release valves, and (3) combination valves. The air/vacuum valve is designed for releasing air while the pipe is being filled and for admitting air when the pipe is being drained. The valve must be large enough that it can admit and expel large quantities of air at a low pressure differential. The outlet orifice is generally the same diameter as the inlet pipe.

These valves typically contain a float, which rises and closes the orifice as the valve body fills with water. Once the line is pressurized, this type of valve cannot reopen to remove air that may subsequently accumulate until the pressure becomes negative, allowing the float to drop. If the pressure becomes negative during a transient or while draining, the float drops and admits air into the line. For thin-walled pipes that can collapse under internal vacuums, the air/vacuum valves should be sized for a full pipe break at the lowest pipe elevation. The vacuum valve must supply an air flow equal to the maximum drainage rate of the water from the pipe break and at an internal pipe pressure above the pipe collapse pressure.

The critical factor in sizing air/vacuum valves is usually the air flow rate to protect the pipe from a full pipe break. Since a pipe is filled much slower than it would drain during a full break, the selected valve will be sized so that the air is expelled during filling without pressurizing the pipe. Sizing charts are provided by manufacturers.

Air release valves contain a small orifice and are designed to release small quantities of pressurized air that are trapped during filling and that accumulate after initial filling and pressurization. The small orifice is controlled by a plunger activated by a float at the end of a lever arm. As air accumulates in the valve body, the float drops and opens the orifice. As the air is expelled, the float rises and closes off the orifice. Sizing air release valves requires an estimate of the amount of pressurized air that must be expelled. This is determined by the filling procedure and any source of air that can be admitted into the pipe or be degassed from the liquid during operation.

The combination valve is actually two valves: a large valve that functions as an air/vacuum valve and a small one that functions as an air release valve. The installation can either consist of an air/vacuum valve and an air release valve plumbed in parallel, or the two can be housed in a single valve body. Most air valve installations require combination valves.

One caution is that manual air release valves should be avoided because improper operation of them can be very dangerous. If the system is pressurized with the manual air valves closed, the trapped air

will be pressurized to full system pressure. When the air valve is manually opened, the pressurized air can cause rapid acceleration of the liquid and generate serious transients when the water is decelerated as it hits the air valve. If manual air valves are installed, they should be very small so the air release rate is controlled to a safe rate.

Locating air valves in a piping system depends on the pipe profile, pipe length, and flow rates. Preferably, pipes should be laid to grade with valves placed at the high points or at intervals if there are no high points. One should use engineering judgment when defining a high point. If the pipe has numerous high points that are close together, or if the high points are not pronounced, it will not be necessary to have an air valve at each high point. If the liquid flow velocity is above about 1 m/sec (3 ft/sec), the flowing water can move the entrained air past intermediate high points to a downstream air valve. Releasing the air through an air valve prevents any sizable air pockets under high pressure from forming in the pipe. Trapped air under high pressure is extremely dangerous.

Velocity of the flow during filling is important. A safe way to fill a pipe is to limit the initial fill rate to an average flow velocity of about 0.3 m/sec (1 ft/sec) until most of the air is released and the air/vacuum valves close. The next step is to flush the system at about 1 m/sec (3 ft/sec), at a low system pressure, to flush the remaining air to an air release valve. It is important that the system not be pressurized until the air has been removed. Allowing large quantities of air under high pressure to accumulate and move through the pipe can generate severe transients. This is especially true if the compressed air is allowed to pass through a control valve or manual air release valve. When pressurized air flows through a partially open valve, the sudden acceleration and deceleration of the air and liquid can generate high pressure transients.

Pump Selection

Optimizing the life of a water supply system requires proper selection, operation, and maintenance of the pumps. During the selection process, the designer must be concerned about matching the pump performance to the system requirements and must anticipate problems that will be encountered when the pumps are started or stopped and when the pipe is filled and drained. The design should also consider the effect of variations in flow requirements, and also anticipate problems that will be encountered due to increased future demands and details of the installation of the pumps.

Selecting a pump for a particular service requires matching the system requirements to the capabilities of the pump. The process consists of developing a system equation by applying the energy equation to evaluate the pumping head required to overcome the elevation difference, friction, and minor losses. For a pump supplying water between two reservoirs, the pump head required to produce a given discharge can be expressed as

$$Hp = \Delta Z + H_f \tag{3.4.17}$$

or

$$Hp = \Delta Z + CQ^2 \tag{3.4.18}$$

in which the constant C is defined by Equation (3.4.13).

Figure 3.4.2 shows a system curve for a pipe having an 82-m elevation lift and moderate friction losses. If the elevation of either reservoir is a variable, then there is not a single curve but a family of curves corresponding to differential reservoir elevations.

The three pump curves shown in Figure 3.4.2 represent different impeller diameters. The intersections of the system curve with the pump curves identify the flow that each pump will supply if installed in that system. For this example both A and B pumps would be a good choice because they both operate at or near their best efficiency range. Figure 3.4.2 shows the head and flow that the B pump will produce

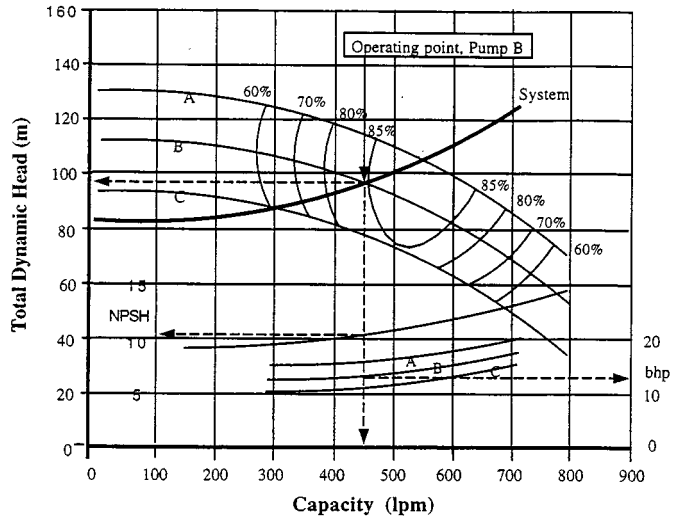


FIGURE 3.4.2 Pump selection for a single pump.

when operating in that system are 97 m and 450 L/m. The net positive suction head (NPSH) and brake horsepower (bhp) are obtained as shown in the figure.

The selection process is more complex when the system demand varies, either due to variations in the water surface elevation or to changing flow requirements. If the system must operate over a range of reservoir elevations, the pump should be selected so that the system curve, based on the mean (or the most frequently encountered) water level, intersects the pump curve near the midpoint of the best efficiency range. If the water level variation is not too great, the pump may not be able to operate efficiently over the complete flow range.

The problem of pump selection also becomes more difficult when planning for future demands or if the pumps are required to supply a varying flow. If the flow range is large, multiple pumps or a variable-speed drive may be needed. Recent developments in variable-frequency drives for pumps make them a viable alternative for systems with varying flows. Selection of multiple pumps and the decision about installing them in parallel or in series depend on the amount of friction in the system. Parallel installations are most effective for low-friction systems. Series pumps work best in high-friction systems.

For parallel pump operation the combined two pump curve is constructed by adding the flow of each pump. Such a curve is shown in Figure 3.4.3 (labeled 2 pumps). The intersection of the two-pump curve with the system curve identifies the combined flow for the two pumps. The pump efficiency for each pump is determined by projecting horizontally to the left to intersect the single-pump curve. For this example, a C pump, when operating by itself, will be have an efficiency of 83%. With two pumps operating, the efficiency of each will be about 72%. For the two pumps to operate in the most efficient way, the selection should be made so the system curve intersects the single-pump curve to the right of its best efficiency point.

Starting a pump with the pipeline empty will result in filling at a very rapid rate because initially there is little friction to build backpressure. As a result, the pump will operate at a flow well above the design flow. This may cause the pump to cavitate, but the more serious problem is the possibility of high pressures generated by the rapid filling of the pipe. Provisions should be made to control the rate of filling to a safe rate. Start-up transients are often controlled by starting the pump against a partially open discharge valve located near the pump and using a bypass line around the pump. This allows the system to be filled slowly and safely. If the pipe remains full and no air is trapped, after the initial filling, subsequent start-up of the pumps generally does not create any serious problem. Adequate air release valves should be installed to release the air under low pressure.

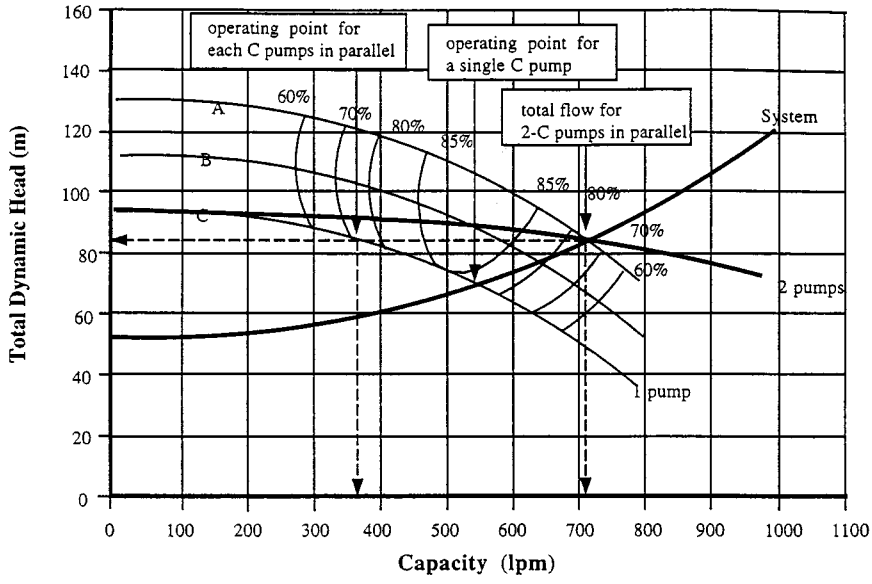


FIGURE 3.4.3 Selection of parallel pumps.

For some systems, stopping the pump, either intentionally or accidentally, can generate high pressures that can damage the pipe and controls. If the design process does not consider these potential problems, the system may not function trouble free. Downtime and maintenance costs may be high. Not all systems will experience start-up and shutdown problems, but the design should at least consider the possibility. The problem is more severe for pipelines that have a large elevation change and multiple high points. The magnitude of the transient is related to the length and profile of the pipeline, the pump characteristics, the magnitude of the elevation change, and the type of check valve used. The downsurge caused by stopping the pump can cause column separation and high pressures due to flow reversals and closure of the check valves. Surge-protection equipment can be added to such systems to prevent damage and excessive maintenance.

Another operational problem occurs with parallel pumps. Each pump must have a check valve to prevent reverse flow. When one of the pumps is turned off, the flow reverses almost immediately in that line because of the high manifold pressure supplied by the operating pumps. This causes the check valve to close. If a slow-closing check valve is installed, the flow can attain a high reverse velocity before the valve closes, generating high pressure transients.

Numerous mechanical devices and techniques have been used to suppress pump shutdown transients. These include increasing the rotational inertia of the pump, use of surge tanks or air chambers near the pump, pressure relief valves, vacuum-breaking valves, and surge-anticipating valves. Selection of the proper transient control device will improve reliability, extend the economic life of the system, and reduce maintenance. Failure to complete a transient analysis and include the required controls will have the opposite effect. A system is only as good as it is designed to be.

Other Considerations

Feasibility Study

Designing pipelines, especially long transmission lines, involves more than just determining the required type and size of pipe. A starting point for major projects is usually a feasibility study which involves

social, environmental, political, and legal issues, as well as an economic evaluation of the engineering alternatives developed during the preliminary design. The preliminary design should identify the scope of the project and all major features that influence the cost or viability. Since local laws, social values, and environmental concerns vary significantly between geographic areas, the engineer must be aware of the problems unique to the area.

Choices that affect the economics of the project include alternative pipe routes, amount of storage and its effect on reliability and controllability of flow, choice of pipe material, diameter and pressure class, provision for future demands, etc. In making decisions one must consider both the engineering and economic advantages of the alternatives. Reliability, safety, maintenance, operating, and replacement costs must all be given their proper value. The analysis should consider (1) expected life of the pipe, which is a function of the type of pipe material and the use of linings or protective coatings; (2) economic life, meaning how long the pipe will supply the demand; (3) planning for future demand; (4) pumping cost vs. pipe cost; and (5) provisions for storage.

During the feasibility study only a general design has been completed so a detailed analysis of all hydraulic problems and their solutions is not available. Even so, it is necessary to anticipate the need for special facilities or equipment and problems such as safe filling, provisions for draining, cavitation at control valves, and transient problems caused by valve or pump operation. Provisions should be made for the cost of the detailed analysis, design, and construction costs required to control special operational problems. Attention should also be given to costs associated with winterizing, stream crossings, highways crossing, special geologic or topographic problems, and any other items that would have a significant influence on the cost, reliability, or safety of the project.

Storage

The purposes of storage tanks and intermediate reservoirs include (1) to supply water when there is a temporary interruption of flow from the supply, (2) to provide supplemental water during peak periods, (3) to sectionalize the pipe to reduce mean and transient pressures, (4) to maintain pressure (elevated storage), and (5) to simplify control. Storage also has a significant impact on the control structures, pumping plants, and general operation of the pipeline. If there is adequate storage, large fluctuations in demand can be tolerated. Any mismatch in supply and demand is made up for by an increase or decrease in storage, and valves in the transmission main will require only infrequent adjustments to maintain storage. Pumps can be activated by level controls at the storage tank and not by fluctuations in demand so they can operate for long periods near their design point.

If there is no storage, the system may have to provide continuous fine adjustment of the flow to provide the required flow within safe pressure limits. For gravity systems this may require automatic pressure- or flow-regulating valves. For pumped systems, the variations in flow can cause constant-speed centrifugal pumps to operate both below and above their design point where power consumption is high, efficiency is low, and where there is more chance of operational problems. Selection of a variable-frequency drive can avoid these problems. The selection of multiple pumps vs. a variable-speed pump is primarily a economic decision.

Thrust Blocks

Any time there is a change of pipe alignment, an unbalanced force is developed. The force required to restrain the pipe can be calculated with the two-dimensional, steady state momentum equation. For buried pipelines, this force can be transmitted to the soil with a thrust block. Determining the size of the block and, consequently, the bearing surface area depends on pipe diameter, fluid pressure, deflection angle of the pipe, and bearing capacity of the soil. A convenient monograph for sizing thrust blocks was published in the *Civil Engineering* in 1969 (Morrison, 1969).

References

ASCE. 1992. *Pressure Pipeline Design for Water and Wastewater.* Prepared by the Committee on Pipeline Planning of the Pipeline Division of the American Society of Civil Engineers, New York.

Kalsi Engineering and Tullis Engineering Consultants. 1993. *Application Guide for Check Valves in Nuclear Power Plants*, Revision 1, NP-5479. Prepared for Nuclear Maintenance Applications Center, Charlotte, NC.

Miller, D.S. 1990. *Internal Flow Systems — Design and Performance Prediction*, 2nd ed. Gulf Publishing Company, Houston.

Morrison, E.B. 1969. Monograph for the design of thrust blocks. *Civil Eng.,* 39, June, 55–51.

Spanger, M.G. and Handy, R.L. 1973. *Soil Engineering,* 3rd ed. Intext Eductational Publishers, New York, Chap. 25 and 26.

Streeter, V.L. and Wylie, E.B. 1975. *Fluid Mechanics,* 6th ed. McGraw-Hill, New York, 752 pp.

Thorley, A.R.D. 1989. Check valve behavior under transient flow conditions: a state-of-the-art review. *ASME*, 111, Vol. 2, June. *J. Fluids Engineering: Transactions of the ASME,* pp. 173–183.

Tullis, J.P. 1989. *Hydraulics of Pipelines — Pumps, Valves, Cavitation, Transients*, John Wiley and Sons, New York.

Tullis, J.P. 1993. Cavitation Guide for Control Valves, NUREG/CR-6031, U.S. Nuclear Regulatory Commission, Washington, D.C.

Wood, D.J. 1966. An explicit friction factor relationship. *Civil Eng.,* 36, December, 60–61.

Wylie, E.B. and Streeter, V.L. 1993. *Fluid Transients in Systems,* Prentice-Hall, Englewood Cliffs, N.J.

Further Information

Bean, H.S., ed. 1971. *Fluid Meters, Their Theory and Application,* 6th ed. The American Society of Mechanical Engineers, New York.

Handbook of PVC-Design and Construction. 1979. Uni-Bell Plastic Pipe Association, Dallas.

King, H.W. 1954. *Handbook of Hydraulics — For the Solution of Hydraulic Problems,* 4th ed. Revised by E.F. Brater. McGraw-Hill, New York.

Stephenson, D. 1981. *Pipeline Design for Water Engineers,* 2nd ed. Elsevier Scientific Publishing Company, Distributed in the U.S. by Gulf Publishing Company, Houston.

Stutsman, R.D. 1993. Steel Penstocks, ASCE Manuals and Reports on Engineering Practice No. 79, Energy Division, American Society of Civil Engineers, New York.

Watkins, R.K. and Spangler, M.G., Some Characteristics of the Modulus of Passive Resistance of Soil: A Study in Similitude. Highway Research Board Proceedings Vol. 37, 1958, pp. 576–583.

Tullis, J.P. 1996. Valves, in Dorf, R.C., Ed., *The Engineering Handbook,* CRC Press, Boca Raton, FL.

3.5 Open Channel Flow*

Frank M. White

Definition

The term *open channel flow* denotes the gravity-driven flow of a liquid with a free surface. Technically, we may study any flowing liquid and any gas interface. In practice, the vast majority of open channel flows concern water flowing beneath atmospheric air in artificial or natural channels.

The geometry of an arbitrary channel is shown in Figure 3.5.1. The area A is for the water cross section only, and b is its top width. The wetted perimeter P covers only the bottom and sides, as shown, not the surface (whose air resistance is neglected). The water depth at any location is y, and the channel slope is θ, often denoted as $S_o = \sin\theta$. All of these parameters may vary with distance x along the channel. In unsteady flow (not discussed here) they may also vary with time.

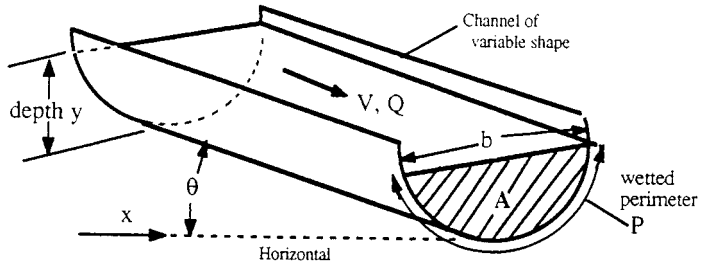


FIGURE 3.5.1 Definition sketch for an open channel.

Uniform Flow

A simple reference condition, called *uniform flow,* occurs in a long straight prismatic channel of constant slope S_o. There is no acceleration, and the water flows at constant depth with fluid weight exactly balancing the wetted wall shear force: $\rho g L A \sin\theta = \tau_w P L$, where L is the channel length. Thus, $\tau_w = \rho g R_h S_o$, where $R_h = A/P$ is called the *hydraulic radius* of the channel. If we relate wall shear stress to the Darcy friction factor f, $\tau_w = (f/8)\rho V^2$, we obtain the basic uniform flow open channel relation:

$$\text{Uniform flow:}\quad V = \sqrt{\frac{8g}{f}}\sqrt{R_h S_o}, \quad \text{where} \quad \sqrt{\frac{8g}{f}} = C = \text{Chézy coefficient} \tag{3.5.1}$$

Antoine Chézy first derived this formula in 1769. It is satisfactory to base f upon the pipe-flow Moody diagram (Figure 3.4.1) using the hydraulic diameter, $D_h = 4R_h$, as a length scale. That is, $f = fcn\,(VD_h/\nu, \varepsilon/D_h)$ from the Moody chart. In ordinary practice, however, engineers assume fully rough, high-Reynolds-number flow and use Robert Manning's century-old correlation:

$$C \approx \frac{\zeta}{n} R_h^{1/6}, \quad \text{or} \quad V_{\text{uniform}} \approx \frac{\zeta}{n} R_h^{2/3} S_o^{1/2} \quad \text{and} \quad Q = VA \tag{3.5.2}$$

where ζ is a conversion factor equal to 1.0 in SI units and 1.486 in English units. The quantity n is Manning's roughness parameter, with typical values, along with the associated roughness heights ε, listed in Table 3.5.1.

* From White, F., *Fluid Mechanics,* 3rd ed., 1994; reproduced with permission of MacGraw-Hill, Inc. Nomenclature appears at end of this section.

TABLE 3.5.1 Average Roughness Parameters for Various Channel Surfaces

		Average Roughness Height ε	
	n	**ft**	**mm**
Artificial lined channels			
Glass	0.010 ± 0.002	0.0011	0.3
Brass	0.011 ± 0.002	0.0019	0.6
Steel; smooth	0.012 ± 0.002	0.0032	1.0
Painted	0.014 ± 0.003	0.0080	2.4
Riveted	0.015 ± 0.002	0.012	3.7
Cast iron	0.013 ± 0.003	0.0051	1.6
Cement; finished	0.012 ± 0.002	0.0032	1.0
Unfinished	0.014 ± 0.002	0.0080	2.4
Planed wood	0.012 ± 0.002	0.0032	1.0
Clay tile	0.014 ± 0.003	0.0080	2.4
Brickwork	0.015 ± 0.002	0.012	3.7
Asphalt	0.016 ± 0.003	0.018	5.4
Corrugated metal	0.022 ± 0.005	0.12	37
Rubble masonry	0.025 ± 0.005	0.26	80
Excavated earth channels			
Clean	0.022 ± 0.004	0.12	37
Gravelly	0.025 ± 0.005	0.26	80
Weedy	0.030 ± 0.005	0.8	240
Stony; cobbles	0.035 ± 0.010	1.5	500
Natural channels			
Clean and straight	0.030 ± 0.005	0.8	240
Sluggish, deep pools	0.040 ± 0.010	3	900
Major rivers	0.035 ± 0.010	1.5	500
Floodplains			
Pasture, farmland	0.035 ± 0.010	1.5	500
Light brush	0.05 ± 0.02	6	2000
Heavy brush	0.075 ± 0.025	15	5000
Trees	0.15 ± 0.05	?	?

Critical Flow

Since the surface is always atmospheric, pressure head is not important in open channel flows. Total energy E relates only to velocity and elevation:

$$\text{Specific energy } E = y + \frac{V^2}{2g} = y + \frac{Q^2}{2gA^2}$$

At a given volume flow rate Q, the energy passes through a minimum at a condition called *critical flow*, where $dE/dy = 0$, or $dA/dy = b = gA^3/Q^2$:

$$A_{crit} = \left(\frac{bQ^2}{g}\right)^{1/3} \qquad V_{crit} = \frac{Q}{A_{crit}} = \left(\frac{gA_{crit}}{b}\right)^{1/2} \tag{3.5.3}$$

where b is the top-surface width as in Figure 3.5.1. The velocity V_{crit} equals the speed of propagation of a surface wave along the channel. Thus, we may define the Froude number Fr of a channel flow, for any cross section, as $Fr = V/V_{crit}$. The three regimes of channel flow are

$$\text{Fr} < 1\text{: subcritical flow;} \quad \text{Fr} = 1\text{: critical flow;} \quad \text{Fr} > 1\text{: supercritical flow}$$

There are many similarities between Froude number in channel flow and Mach number in variable-area duct flow (see Section 3.6).

For a rectangular duct, $A = by$, we obtain the simplified formulas

$$V_{crit} = \sqrt{gy} \qquad \text{Fr} = \frac{V}{\sqrt{gy}} \tag{3.5.4}$$

independent of the width of the channel.

Example 3.5.1

Water (ρ = 998 kg/m^3, μ = 0.001 kg/m · sec) flows uniformly down a half-full brick 1-m-diameter circular channel sloping at 1°. Estimate (a) Q; and (b) the Froude number.

Solution 3.5.1 (a). First compute the geometric properties of a half-full circular channel:

$$A = \frac{\pi}{8}(1\text{ m})^2 = 0.393\text{ m}^2; \quad P = \frac{\pi}{2}(1\text{ m}) = 1.57\text{ m}; \quad R = \frac{A}{P} = \frac{0.393}{1.57} = 0.25\text{ m}$$

From Table 3.5.1, for brickwork, $n \approx 0.015$. Then, Manning's formula, Equation (3.5.2) predicts

$$V = \frac{\zeta}{n} R_h^{1/6} S_o^{1/2} = \frac{1.0}{0.015}(0.25)^{1/6}(\sin 1°)^{1/2} \approx 3.49\frac{\text{m}}{\text{sec}}; \quad Q = 3.49(0.393) \approx \mathbf{1.37}\frac{\mathbf{m^3}}{\mathbf{sec}} \quad \textit{Solution 3.5.1(a)}$$

The uncertainty in this result is about ±10%. The flow rate is quite large (21,800 gal/min) because 1°, although seemingly small, is a substantial slope for a water channel.

One can also use the Moody chart. with $V \approx 3.49$ m/sec, compute Re = $\rho V D_h/\mu \approx 3.49$ E6 and $\varepsilon/D_h \approx 0.0037$, then compute $f \approx 0.0278$ from the Moody chart. Equation (3.5.1) then predicts

$$V = \sqrt{\frac{8g}{f} R_h S_o} = \sqrt{\frac{8(9.81)}{0.0278}(0.25)(\sin 1°)} \approx 3.51\frac{\text{m}}{\text{sec}}; \quad Q = VA \approx \mathbf{1.38}\frac{\mathbf{m^3}}{\mathbf{sec}}$$

Solution 3.5.1 (b). With Q known from part (a), compute the critical conditions from Equation (3.5.3):

$$A_{crit} = \left(\frac{bQ^2}{g}\right)^{1/3} = \left[\frac{1.0(1.37)^2}{9.81}\right]^{1/3} = 0.576\text{ m}^2, \quad V_{crit} = \frac{Q}{A_{crit}} = \frac{1.37}{0.576} = 2.38\ \frac{\text{m}}{\text{sec}}$$

Hence

$$\text{Fr} = \frac{V}{V_{crit}} = \frac{3.49}{2.38} \approx \mathbf{1.47}\square(\text{supercritical}) \quad \textit{Solution 3.5.1(b)}$$

Again the uncertainty is approximately ±10%, primarily because of the need to estimate the brick roughness.

Hydraulic Jump

In gas dynamics (Section 3.6), a supersonic gas flow may pass through a thin normal shock and exit as a subsonic flow at higher pressure and temperature. By analogy, a supercritical open channel flow may pass through a *hydraulic jump* and exit as a subcritical flow at greater depth, as in Figure 3.5.2. Application of continuity and momentum to a jump in a rectangular channel yields

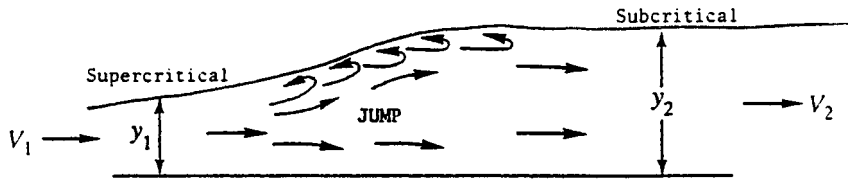


FIGURE 3.5.2 A two-dimensional hydraulic jump.

$$V_2 = V_1 \frac{y_1}{y_2} \qquad y_2 = \frac{y_1}{2}\left[-1 + \sqrt{1 + 8\mathrm{Fr}_1^2}\right] \quad \text{where} \quad \mathrm{Fr}_1 = \frac{V_1}{\sqrt{g y_1}} > 1 \tag{3.5.5}$$

Both the normal shock and the hydraulic jump are dissipative processes: the entropy increases and the effective energy decreases. For a rectangular jump,

$$\Delta E = E_1 - E_2 = \frac{\left(y_2 - y_1\right)^3}{4 y_1 y_2} > 0 \tag{3.5.6}$$

For strong jumps, this loss in energy can be up to 85% of E_1. The second law of thermodynamics requires $\Delta E > 0$ and $y_2 > y_1$ or, equivalently, $\mathrm{Fr}_1 > 1$,

Note from Figure 3.5.2 that a hydraulic jump is not thin. Its total length is approximately four times the downstream depth. Jumps also occur in nonrectangular channels, and the theory is much more algebraically laborious.

Weirs

If an open channel flow encounters a significant obstruction, it will undergo rapidly varied changes which are difficult to model analytically but can be correlated with experiment. An example is the *weir* in Figure 3.5.3 (colloquially called a *dam*), which forces the flow to deflect over the top. If $L \ll Y$, the weir is termed *sharp-crested*; if $L = \mathrm{O}(Y)$ it is *broad-crested*. Small details, such as the upper front corner radius or the crest roughness, may be significant. The crest is assumed level and of width b into the paper.

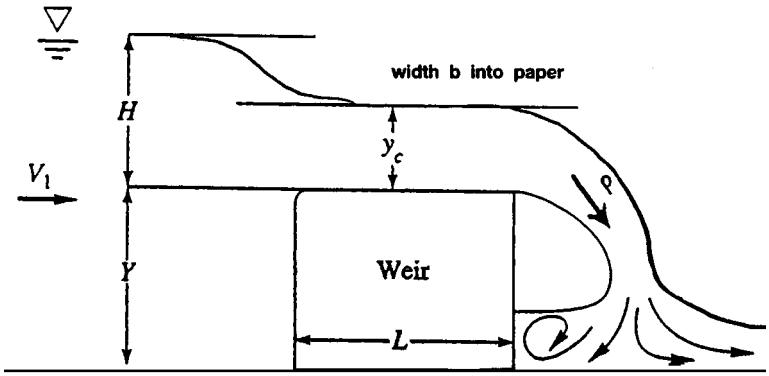


FIGURE 3.5.3 Geometry and notation for flow over a weir.

If there is a free overfall, as in Figure 3.5.3, the flow accelerates from subcritical upstream to critical over the crest to supercritical in the overfall. There is no flow when the excess upstream depth $H = 0$. A simple Bernoulli-type analysis predicts that the flow rate Q over a wide weir is approximately proportional to $bg^{1/2}H^{3/2}$. An appropriate correlation is thus

$$Q_{\text{weir}} = C_d b g^{1/2} H^{3/2}, \quad \text{where } C_d = \text{dimensionless weir coefficient} \tag{3.5.7}$$

If the upstream flow is turbulent, the weir coefficient depends only upon geometry, and Reynolds number effects are negligible. If the weir has sidewalls and is narrow, replace width b by $(b - 0.1H)$.

Two recommended empirical correlations for Equation (3.5.7) are as follows:

$$\begin{aligned} &\text{Sharp-crested:} \quad C_d \approx 0.564 + 0.0846\frac{H}{Y} \quad \text{for} \quad \frac{L}{Y} < 0.07 \\ &\text{Broad-crested:} \quad C_d \approx 0.462 \quad \text{for} \quad 0.08 < \frac{H}{L} < 0.33 \end{aligned} \tag{3.5.8}$$

These data are for wide weirs with a sharp upper corner in front. Many other weir geometries are discussed in the references for this section. Of particular interest is the sharp-edged vee-notch weir, which has no length scale b. If 2θ is the total included angle of the notch, the recommended correlation is

$$\text{Vee-notch, angle } 2\theta: \quad Q \approx 0.44 \tan\theta\, g^{1/2} H^{5/2} \quad \text{for} \quad 10° < \theta \le 50° \tag{3.5.9}$$

The vee-notch is more sensitive at low flow rates (large H for a small Q) and thus is popular in laboratory measurements of channel flow rates.

A weir in the field will tend to spring free and form a natural *nappe*, or air cavity, as in Figure 3.5.3. Narrow weirs, with sidewalls, may need to be aerated artificially to form a nappe and keep the flow from sliding down the face of the weir. The correlations above assume nappe formation.

Gradually Varied Flow

Return to Figure 3.5.1 and suppose that (y, A, b, P, S_o) are all variable functions of horizontal position x. If these parameters are slowly changing, with no hydraulic jumps, the flow is termed *gradually varied* and satisfies a simple one-dimensional first-order differential equation if Q = constant:

$$\frac{dy}{dx} \approx \frac{S_o - S}{1 - \dfrac{V^2 b}{gA}}, \quad \text{where } V = \frac{Q}{A} \quad \text{and} \quad S = \frac{f}{D_h}\frac{V^2}{2g} = \frac{n^2 V^2}{\zeta^2 R_h^{4/3}} \tag{3.5.10}$$

The conversion factor $\zeta^2 = 1.0$ for SI units and 2.208 for English units. If flow rate, bottom slope, channel geometry, and surface roughness are known, we may solve for $y(x)$ for any given initial condition $y = y_o$ at $x = x_o$. The solution is computed by any common numerical method, e.g., Runge–Kutta.

Recall from Equation (3.5.3) that the term $V^2b/(gA) \equiv \text{Fr}^2$, so the sign of the denominator in Equation (3.5.10) depends upon whether the flow is sub- or supercritical. The mathematical behavior of Equation (3.5.10) differs also. If Fr is near unity, the change dy/dx will be very large, which probably violates the basic assumption of "gradual" variation.

For a given flow rate and local bottom slope, two reference depths are useful and may be computed in advance:

(a) The *normal* depth y_n for which Equation (3.5.2) yields the flow rate:
(b) The *critical* depth y_c for which Equation (3.5.3) yields the flow rate.

Comparison of these two, and their relation to the actual local depth y, specifies the type of solution curve being computed. The five bottom-slope regimes (mild M, critical C, steep S, horizontal H, and adverse A) create 12 different solution curves, as illustrated in Figure 3.5.4. All of these may be readily generated by a computer solution of Equation 3.5.10. The following example illustrates a typical solution to a gradually varied flow problem.

Example 3.5.2

Water, flowing at 2.5 m^3/sec in a rectangular gravelly earth channel 2 m wide, encounters a broad-crested weir 1.5 m high. Using gradually varied theory, estimate the water depth profile back to 1 km upstream of the weir. The bottom slope is 0.1^0.

Solution. We are given Q, $Y = 1.5$ m, and $b = 2$ m. We may calculate excess water level H at the weir (see Figure 3.5.3) from Equations (3.5.7) and (3.5.8):

$$Q = 2.5\frac{\text{m}^3}{\text{sec}} = C_d b_{\text{eff}} g^{1/2} H^{3/2} = 0.462(2.0 - 0.1H)(9.81)^{1/2} H^{3/2}, \quad \text{solve for } H \approx \mathbf{0.94}\ \text{m}$$

Since the weir is not too wide, we have subtracted 0.1 H from b as recommended. The weir serves as a "control structure" which sets the water depth just upstream. This is our initial condition for gradually varied theory: $y(0) = Y + H = 1.5 + 0.94 \approx 2.44$ m at $x = 0$. Before solving Equation (3.5.10), we find the normal and critical depths to get a feel for the problem:

$$\text{Normal depth:} \qquad Q = 2.5\frac{\text{m}^3}{\text{sec}} = \frac{1.0}{0.025}(2.0y_n)\left(\frac{2.0y_n}{2.0 + 2y_n}\right)^{3/2}\sqrt{\sin 0.1^\circ}, \quad \text{solve } y_n \approx \mathbf{1.14}\ \text{m}$$

$$\text{Critical depth:} \qquad A_c = 2.0y_c - \left(\frac{bQ^2}{g}\right)^{1/3} = \left[\frac{2.0(2.5)^2}{9.81}\right]^{1/3}, \quad \text{solve } y_c \approx \mathbf{0.54}\ \text{m}$$

We have taken $n \approx 0.025$ for gravelly earth, from Table 3.5.1. Since $y(0) > y_n > y_c$, we are on a mild slope $M - 1$ "backwater" curve, as in Figure 3.5.4. For our data, Equation (3.5.10) becomes

$$\frac{dy}{dx} \approx \frac{S_o - n^2Q^2/\left(\zeta^2 A^2 R_h^{4/3}\right)}{1 - Q^2 b/\left(gA^3\right)}$$

where $Q = 2.5$, $b = 2$, $\zeta = 1$, $A = 2y$, $S_o = \sin 0.1^\circ$, $R_h = 2y/(2 + 2y)$, $g = 9.81$, $y(0) = 2.44$ at $x = 0$.

Integrate numerically backward, that is, for $\Delta x < 0$, until $x = -1$ km $= -1000$ m. The complete solution curve is shown in Figure 3.5.5. The water depth decreases upstream and is approximately $y \approx 1.31$ m at $x = -1000$ m. If slope and channel width remain constant, the water depth asymptotically approaches the normal depth y_n far upstream.

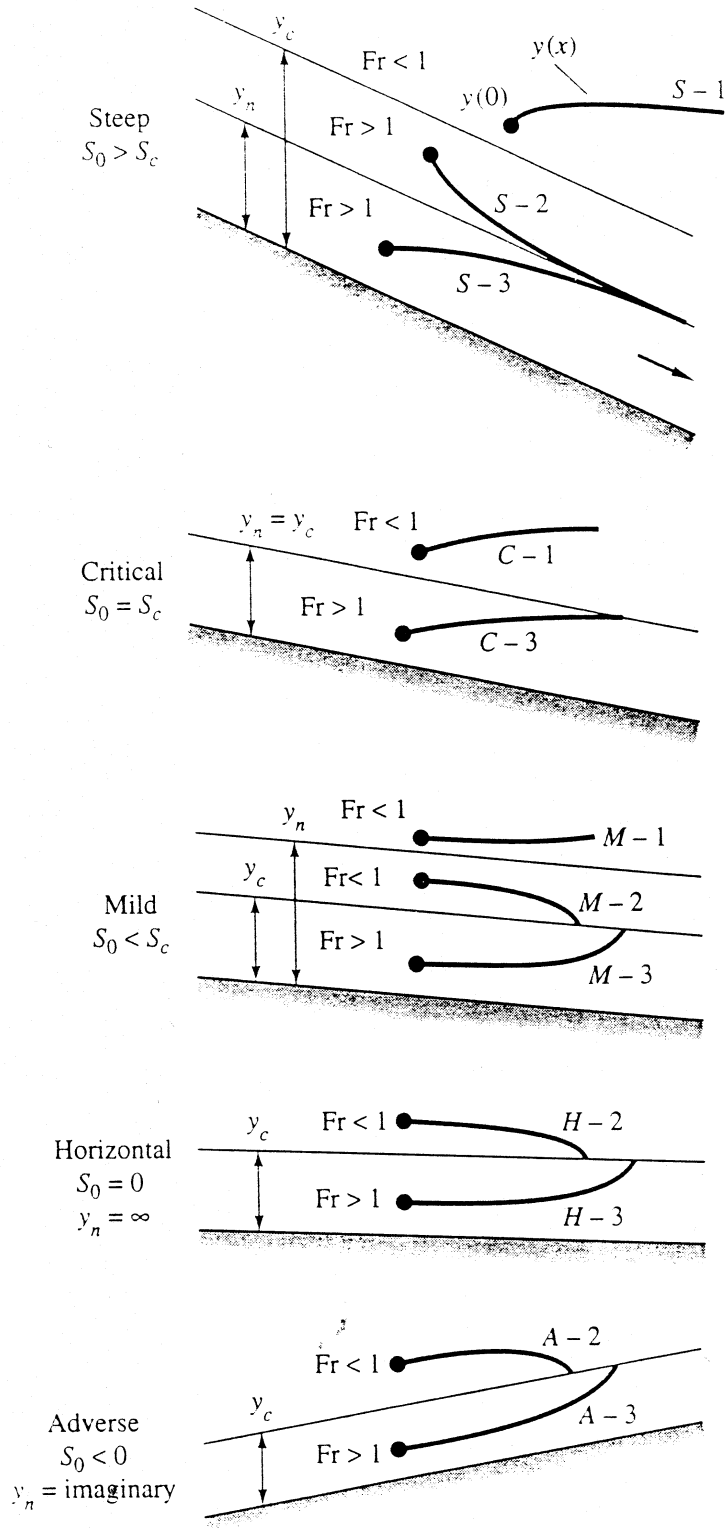


FIGURE 3.5.4 Classification of solution curves for gradually varied flow.

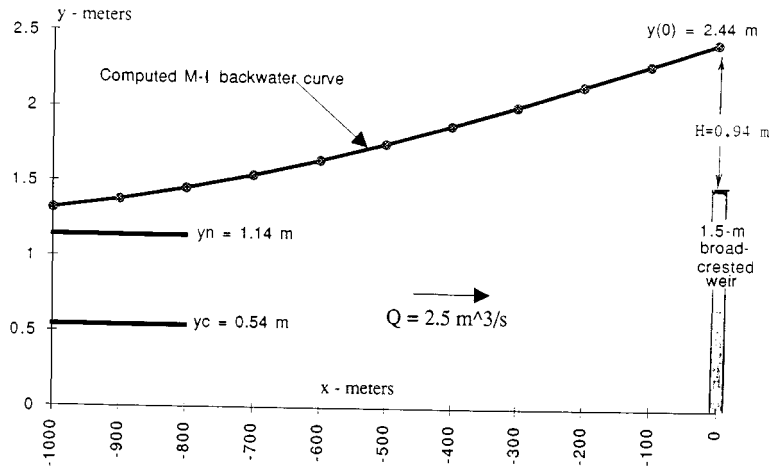


FIGURE 3.5.5 Backwater solution curve for Example 3.5.2.

Nomenclature

English symbols

A = water cross section area
b = channel upper-surface width
C = Chézy coefficient, Equation (3.5.1)
C_d = weir discharge coefficient, Equation (3.5.7)
D_h = hydraulic diameter, = $4R_h$
E = specific energy, = $y + V^2/2g$
f = Moody friction factor
Fr = Froude number, = V/V_{crit}
g = acceleration of gravity
H = excess water level above weir, Figure 3.5.3
L = weir length, Figure 3.5.3
n = Manning roughness factor, Table 3.5.1
P = wetted perimeter
Q = volume flow rate
R_h = hydraulic radius, = A/P
S = frictional slope, Equation (3.5.10)
S_o = bottom slope
V = average velocity
x = horizontal distance along the channel
y = water depth
Y = weir height, Figure (3.5.3

Greek Symbols

ε = wall roughness height, Table 3.5.1
ρ = fluid density
μ = fluid viscosity
ν = fluid kinematic viscosity, = μ/ρ
ζ = conversion factor, = 1.0 (SI) and 1.486 (English)

Subscripts

c,crit = critical, at Fr = 1
n = normal, in uniform flow

References

Ackers, P. et al. 1978. *Weirs and Flumes for Flow Measurement*, John Wiley, New York.

Bos, M.G. 1985. *Long-Throated Flumes and Broad-Crested Weirs,* Martinus Nijhoff (Kluwer), Dordrecht, The Netherlands.

Bos, M.G., Replogle, J.A., and Clemmens, A.J. 1984. *Flow-Measuring Flumes for Open Channel Systems,* John Wiley, New York.

Brater, E.F. 1976. *Handbook of Hydraulics,* 6th ed., McGraw-Hill, New York.

Chow, V.T. 1959. *Open Channel Hydraulics,* McGraw-Hill, New York.

French, R.H. 1985. *Open Channel Hydraulics,* McGraw-Hill, New York.

Henderson, F.M. 1966. *Open Channel Flow,* Macmillan, New York.

Sellin, R.H.J. 1970. *Flow in Channels*, Gordon & Breach, London.

Spitzer, D.W. (Ed.). 1991. *Flow Measurement: Practical Guides for Measurement and Control*, Instrument Society of America, Research Triangle Park, NC.

3.6 External Incompressible Flows

Alan T. McDonald

Introduction and Scope

Potential flow theory (Section 3.2) treats an incompressible *ideal fluid* with zero viscosity. There are no shear stresses; pressure is the only stress acting on a fluid particle. Potential flow theory predicts no drag force when an object moves through a fluid, which obviously is not correct, because all real fluids are viscous and cause drag forces. The objective of this section is to consider the behavior of viscous, incompressible fluids flowing over objects.

A number of phenomena that occur in external flow at high Reynolds number over an object are shown in Figure 3.6.1. The freestream flow divides at the stagnation point and flows around the object. Fluid at the object surface takes on the velocity of the body as a result of the no-slip condition. Boundary layers form on the upper and lower surfaces of the body; flow in the boundary layers is initially laminar, then **transition** to turbulent flow may occur (points "T").

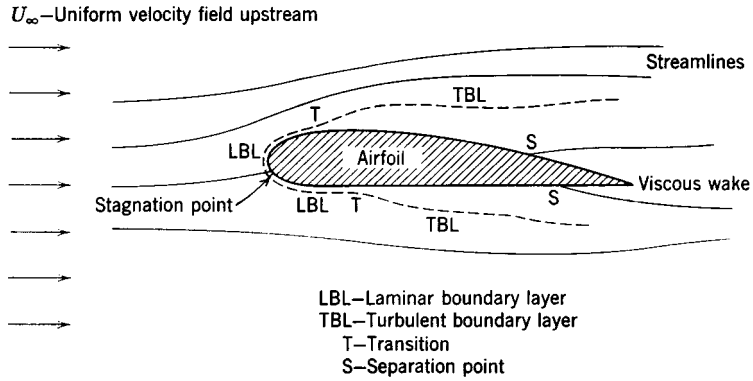


FIGURE 3.6.1 Viscous flow around an airfoil (boundary layer thickness exaggerated for clarity).

Boundary layers thickening on the surfaces cause only a slight displacement of the streamlines of the external flow (their thickness is greatly exaggerated in the figure). **Separation** may occur in the region of increasing pressure on the rear of the body (points "S"); after separation boundary layer fluid no longer remains in contact with the surface. Fluid that was in the boundary layers forms the viscous *wake* behind the object.

The Bernoulli equation is valid for steady, incompressible flow without viscous effects. It may be used to predict pressure variations outside the boundary layer. Stagnation pressure is constant in the uniform inviscid flow far from an object, and the Bernoulli equation reduces to

$$p_\infty + \frac{1}{2}\rho V^2 = \text{constant} \tag{3.6.1}$$

where p_∞ is pressure far upstream, ρ is density, and V is velocity. Therefore, the local pressure can be determined if the local freestream velocity, U, is known.

Boundary Layers

The Boundary Layer Concept

The **boundary layer** is the thin region near the surface of a body in which viscous effects are important. By recognizing that viscous effects are concentrated near the surface of an object, Prandtl showed that only the Euler equations for inviscid flow need be solved in the region outside the boundary layer. Inside the boundary layer, the elliptic Navier-Stokes equations are simplified to boundary layer equations with parabolic form that are easier to solve. The thin boundary layer has negligible pressure variation across it; pressure from the freestream is impressed upon the boundary layer.

Basic characteristics of all laminar and turbulent boundary layers are shown in the developing flow over a flat plate in a semi-infinite fluid. Because the boundary layer is thin, there is negligible disturbance of the inviscid flow outside the boundary layer, and the **pressure gradient** along the surface is close to zero. Transition from laminar to turbulent boundary layer flow on a flat plate occurs when Reynolds number based on x exceeds $\mathrm{Re}_x = 500{,}000$. Transition may occur earlier if the surface is rough, pressure increases in the flow direction, or separation occurs. Following transition, the turbulent boundary layer thickens more rapidly than the laminar boundary layer as a result of increased shear stress at the body surface.

Boundary Layer Thickness Definitions

Boundary layer disturbance thickness, δ, is usually defined as the distance, y, from the surface to the point where the velocity within the boundary layer, u, is within 1% of the local freestream velocity, U. As shown in Figure 3.6.2, the boundary layer velocity profile merges smoothly and asymptotically into the freestream, making δ difficult to measure. For this reason and for their physical significance, we define two integral measures of boundary layer thickness. Displacement thickness, δ^*, is defined as

$$\frac{\delta^*}{\delta} = \int_0^\infty \left(1 - \frac{u}{U}\right) d\left(\frac{y}{\delta}\right) \tag{3.6.2}$$

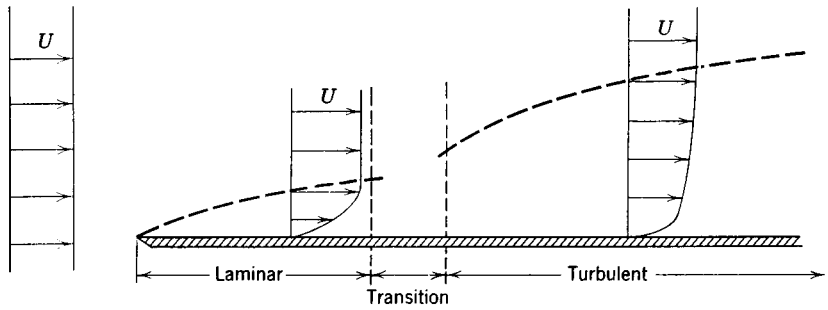


FIGURE 3.6.2 Boundary layer on a flat plate (vertical thickness exaggerated for clarity).

Physically, δ^* is the distance the solid boundary would have to be displaced into the freestream in a frictionless flow to produce the mass flow deficit caused by the viscous boundary layer. Momentum thickness, θ, is defined as

$$\frac{\theta}{\delta} = \int_0^\infty \frac{u}{U}\left(1 - \frac{u}{U}\right) d\left(\frac{y}{\delta}\right) \tag{3.6.3}$$

Physically, θ is the thickness of a fluid layer, having velocity U, for which the momentum flux is the same as the deficit in momentum flux within the boundary layer (momentum flux is momentum per unit time passing a cross section).

Because δ^* and θ are defined in terms of integrals for which the integrand vanishes in the freestream, they are easier to evaluate experimentally than disturbance thickness δ.

Exact Solution of the Laminar Flat-Plate Boundary Layer

Blasius obtained an exact solution for laminar boundary layer flow on a flat plate. He assumed a thin boundary layer to simplify the streamwise momentum equation. He also assumed *similar* velocity profiles in the boundary layer, so that when written as $u/U = f(y/\delta)$, velocity profiles do not vary with x. He used a similarity variable to reduce the partial differential equations of motion and continuity to a single third-order ordinary differential equation.

Blasius used numerical methods to solve the ordinary differential equation. Unfortunately, the velocity profile must be expressed in tabular form. The principal results of the Blasius solution may be expressed as

$$\frac{\delta}{x} = \frac{5}{\sqrt{\mathrm{Re}_x}} \tag{3.6.4}$$

and

$$C_f = \frac{\tau_w}{\frac{1}{2}\rho U^2} = \frac{0.664}{\sqrt{\mathrm{Re}_x}} \tag{3.6.5}$$

These results characterize the laminar boundary layer on a flat plate; they show that laminar boundary layer thickness varies as $x^{1/2}$ and wall shear stress varies as $1/x^{1/2}$.

Approximate Solutions

The Blasius solution cannot be expressed in closed form and is limited to laminar flow. Therefore, approximate methods that give solutions for both laminar and turbulent flow in closed form are desirable. One such method is the *momentum integral equation* (MIE), which may be developed by integrating the boundary layer equation across the boundary layer or by applying the streamwise momentum equation to a differential control volume (Fox and McDonald, 1992). The result is the ordinary differential equation

$$\frac{d\theta}{dx} = \frac{\tau_w}{\rho U^2} - \left(\frac{\delta^*}{\theta} + 2\right)\frac{\theta}{U}\frac{dU}{dx} \tag{3.6.6}$$

The first term on the right side of Equation (3.6.6) contains the influence of wall shear stress. Since τ_w is always positive, it always causes θ to increase. The second term on the right side contains the pressure gradient, which can have either sign. Therefore, the effect of the pressure gradient can be to either increase or decrease the rate of growth of boundary layer thickness.

Equation (3.6.6) is an ordinary differential equation that can be solved for θ as a function of x on a flat plate (zero pressure gradient), provided a reasonable shape is assumed for the boundary layer velocity profile and shear stress is expressed in terms of the other variables. Results for laminar and turbulent flat-plate boundary layer flows are discussed below.

Laminar Boundary Layers. A reasonable approximation to the laminar boundary layer velocity profile is to express u as a polynomial in y. The resulting solutions for δ and τ_w have the same dependence on x as the exact Blasius solution. Numerical results are presented in Table 3.6.1. Comparing the approximate and exact solutions shows remarkable agreement in view of the approximations used in the analysis. The trends are predicted correctly and the approximate values are within 10% of the exact values.

Turbulent Boundary Layers. The turbulent velocity profile may be expressed well using a power law, $u/U = (y/\delta)^{1/n}$, where n is an integer between 6 and 10 (frequently 7 is chosen). For turbulent flow it is

TABLE 3.6.1 Exact and Approximate Solutions for Laminar Boundary Layer Flow over a Flat Plate at Zero Incidence

Velocity Distribution $\frac{u}{U} = f\left(\frac{y}{\delta}\right) = f(\eta)$	$\frac{\theta}{\delta}$	$\frac{\delta^*}{\delta}$	$a = \frac{\delta}{x}\sqrt{\mathrm{Re}_x}$	$b = C_f\sqrt{\mathrm{Re}_x}$
$f(\eta) = 2\eta - \eta^2$	2/15	1/3	5.48	0.730
$f(\eta) = 3/2\ \eta - 1/2\ \eta^3$	39/280	3/8	4.64	0.647
$f(\eta) = \sin(\pi/2\ \eta)$	$(4-\pi)/2\pi$	$(\pi-2)/\pi$	4.80	0.654
Exact	0.133	0.344	5.00	0.664

not possible to express shear stress directly in terms of a simple velocity profile; an empirical correlation is required. Using a pipe flow data correlation gives

$$\frac{\delta}{x} = \frac{0.382}{\mathrm{Re}_x^{1/5}} \tag{3.6.7}$$

and

$$C_f = \frac{\tau_w}{\frac{1}{2}\rho U^2} = \frac{0.0594}{\mathrm{Re}_x^{1/5}} \tag{3.6.8}$$

These results characterize the turbulent boundary layer on a flat plate. They show that turbulent boundary layer thickness varies as $x^{4/5}$ and wall shear stress varies as $1/x^{1/5}$.

Approximate results for laminar and turbulent boundary layers are compared in Table 3.6.2. At a Reynolds number of 1 million, wall shear stress for the turbulent boundary layer is nearly six times as large as for the laminar layer. For a turbulent boundary layer, thickness increases five times faster with distance along the surface than for a laminar layer. These approximate results give a physical feel for relative magnitudes in the two cases.

TABLE 3.6.2 Thickness and Skin Friction Coefficient for Laminar and Turbulent Boundary Layers on a Flat Plate

Reynolds Number	Boundary Layer Thickness/x		Skin Friction Coefficient		Turbulent/Laminar Ratio	
	Laminar BL	Turbulent BL	Laminar BL	Turbulent BL	BL Thickness	Skin Friction
2E + 05	0.0112	0.0333	0.00148	0.00517	2.97	3.48
5E + 05	0.00707	0.0277	0.000939	0.00431	3.92	4.58
1E + 06	0.00500	0.0241	0.000664	0.00375	4.82	5.64
2E + 06	0.00354	0.0210	0.000470	0.00326	5.93	6.95
5E + 06	0.00224	0.0175	0.000297	0.00272	7.81	9.15
1E + 07	0.00158	0.0152	0.000210	0.00236	9.62	11.3
2E + 07	0.00112	0.0132	0.000148	0.00206	11.8	13.9
5E + 07	0.000707	0.0110	0.0000939	0.00171	15.6	18.3

Note: BL = boundary layer.

The MIE cannot be solved in closed form for flows with nonzero pressure gradients. However, the role of the pressure gradient can be understood qualitatively by studying the MIE.

Effect of Pressure Gradient

Boundary layer flow with favorable, zero, and adverse pressure gradients is depicted schematically in Figure 3.6.3. (Assume a thin boundary layer, so flow on the lower surface behaves as external flow on

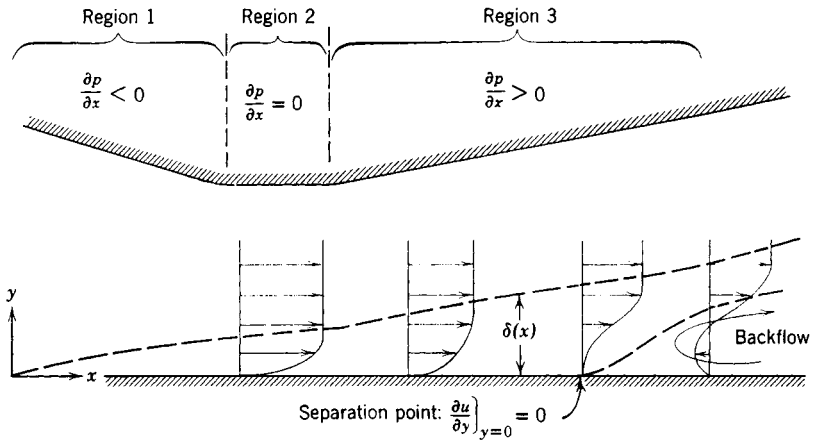


FIGURE 3.6.3 Boundary layer flow with presssure gradient (thickness exaggerated for clarity).

a surface, with the pressure gradient impressed on the boundary layer.) The pressure gradient is favorable when $\partial p/\partial x < 0$, zero when $\partial p/\partial x = 0$, and adverse when $\partial p/\partial x > 0$, as indicated for Regions 1, 2, and 3.

Viscous shear always causes a net retarding force on any fluid particle within the boundary layer. For zero pressure gradient, shear forces alone can never bring the particle to rest. (Recall that for laminar and turbulent boundary layers the shear stress varied as $1/x^{1/2}$ and $1/x^{1/5}$, respectively; shear stress never becomes zero for finite x.) Since shear stress is given by $\tau_w = \mu\ \partial u/\partial y)_{y=0}$, the velocity gradient cannot be zero. Therefore, flow cannot separate in a zero pressure gradient; shear stresses alone can never cause flow separation.

In the favorable pressure gradient of Region 1, pressure forces tend to maintain the motion of the particle, so flow cannot separate. In the adverse pressure gradient of Region 3, pressure forces oppose the motion of a fluid particle. An adverse pressure gradient is a necessary condition for flow separation.

Velocity profiles for laminar and turbulent boundary layers are shown in Figure 3.6.2. It is easy to see that the turbulent velocity profile has much more momentum than the laminar profile. Therefore, the turbulent velocity profile can resist separation in an adverse pressure gradient better than the laminar profile.

The freestream velocity distribution must be known before the MIE can be applied. We obtain a first approximation by applying potential flow theory to calculate the flow field around the object. Much effort has been devoted to calculation of velocity distributions over objects of known shape (the "direct" problem) and to determination of shapes to produce a desired pressure distribution (the "inverse" problem). Detailed discussion of such calculation schemes is beyond the scope of this section; the state of the art continues to progress rapidly.

Drag

Any object immersed in a viscous fluid flow experiences a net force from the shear stresses and pressure differences caused by the fluid motion. *Drag* is the force component parallel to, and *lift* is the force component perpendicular to, the flow direction. *Streamlining* is the art of shaping a body to reduce fluid dynamic drag. Airfoils (hydrofoils) are designed to produce lift in air (water); they are streamlined to reduce drag and thus to attain high lift–drag ratios.

In general, lift and drag cannot be predicted analytically for flows with separation, but progress continues on computational fluid dynamics methods. For many engineering purposes, drag and lift forces are calculated from experimentally derived coefficients, discussed below.

Drag coefficient is defined as

$$C_D = \frac{F_D}{\frac{1}{2}\rho V^2 A} \tag{3.6.9}$$

where $^1/_2\rho V^2$ is dynamic pressure and A is the area upon which the coefficient is based. Common practice is to base drag coefficients on projected *frontal area* (Fox and McDonald, 1992).

Similitude was treated in Section 3.3. In general, the drag coefficient may be expressed as a function of Reynolds number, Mach number, Froude number, relative roughness, submergence divided by length, and so forth. In this section we consider neither high-speed flow nor free-surface effects, so we will consider only Reynolds number and roughness effects on drag coefficient.

Friction Drag

The total friction drag force acting on a plane surface aligned with the flow direction can be found by integrating the shear stress distribution along the surface. The drag coefficient for this case is defined as friction force divided by dynamic pressure and *wetted area* in contact with the fluid. Since shear stress is a function of Reynolds number, so is drag coefficient (see Figure 3.6.4). In Figure 3.6.4, transition occurs at Re_x = 500,000; the dashed line represents the drag coefficient at larger Reynolds numbers. A number of empirical correlations may be used to model the variation in C_D shown in Figure 3.6.4 (Schlichting, 1979).

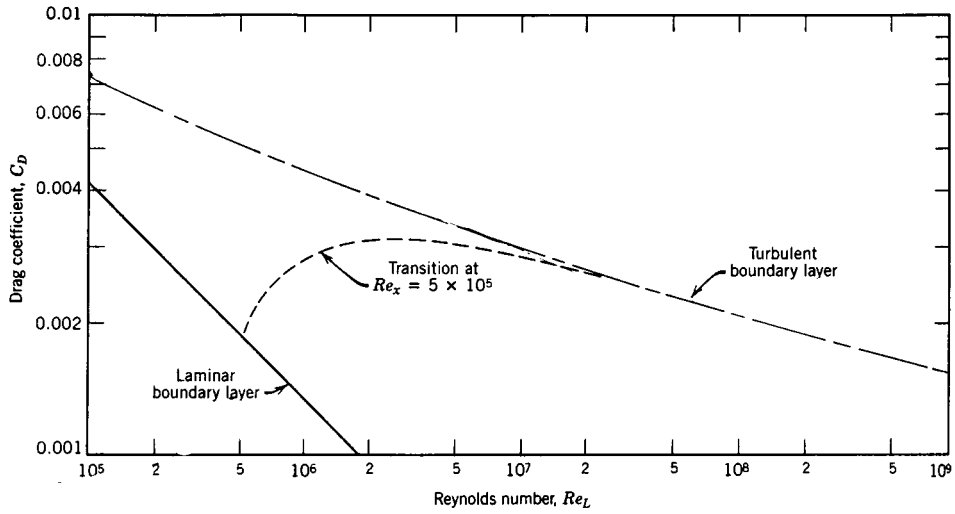


FIGURE 3.6.4 Drag coefficient vs. Reynolds number for a smooth flat plate parallel to the flow.

Extending the laminar boundary layer line to higher Reynolds numbers shows that it is beneficial to delay transition to the highest possible Reynolds number. Some results are presented in Table 3.6.3; drag is reduced more than 50% by extending laminar boundary layer flow to $Re_L = 10^6$.

Pressure Drag

A thin flat surface normal to the flow has no area parallel to the flow direction. Therefore, there can be no friction force parallel to the flow; all drag is caused by pressure forces. Drag coefficients for objects with sharp edges tend to be independent of Reynolds number (for Re > 1000), because the separation points are fixed by the geometry of the object. Drag coefficients for selected objects are shown in Table 3.6.4.

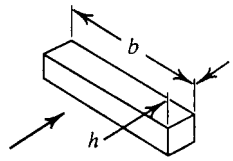
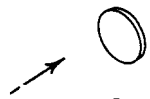
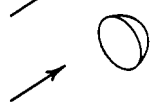
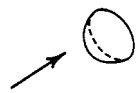
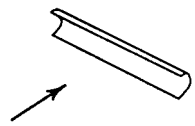
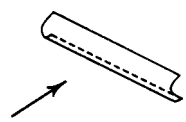
Rounding the edges that face the flow reduces drag markedly. Compare the drag coefficients for the hemisphere and C-section shapes facing into and away from the flow. Also note that the drag coefficient

TABLE 3.6.3 Drag Coefficients for Laminar, Turbulent, and Transition Boundary Layers on a Flat Plate

Reynolds Number	Drag Coefficient			Laminar/ Transition	% Drag Reduction
	Laminar BL	Turbulent BL	Transition		
2E + 05	0.00297	0.00615	—	—	—
5E + 05	0.00188	0.00511	0.00189	—	—
1E + 06	0.00133	0.00447	0.00286	0.464	53.6
2E + 06	0.000939	0.00394	0.00314	0.300	70.0
5E + 06	0.000594	0.00336	0.00304	0.195	80.5
1E + 07	0.000420	0.00300	0.00284	0.148	85.2
2E + 07	0.000297	0.00269	0.00261	0.114	88.6
5E + 07	0.000188	0.00235	0.00232	0.081	9.19

Note: BL = Boundary layer.

TABLE 3.6.4 Drag Coefficient Data for Selected Objects (Re > 1000)

Object	Diagram		$C_D(RE^* \gtrsim 10^3)$
Square prism		$b/h = \infty$	2.05
		$b/h = 1$	1.05
Disk			1.17
Ring			1.20[b]
Hemisphere (open end facing flow)			1.42
Hemisphere (open end facing downstream)			0.38
C-section (open side facing flow)			2.30
C-section (open side facing downstream)			1.20

[a] Data from Hoerner, 1965.
[b] Based on ring area.

for a two-dimensional object (long square cylinder) is about twice that for the corresponding three-dimensional object (square cylinder with $b/h = 1$).

Friction and Pressure Drag: Bluff Bodies

Both friction and pressure forces contribute to the drag of *bluff bodies* (see Shapiro, 1960, for a good discussion of the mechanisms of drag). As an example, consider the drag coefficient for a smooth sphere shown in Figure 3.6.5. Transition from laminar to turbulent flow in the boundary layers on the forward portion of the sphere causes a dramatic dip in drag coefficient at the *critical Reynolds number* ($Re_D \approx 2 \times 10^5$). The turbulent boundary layer is better able to resist the adverse pressure gradient on the rear of the sphere, so separation is delayed and the wake is smaller, causing less pressure drag.

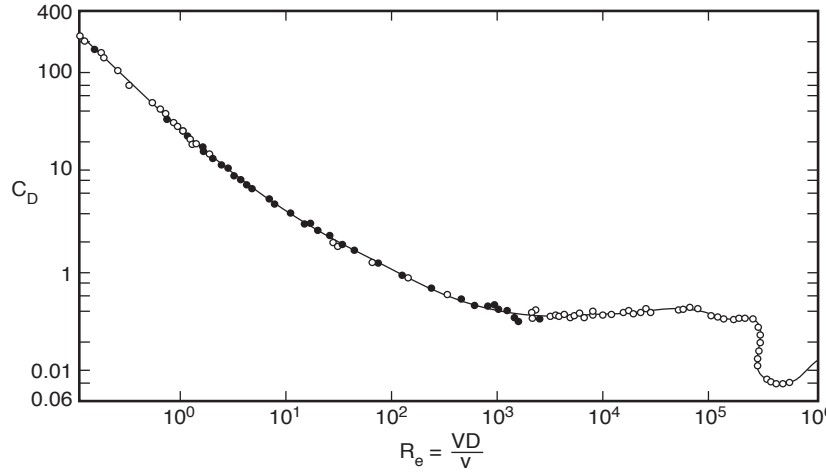


FIGURE 3.6.5 Drag coefficient vs. Reynolds number for a smooth sphere.

Surface roughness (or freestream disturbances) can reduce the critical Reynolds number. Dimples on a golf ball cause the boundary layer to become turbulent and, therefore, lower the drag coefficient in the range of speeds encountered in a drive.

Streamlining

Streamlining is adding a faired tail section to reduce the extent of separated flow on the downstream portion of an object (at high Reynolds number where pressure forces dominate drag). The adverse pressure gradient is taken over a longer distance, delaying separation. However, adding a faired tail increases surface area, causing skin friction drag to increase. Thus, streamlining must be optimized for each shape.

Front contours are of principal importance in road vehicle design; the angle of the back glass also is important (in most cases the entire rear end cannot be made long enough to control separation and reduce drag significantly).

Lift

Lift coefficient is defined as

$$C_L = \frac{F_L}{\frac{1}{2}\rho V^2 A} \tag{3.6.10}$$

Note that lift coefficient is based on projected *planform area*.

Airfoils

Airfoils are shaped to produce lift efficiently by accelerating flow over the upper surface to produce a low-pressure region. Because the flow must again decelerate, inevitably there must be a region of adverse pressure gradient near the rear of the upper surface (pressure distributions are shown clearly in Hazen, 1965).

Lift and drag coefficients for airfoil sections depend on Reynolds number and *angle of attack* between the chord line and the undisturbed flow direction. The *chord line* is the straight line joining the leading and trailing edges of the airfoil (Abbott and von Doenhoff, 1959).

As the angle of attack is increased, the minimum pressure point moves forward on the upper surface and the minimum pressure becomes lower. This increases the adverse pressure gradient. At some angle of attack, the adverse pressure gradient is strong enough to cause the boundary layer to separate completely from the upper surface, causing the airfoil to *stall*. The separated flow alters the pressure distribution, reducing lift sharply.

Increasing the angle of attack also causes the the drag coefficient to increase. At some angle of attack below stall the ratio of lift to drag, the *lift–drag* ratio, reaches a maximum value.

Drag Due to Lift

For wings (airfoils of finite span), lift and drag also are functions of aspect ratio. Lift is reduced and drag increased compared with infinite span, because end effects cause the lift vector to rotate rearward. For a given geometric angle of attack, this reduces effective angle of attack, reducing lift. The additional component of lift acting in the flow direction increases drag; the increase in drag due to lift is called *induced drag*.

The effective aspect ratio includes the effect of planform shape. When written in terms of effective aspect ratio, the drag of a finite-span wing is

$$C_D = C_{D,\infty} + \frac{C_L^2}{\pi ar} \tag{3.6.11}$$

where ar is effective aspect ratio and the subscript ∞ refers to the infinite section drag coefficient at C_L. For further details consult the references.

The lift coefficient must increase to support aircraft weight as speed is reduced. Therefore, induced drag can increase rapidly at low flight speeds. For this reason, minimum allowable flight speeds for commercial aircraft are closely controlled by the FAA.

Boundary Layer Control

The major part of the drag on an airfoil or wing is caused by skin friction. Therefore, it is important to maintain laminar flow in the boundary layers as far aft as possible; laminar flow sections are designed to do this. It also is important to prevent flow separation and to achieve high lift to reduce takeoff and landing speeds. These topics fall under the general heading of boundary layer control.

Profile Shaping

Boundary layer transition on a conventional airfoil section occurs almost immediately after the minimum pressure at about 25% chord aft the leading edge. Transition can be delayed by shaping the profile to maintain a favorable pressure gradient over more of its length. The U.S. National Advisory Committee for Aeronautics (NACA) developed several series of profiles that delayed transition to 60 or 65% of chord, reducing drag coefficients (in the design range) 60% compared with conventional sections of the same thickness ratio (Abbott and von Doenhoff, 1959).

Flaps and Slats

Flaps are movable sections near the trailing edge of a wing. They extend and/or deflect to increase wing area and/or increase wing camber (curvature), to provide higher lift than the clean wing. Many aircraft also are fitted with leading edge slats which open to expose a slot from the pressure side of the wing to the upper surface. The open slat increases the effective radius of the leading edge, improving maximum lift coefficient. The slot allows energized air from the pressure surface to flow into the low-pressure region atop the wing, energizing the boundary layers and delaying separation and stall.

Suction and Blowing

Suction removes low-energy fluid from the boundary layer, reducing the tendency for early separation. Blowing via high-speed jets directed along the surface reenergizes low-speed boundary layer fluid. The objective of both approaches is to delay separation, thus increasing the maximum lift coefficient the wing can achieve. Powered systems add weight and complexity; they also require bleed air from the engine compressor, reducing thrust or power output.

Moving Surfaces

Many schemes have been proposed to utilize moving surfaces for boundary layer control. Motion in the direction of flow reduces skin friction, and thus the tendency to separate; motion against the flow has the opposite effect. The aerodynamic behavior of sports balls — baseballs, golf balls, and tennis balls — depends significantly on aerodynamic side force (lift, down force, or side force) produced by spin. These effects are discussed at length in Fox and McDonald (1992) and its references.

Computation vs. Experiment

Experiments cannot yet be replaced completely by analysis. Progress in modeling, numerical techniques, and computer power continues to be made, but the role of the experimentalist likely will remain important for the foreseeable future.

Computational Fluid Dynamics (CFD)

Computation of fluid flow requires accurate mathematical modeling of flow physics and accurate numerical procedures to solve the equations. The basic equations for laminar boundary layer flow are well known. For turbulent boundary layers generally it is not possible to resolve the solution space into sufficiently small cells to allow direct numerical simulation. Instead, empirical models for the turbulent stresses must be used. Advances in computer memory storage capacity and speed (e.g., through use of massively parallel processing) continue to increase the resolution that can be achieved.

A second source of error in CFD work results from the numerical procedures required to solve the equations. Even if the equations are exact, approximations must be made to discretize and solve them using finite-difference or finite-volume methods. Whichever is chosen, the solver must guard against introducing numerical instability, round-off errors, and numerical diffusion (Hoffman, 1992).

Role of the Wind Tunnel

Traditionally, wind tunnel experiments have been conducted to verify the design and performance of components and complete aircraft. Design verification of a modern aircraft may require expensive scale models, several thousand hours of wind tunnel time at many thousands of dollars an hour, and additional full-scale flight testing.

New wind tunnel facilities continue to be built and old ones refurbished. This indicates a need for continued experimental work in developing and optimizing aircraft configurations.

Many experiments are designed to produce baseline data to validate computer codes. Such systematic experimental data can help to identify the strengths and weaknesses of computational methods.

CFD tends to become only indicative of trends when massive zones of flow separation are present. Takeoff and landing configurations of conventional aircraft, with landing gear, high-lift devices, and

flaps extended, tend to need final experimental confirmation and optimization. Many studies of vertical takeoff and vectored thrust aircraft require testing in wind tunnels.

Defining Terms

Boundary layer: Thin layer of fluid adjacent to a surface where viscous effects are important; viscous effects are negligible outside the boundary layer.

Drag coefficient: Force in the flow direction exerted on an object by the fluid flowing around it, divided by dynamic pressure and area.

Lift coefficient: Force perpendicular to the flow direction exerted on an object by the fluid flowing around it, divided by dynamic pressure and area.

Pressure gradient: Variation in pressure along the surface of an object. For a *favorable* pressure gradient, pressure *decreases* in the flow direction; for an *adverse* pressure gradient, pressure *increases* in the flow direction.

Separation: Phenomenon that occurs when fluid layers adjacent to a solid surface are brought to rest and boundary layers depart from the surface contour, forming a low-pressure *wake* region. Separation can occur only in an *adverse pressure gradient*.

Transition: Change from laminar to turbulent flow within the boundary layer. The location depends on distance over which the boundary layer has developed, pressure gradient, surface roughness, freestream disturbances, and heat transfer.

References

Abbott, I.H. and von Doenhoff, A.E. 1959. *Theory of Wing Sections, Including a Summary of Airfoil Data*. Dover, New York.

Fox, R.W. and McDonald, A.T. 1992. *Introduction to Fluid Mechanics,* 4th ed. John Wiley & Sons, New York.

Hazen, D.C. 1965. *Boundary Layer Control,* film developed by the National Committee for Fluid Mechanics Films (NCFMF) and available on videotape from Encyclopaedia Britannica Educational Corporation, Chicago.

Hoerner, S.F. 1965. *Fluid-Dynamic Drag,* 2nd ed. Published by the author, Midland Park, NJ.

Hoffman, J.D. 1992. *Numerical Methods for Engineers and Scientists*. McGraw-Hill, New York.

Schlichting, H. 1979. *Boundary-Layer Theory*, 7th ed. McGraw-Hill, New York.

Shapiro, A.H. 1960. *The Fluid Dynamics of Drag*, film developed by the National Committee for Fluid Mechanics Film (NCFMF) and available on videotape from Encyclopaedia Britannica Educational Corporation, Chicago.

Further Information

A comprehensive source of basic information is the *Handbook of Fluid Dynamics,* edited by Victor L. Streeter (McGraw-Hill, New York, 1960).

Timely reviews of important topics are published in the *Annual Review of Fluid Mechanics* series (Annual Reviews, Inc., Palo Alto, CA.). Each volume contains a cumulative index.

ASME (American Society of Mechanical Engineers, New York, NY) publishes the *Journal of Fluids Engineering* quarterly. *JFE* contains fluid machinery and other engineering applications of fluid mechanics.

The monthly *AIAA Journal* and bimonthly *Journal of Aircraft* (American Institute for Aeronautics and Astronautics, New York) treat aerospace applications of fluid mechanics.

3.7 Compressible Flow

Ajay Kumar

Introduction

This section deals with compressible flow. Only one- or two-dimensional steady, inviscid flows under perfect gas assumption are considered. Readers are referred to other sources of information for unsteady effects, viscous effects, and three-dimensional flows.

The term *compressible flow* is routinely used to define variable density flow which is in contrast to incompressible flow, where the density is assumed to be constant throughout. In many cases, these density variations are principally caused by the pressure changes from one point to another. Physically, the *compressibility* can be defined as the fractional change in volume of the gas element per unit change in pressure. It is a property of the gas and, in general, can be defined as

$$\tau = \frac{1}{\rho}\frac{d\rho}{dp}$$

where τ is the compressibility of the gas, ρ is the density, and p is the pressure being exerted on the gas. A more precise definition of compressibility is obtained if we take into account the thermal and frictional losses. If during the compression the temperature of the gas is held constant, it is called the isothermal compressibility and can be written as

$$\tau_T = \frac{1}{\rho}\left(\frac{\partial\rho}{\partial p}\right)_T$$

However, if the compression process is reversible, it is called the isentropic compressibility and can be written as

$$\tau_s = \frac{1}{\rho}\left(\frac{\partial\rho}{\partial p}\right)_s$$

Gases in general have high compressibility (τ_T for air is 10^{-5} m^2/N at 1 atm) as compared with liquids (τ_T for water is 5×10^{-10} m^2/N at 1 atm).

Compressibility is a very important parameter in the analysis of compressible flow and is closely related to the *speed of sound*, a, which is the velocity of propagation of small pressure disturbances and is defined as

$$a^2 = \left(\frac{\partial p}{\partial\rho}\right)_s \quad \text{or} \quad a = \sqrt{\left(\frac{\partial p}{\partial\rho}\right)_s}$$

In an isentropic process of a perfect gas, the pressure and density are related as

$$\frac{p}{\rho^\gamma} = \text{constant}$$

Using this relation along with the perfect gas relation $p = \rho RT$, we can show that for a perfect gas

$$a = \sqrt{\gamma RT} = \sqrt{\frac{\gamma p}{\rho}}$$

where γ is the ratio of specific heats at constant pressure and constant volume, R is the gas constant, and T is the temperature. For air under normal conditions, γ is 1.4 and R is 287 m^2/sec^2 K so that the speed of sound for air becomes $a = 20.045\sqrt{T}$ m/sec where T is in kelvin.

Another important parameter in compressible flows is the *Mach number, M,* which is defined as the ratio of the gas velocity to the speed of sound or

$$M = \frac{V}{a}$$

where V is the velocity of gas. Depending upon the Mach number of the flow, we can define the following flow regimes:

$M \ll 1$ Incompressible flow

$M < 1$ Subsonic flow

$M \approx 1$ Transonic flow

$M > 1$ Supersonic flow

$M \gg 1$ Hypersonic flow

Subsonic through hypersonic flows are compressible in nature. In these flows, the velocity is appreciable compared with the speed of sound, and the fractional changes in pressure, temperature, and density are all of significant magnitude. We will restrict ourselves in this section to subsonic through flows only.

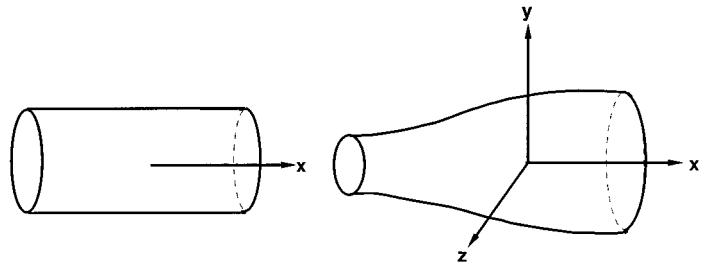
Before we move on to study these flows, let us define one more term. Let us consider a gas with static pressure p and temperature T, traveling at some velocity V and corresponding Mach number M. If this gas is brought isentropically to stagnation or zero velocity, the pressure and temperature which the gas achieves are defined as *stagnation pressure* p_0 and *stagnation temperature* T_0 (also called total pressure and total temperature). The speed of sound at stagnation conditions is called the *stagnation speed of sound* and is denoted as a_0.

One-Dimensional Flow

In one-dimensional flow, the flow properties vary only in one coordinate direction. Figure 3.7.1 shows two streamtubes in a flow. In a *truly one-dimensional flow* illustrated in Figure 3.7.1(a), the flow variables are a function of x only and the area of the stream tube is constant. On the other hand, Figure 3.7.1(b) shows a flow where the area of the stream tube is also a function of x but the flow variables are still a function of x only. This flow is defined as the *quasi-one-dimensional flow.* We will first discuss the truly one-dimensional flow.

In a steady, truly one-dimensional flow, conservation of mass, momentum, and energy leads to the following simple algebraic equations.

$$\begin{aligned} \rho u &= \text{constant} \\ p + \rho u^2 &= \text{constant} \\ h + \frac{u^2}{2} + q &= \text{constant} \end{aligned} \qquad (3.7.1)$$



a b

FIGURE 3.7.1 (a) One-dimensional flow; (b) quasi-one-dimensional flow.

where q is the heat added per unit mass of the gas. These equations neglect body forces, viscous stresses, and heat transfer due to thermal conduction and diffusion. These relations given by Equation 3.7.1, when applied at points 1 and 2 in a flow with no heat addition, become

$$\begin{aligned} \rho_1 u_1 &= \rho_2 u_2 \\ p_1 + \rho_1 u_1^2 &= p_2 + \rho_2 u_2^2 \\ h_1 + \frac{u_1^2}{2} &= h_2 + \frac{u_2^2}{2} \end{aligned} \tag{3.7.2}$$

The energy equation for a calorically perfect gas, where $h = c_p T$, becomes

$$c_p T_1 + \frac{u_1^2}{2} = c_p T_2 + \frac{u_2^2}{2}$$

Using $c_p = \gamma R/(\gamma - 1)$ and $a^2 = \gamma RT$, the above equation can be written as

$$\frac{a_1^2}{\gamma - 1} + \frac{u_1^2}{2} = \frac{a_2^2}{\gamma - 1} + \frac{u_2^2}{2} \tag{3.7.3}$$

Since Equation (3.7.3) is written for no heat addition, it holds for an adiabatic flow. If the energy equation is applied to the stagnation conditions, it can be written as

$$c_p T + \frac{u^2}{2} = c_p T_0$$

$$\frac{T_0}{T} = 1 + \frac{\gamma - 1}{2} M^2 \tag{3.7.4}$$

It is worth mentioning that in arriving at Equation (3.7.4), only adiabatic flow condition is used whereas stagnation conditions are defined as those where the gas is brought to rest isentropically. Therefore, the definition of stagnation temperature is less restrictive than the general definition of stagnation conditions. According to the general definition of isentropic flow, it is a reversible adiabatic flow. This definition is needed for the definition of stagnation pressure and density. For an isentropic flow,

$$\frac{p_0}{p} = \left(\frac{\rho_0}{\rho}\right)^{\gamma} = \left(\frac{T_0}{T}\right)^{\gamma/(\gamma-1)} \tag{3.7.5}$$

From Equations 3.7.4 and 3.7.5, we can write

$$\frac{p_0}{p} = \left(1 + \frac{\gamma - 1}{2} M^2\right)^{\gamma/(\gamma-1)} \tag{3.7.6}$$

$$\frac{\rho_0}{\rho} = \left(1 + \frac{\gamma - 1}{2} M^2\right)^{1/(\gamma-1)} \tag{3.7.7}$$

Values of stagnation conditions are tabulated in Anderson (1982) as a function of M for $\gamma = 1.4$.

Normal Shock Wave

A shock wave is a very thin region (of the order of a few molecular mean free paths) across which the static pressure, temperature, and density increase whereas the velocity decreases. If the shock wave is perpendicular to the flow, it is called a *normal shock wave*. The flow is supersonic ahead of the normal shock wave and subsonic behind it. Figure 3.7.2 shows the flow conditions across a normal shock wave which is treated as a discontinuity. Since there is no heat added or removed, the flow across the shock wave is adiabatic. By using Equations 3.7.2 the normal shock equations can be written as

$$\begin{aligned} \rho_1 u_1 &= \rho_2 u_2 \\ p_1 + \rho_1 u_1^2 &= p_2 + \rho_2 u_2^2 \\ h_1 + \frac{u_1^2}{2} &= h_2 + \frac{u_2^2}{2} \end{aligned} \tag{3.7.8}$$

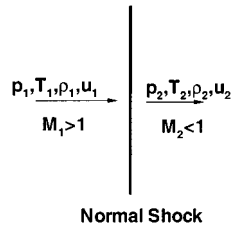


FIGURE 3.7.2 Flow conditions across a normal shock.

Equations (3.7.8) are applicable to a general type of flow; however, for a calorically perfect gas, we can use the relations $p = \rho RT$ and $h = c_p T$ to derive a number of equations relating flow conditions downstream of the normal shock to those at upstream. These equations (also known as Rankine–Hugoniot relations) are

$$\frac{p_2}{p_1} = 1 + \frac{2\gamma}{\gamma + 1}\left(M_1^2 - 1\right)$$

$$\frac{\rho_2}{\rho_1} = \frac{u_1}{u_2} = \frac{(\gamma + 1) M_1^2}{2 + (\gamma - 1) M_1^2} \tag{3.7.9}$$

$$\frac{T_2}{T_1} = \frac{h_2}{h_1} = \left[1 + \frac{2\gamma}{\gamma+1}\left(M_1^2 - 1\right)\right]\left[\frac{2 + (\gamma-1)M_1^2}{(\gamma+1)M_1^2}\right]$$

$$M_2^2 = \frac{1 + \dfrac{\gamma-1}{2}M_1^2}{\gamma M_1^2 - \dfrac{\gamma-1}{2}}$$

Again, the values of p_2/p_1, ρ_2/ρ_1, T_2/T_1, etc. are tabulated in Anderson (1982) as a function of M_1 for $\gamma = 1.4$. Let us examine some limiting cases. As $M_1 \to 1$, Equations 3.7.9 yield $M_2 \to 1$, $p_2/p_1 \to 1$, $\rho_2/\rho_1 \to 1$, and $T_2/T_1 \to 1$. This is the case of an extremely weak normal shock across which no finite changes occur. This is the same as the sound wave. On the other hand, as $M_1 \to \infty$, Equations (3.7.9) yield

$$M_2 \to \sqrt{\frac{\gamma-1}{2\gamma}} = 0.378; \quad \frac{\rho_2}{\rho_1} \to \frac{\gamma+1}{\gamma-1} = 6; \quad \frac{p_2}{p_1} \to \infty; \quad \frac{T_2}{T_1} \to \infty$$

However, the calorically perfect gas assumption no longer remains valid as $M_1 \to \infty$.

Let us now examine why the flow ahead of a normal shock wave must be supersonic even though Equations (3.7.8) hold for $M_1 < 1$ as well as $M_1 > 1$. From the second law of thermodynamics, the entropy change across the normal shock can be written as

$$s_2 - s_1 = c_p \ln\frac{T_2}{T_1} - R\ln\frac{p_2}{p_1}$$

By using Equations (3.7.9) it becomes

$$s_2 - s_1 = c_p \ln\left\{\left[1 + \frac{2\gamma}{\gamma+1}\left(M_1^2 - 1\right)\right]\left[\frac{2 + (\gamma-1)M_1^2}{(\gamma+1)M_1^2}\right]\right\} - R\ln\left[1 + \frac{2\gamma}{\gamma+1}\left(M_1^2 - 1\right)\right] \qquad (3.7.10)$$

Equation (3.7.10) shows that the entropy change across the normal shock is also a function of M_1 only. Using Equation (3.7.10) we see that

$$\begin{aligned} s_2 - s_1 &= 0 \quad \text{for} \quad M_1 = 1 \\ &< 0 \quad \text{for} \quad M_1 < 1 \\ &> 0 \quad \text{for} \quad M_1 > 1 \end{aligned}$$

Since it is necessary that $s_2 - s_1 \geq 0$ from the second law, $M_1 \geq 1$. This, in turn, requires that $p_2/p_1 \geq 1$, $\rho_2/\rho_1 \geq 1$, $T_2/T_1 \geq 1$, and $M_2 \leq 1$.

We now examine how the stagnation conditions change across a normal shock wave. For a calorically perfect gas, the energy equation in Equations (3.7.9) gives

$$c_p T_{01} = c_p T_{02} \quad \text{or} \quad T_{01} = T_{02}$$

In other words, the total temperature remains constant across a stationary normal shock wave.

Let us now apply the entropy change relation across the shock using the stagnation conditions.

$$s_2 - s_1 = c_p \ln \frac{T_{02}}{T_{01}} - R \ln \frac{p_{02}}{p_{01}}$$

Note that entropy at stagnation conditions is the same as at the static conditions since to arrive at stagnation conditions, the gas is brought to rest isentropically. Since $T_{02} = T_{01}$,

$$s_2 - s_1 = -R \ln \frac{p_{02}}{p_{01}}$$

$$\frac{p_{02}}{p_{01}} = e^{-(s_2 - s_1)/R} \tag{3.7.11}$$

Since $s_2 > s_1$ across the normal shockwave, Equation (3.7.11) gives $P_{02} < P_{01}$ or, in other words, the total pressure decreases across a shock wave.

One-Dimensional Flow with Heat Addition

Consider one-dimensional flow through a control volume as shown in Figure 3.7.3. Flow conditions going into this control volume are designated by 1 and coming out by 2. A specified amount of heat per unit mass, q, is added to the control volume. The governing equations relating conditions 1 and 2 can be written as

$$\begin{aligned} \rho_1 u_1 &= \rho_2 u_2 \\ p_1 + \rho_1 u_1^2 &= p_2 + \rho_2 u_2^2 \\ h_1 + \frac{u_1^2}{2} + q &= h_2 + \frac{u_2^2}{2} \end{aligned} \tag{3.7.12}$$

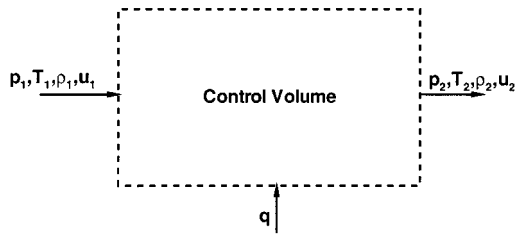


FIGURE 3.7.3 One-dimensional control volume with heat addition.

The following relations can be derived from Equation (3.7.12) for a calorically perfect gas

$$q = c_p \left(T_{02} - T_{01} \right) \tag{3.7.13}$$

$$\frac{p_2}{p_1} = \frac{1 + \gamma M_1^2}{1 + \gamma M_2^2} \tag{3.7.14}$$

$$\frac{T_2}{T_1} = \left(\frac{1+\gamma M_1^2}{1+\gamma M_2^2}\right)^2 \left(\frac{M_2}{M_1}\right)^2 \tag{3.7.15}$$

$$\frac{\rho_2}{\rho_1} = \left(\frac{1+\gamma M_2^2}{1+\gamma M_1^2}\right)^2 \left(\frac{M_1}{M_2}\right)^2 \tag{3.7.16}$$

Equation (3.7.13) indicates that the effect of heat addition is to directly change the stagnation temperature T_0 of the flow. Table 3.7.1 shows some physical trends which can be obtained with heat addition to subsonic and supersonic flow. With heat extraction the trends in Table 3.7.1 are reversed.

TABLE 3.7.1 Effect of Heat Addition on Subsonic and Supersonic Flow

	$M_1 < 1$	$M_1 > 1$
M_2	Increases	Decreases
p_2	Decreases	Increases
T_2	Increases for $M_1 < \gamma^{-1/2}$ and decreases for $M_1 > \gamma^{-1/2}$	Increases
u_2	Increases	Decreases
T_{02}	Increases	Increases
p_{02}	Decreases	Decreases

Figure 3.7.4 shows a plot between enthalpy and entropy, also known as the Mollier diagram, for one-dimensional flow with heat addition. This curve is called the Rayleigh curve and is drawn for a set of given initial conditions. Each point on this curve corresponds to a different amount of heat added or removed. It is seen from this curve that heat addition always drives the Mach numbers toward 1. For a certain amount of heat addition, the flow will become sonic. For this condition, the flow is said to be *choked*. Any further increase in heat addition is not possible without adjustment in initial conditions. For example, if more heat is added in region 1, which is initially supersonic, than allowed for attaining Mach 1 in region 2, then a normal shock will form inside the control volume which will suddenly change the conditions in region 1 to subsonic. Similarly, in case of an initially subsonic flow corresponding to region 1′, any heat addition beyond that is needed to attain Mach 1 in region 2, the conditions in region 1′ will adjust to a lower subsonic Mach number through a series of pressure waves.

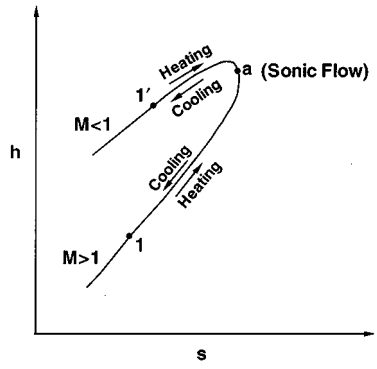


FIGURE 3.7.4 The Rayleigh curve.

Similar to the preceding heat addition or extraction relationships, we can also develop relationships for one-dimensional steady, adiabatic flow but with frictional effects due to viscosity. In this case, the momentum equation gets modified for frictional shear stress. For details, readers are referred to Anderson (1982).

Quasi-One-Dimensional Flow

In quasi-one-dimensional flow, in addition to flow conditions, the area of duct also changes with x. The governing equations for quasi-one-dimensional flow can be written in a differential form as follows using an infinitesimal control volume shown in Figure 3.7.5.

$$d(\rho u A) = 0 \tag{3.7.17}$$

$$dp + \rho u\, du = 0 \tag{3.7.18}$$

$$dh + u\, du = 0 \tag{3.7.19}$$

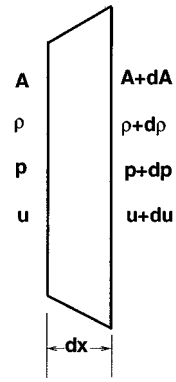


FIGURE 3.7.5 Control volume for quasi-one-dimensional flow.

Equation 3.7.17 can be written as

$$\frac{d\rho}{\rho} + \frac{du}{u} + \frac{dA}{A} = 0 \tag{3.7.20}$$

which can be further written as follows for an isentropic flow:

$$\frac{dA}{A} = \left(M^2 - 1\right)\frac{du}{u} \tag{3.7.21}$$

Some very useful physical insight can be obtained from this area–velocity relation.

- For subsonic flow ($0 \le M < 1$), an increase in area results in decrease in velocity, and vice versa.
- For supersonic flow ($M > 1$), an increase in area results in increase in velocity, and vice versa.
- For sonic flow ($M = 1$), $dA/A = 0$, which corresponds to a minimum or maximum in the area distribution, but it can be shown that a minimum in area is the only physical solution.

Figure 3.7.6 shows the preceding results in a schematic form.

It is obvious from this discussion that for a gas to go isentropically from subsonic to supersonic, and vice versa, it must flow through a convergent–divergent nozzle, also known as the de Laval nozzle. The minimum area of the nozzle at which the flow becomes sonic is called the throat. This physical observation forms the basis of designing supersonic wind tunnels shown schematically in Figure 3.7.7. In general, in a supersonic wind tunnel, a stagnant gas is first expanded to the desired supersonic Mach number. The supersonic flow enters the test section where it passes over a model being tested. The flow

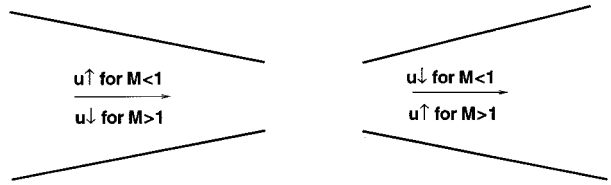


FIGURE 3.7.6 Compressible flow in converging and diverging ducts.

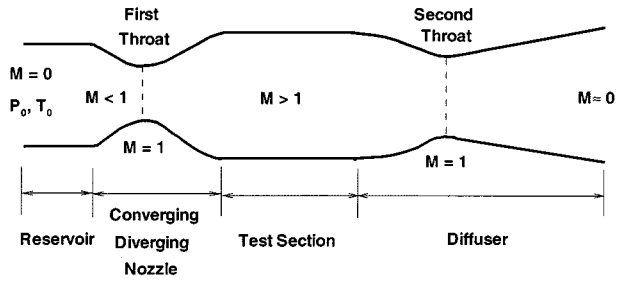


FIGURE 3.7.7 Schematic of a typical supersonic wind tunnel.

then is slowed down by compressing it through a second convergent–divergent nozzle, also known as a diffuser, before it is exhausted to the atmosphere.

Now, using the equations for quasi-one-dimensional flow and the isentropic flow conditions, we can derive a relation for the area ratio that is needed to accelerate or decelerate the gas to sonic conditions. Denoting the sonic conditions by an asterisk, we can write $u^* = a^*$. The area is denoted as A^*, and it is obviously the minimum area for the throat of the nozzle. From Equation (3.7.17) we have

$$\rho u A = \rho^* u^* A^*$$

$$\frac{A}{A^*} = \frac{\rho^* u^*}{\rho u} = \frac{\rho^*}{\rho_0}\frac{\rho_0}{\rho}\frac{u^*}{u} \tag{3.7.22}$$

Under isentropic conditons,

$$\frac{\rho_0}{\rho} = \left(1 + \frac{\gamma - 1}{2} M^2\right)^{1/(\gamma-1)} \tag{3.7.23}$$

$$\frac{\rho_0}{\rho^*} = \left(1 + \frac{\gamma - 1}{2}\right)^{1/(\gamma-1)} = \left(\frac{\gamma + 1}{2}\right)^{1/(\gamma-1)} \tag{3.7.24}$$

Also, $u^*/u = a^*/u$. Let us define a Mach number $M^* = u/a^*$. M^* is known as the *characteristic Mach number* and it is related to the local Mach number by the following relation:

$$M^{*2} = \frac{\dfrac{\gamma + 1}{2} M^2}{1 + \dfrac{\gamma - 1}{2} M^2} \tag{3.7.25}$$

Using Equations (3.7.23) through (3.7.25) in Equation (3.7.22) we can write

$$\left(\frac{A}{A^*}\right)^2 = \frac{1}{M^2}\left[\left(\frac{2}{\gamma+1}\right)\left(1+\frac{\gamma-1}{2}M^2\right)\right]^{(\gamma+1)/(\gamma-1)} \tag{3.7.26}$$

Equation (3.7.26) is called the area Mach number relation. Figure 3.7.8 shows a plot of A/A^* against Mach number. A/A^* is always ≥ 1 for physically viable solutions.

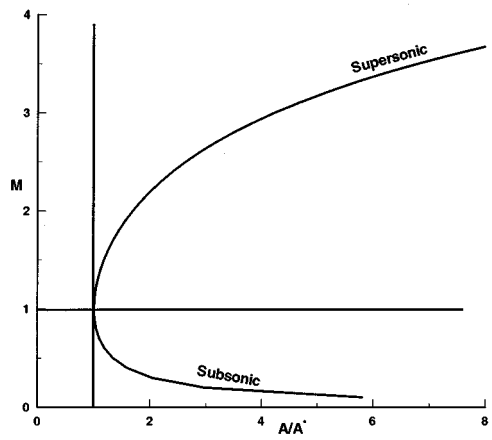


FIGURE 3.7.8 Variation of area ratio A/A^* as a function of Mach number for a quasi-one-dimensional flow.

The area Mach number relation says that for a given Mach number, there is only one area ratio A/A^*. This is a very useful relation and is frequently used to design convergent–divergent nozzles to produce a desired Mach number. Values of A/A^* are tabulated as a function of M in Anderson (1982).

Equation (3.7.26) can also be written in terms of pressure as follows:

$$\frac{A}{A^*} = \frac{\left[1-\left(\frac{p}{p_0}\right)^{(\gamma-1)/\gamma}\right]^{1/2}\left(\frac{p}{p_0}\right)^{1/\gamma}}{\left(\frac{\gamma-1}{2}\right)^{1/2}\left(\frac{2}{\gamma+1}\right)^{(\gamma+1)/2(\gamma-1)}} \tag{3.7.27}$$

Nozzle Flow

Using the area relations, we can now plot the distributions of Mach number and pressure along a nozzle. Figure 3.7.9 shows pressure and Mach number distributions along a given nozzle and the wave configurations for several exit pressures. For curves a and b, the flow stays subsonic throughout and the exit pressure controls the flow in the entire nozzle. On curve c, the throat has just become sonic, and so the pressure at the throat, and upstream of it, can decrease no further. There is another exit pressure corresponding to curve j ($p_j < p_c$) for which a supersonic isentropic solution exists. But if the pressure lies between p_c and p_j, there is no isentropic solution possible. For example, for an exit pressure p_d, a shock will form in the nozzle at location s which will raise the pressure to $p_{d'}$ and turn the flow subsonic. The pressure will then rise to p_d as the subsonic flow goes through an increasing area nozzle. The location, s, depends on the exit pressure. Various possible situations are shown in Figure 3.7.9. It is clear that if the exit pressure is equal to or below p_f, the flow within the nozzle is fully supersonic. This is

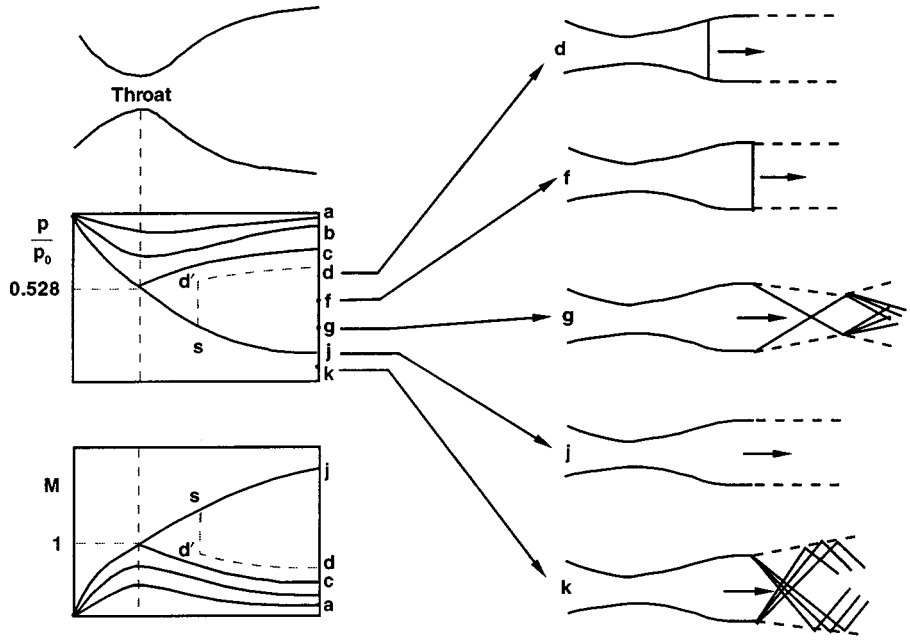


FIGURE 3.7.9 Effect of exit pressure on flow through a nozzle.

the principle used in designing supersonic wind tunnels by operating from a high-pressure reservoir or into a vacuum receiver, or both.

Diffuser

If a nozzle discharges directly into the receiver, the minimum pressure ratio for full supersonic flow in the test section is

$$\left(\frac{p_0}{p_E}\right)_{\min} = \frac{p_0}{p_f}$$

where p_f is the value of p_E at which the normal shock stands right at the nozzle exit. However, by adding an additional diverging section, known as a diffuser, downstream of the test section as shown in Figure 3.7.10 it is possible to operate the tunnel at a lower pressure ratio than p_0/p_f. This happens because the diffuser can now decelerate the subsonic flow downstream of the shock isentropically to a stagnation pressure p_0'. The pressure ratio required then is the ratio of stagnation pressures across a normal shock wave at the test section Mach number. In practice, the diffuser gives lower than expected recovery as a result of viscous losses caused by the interaction of shock wave and the boundary layer which are neglected here.

The operation of supersonic wind tunnels can be made even more efficient; i.e., they can be operated at even lower pressure ratios than p_0/p_0', by using the approach shown in Figure 3.7.7 where the diffuser has a second throat. It can slow down the flow to subsonic Mach numbers isentropically and, ideally, can provide complete recovery, giving $p_0' = p_0$. However, due to other considerations, such as the starting process of the wind tunnel and viscous effects, it is not realized in real life.

Two-Dimensional Supersonic Flow

When supersonic flow goes over a wedge or an expansion corner, it goes through an oblique shock or expansion waves, respectively, to adjust to the change in surface geometry. Figure 3.7.11 shows the two

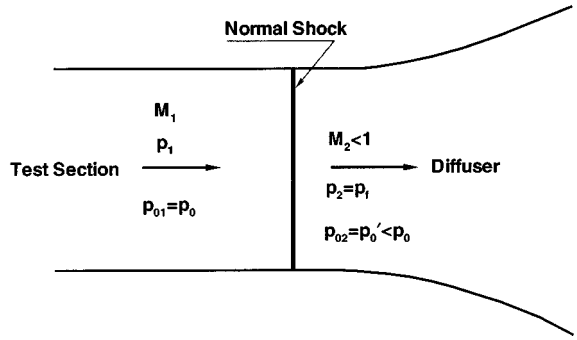


FIGURE 3.7.10 Normal shock diffuser.

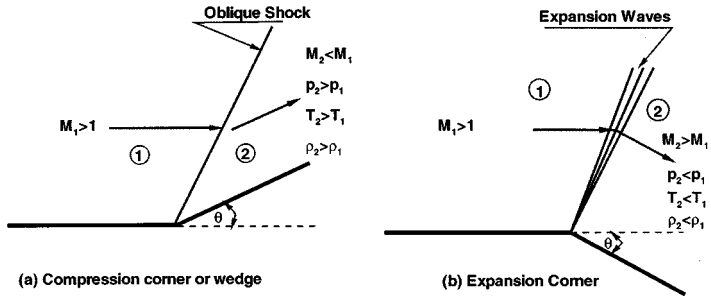


FIGURE 3.7.11 Supersonic flow over a corner.

flow situations. In Figure 3.7.11(a) an oblique shock abruptly turns the flow parallel to the wedge surface. The Mach number behind the shock is less than ahead of it, whereas the pressure, temperature, and density increase. In the case of an expansion corner, oblique expansion waves smoothly turn the flow to become parallel to the surface downstream of the expansion corner. In this case, the Mach number increases, but the pressure, temperature, and density decrease as the flow goes through the expansion corner. Oblique shocks and expansion waves occur in two- and three-dimensional supersonic flows. In this section, we will restrict ourselves to steady, two-dimensional supersonic flows only.

Oblique Shock Waves

The oblique shock can be treated in the same way as the normal shock by accounting for the additional velocity component. If a uniform velocity v is superimposed on the flow field of the normal shock, the resultant velocity ahead of the shock can be adjusted to any flow direction by adjusting the magnitude and direction of v. If v is taken parallel to the shock wave, as shown in Figure 3.7.12, the resultant velocity ahead of the shock is $w_1 = \sqrt{u_1^2 + v_1^2}$ and its direction from the shock is given by $\beta = \tan^{-1}(u_1/v)$. On the downstream side of the shock, since u_2 is less than u_1, the flow always turns toward the shock. The magnitude of u_2 can be determined by the normal shock relations corresponding to velocity u_1 and the magnitude of v is such that the flow downstream of the shock turns parallel to the surface. Since imposition of a uniform velocity does not affect the pressure, temperature, etc., we can use normal shock relations with Mach number replaced in them to correspond to velocity u_1 or u_1/a_1, which is nothing but $M_1 \sin \beta$. Thus, oblique shock relations become

$$\frac{p_2}{p_1} = 1 + \frac{2\gamma}{\gamma + 1}\left(M_1^2 \sin^2 \beta - 1\right) \tag{3.7.28}$$

$$\frac{\rho_2}{\rho_1} = \frac{(\gamma+1)M_1^2 \sin^2\beta}{(\gamma-1)M_1^2 \sin^2\beta + 2} \tag{3.7.29}$$

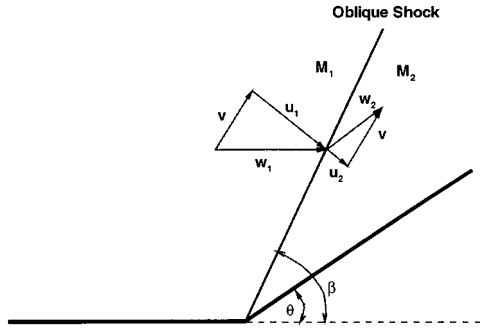


FIGURE 3.7.12 Oblique shock on a wedge.

$$\frac{T_2}{T_1} = \frac{a_2^2}{a_1^2} = \left[1 + \frac{2\gamma}{\gamma+1}\left(M_1^2 \sin^2\beta - 1\right)\right]\left[\frac{2 + (\gamma-1)M_1^2 \sin^2\beta}{(\gamma+1)M_1^2 \sin^2\beta}\right] \tag{3.7.30}$$

The Mach number M_2 (= w_2/a_2) can be obtained by using a Mach number corresponding to velocity u_2 (= $w_2 \sin(\beta - \theta)$) in the normal shock relation for the Mach number. In other words,

$$M_2^2 \sin^2(\beta-\theta) = \frac{1 + \dfrac{\gamma-1}{2}M_1^2 \sin^2\beta}{\gamma M_1^2 \sin^2\beta - \dfrac{\gamma-1}{2}} \tag{3.7.31}$$

To derive a relation between the wedge angle θ and the wave angle β, we have from Figure 3.7.12

$$\tan\beta = \frac{u_1}{v} \quad \text{and} \quad \tan(\beta-\theta) = \frac{u_2}{v}$$

so that

$$\frac{\tan(\beta-\theta)}{\tan\beta} = \frac{u_2}{u_1} = \frac{\rho_1}{\rho_2} = \frac{(\gamma-1)M_1^2 \sin^2\beta + 2}{(\gamma+1)M_1^2 \sin^2\beta}$$

This can be simplified to

$$\tan\theta = 2\cot\beta \frac{M_1^2 \sin^2\beta - 1}{M_1^2(\gamma + \cos 2\beta) + 2} \tag{3.7.32}$$

Dennard and Spencer (1964) have tabulated oblique shock properties as a function of M_1. Let us now make some observations from the preceding relations.

From the normal shock relations, $M_1 \sin\beta \geq 1$. This defines a minimum wave angle for a given Mach number. The maximum wave angle, of course, corresponds to the normal shock or $\beta = \pi/2$. Therefore, the wave angle β has the following range

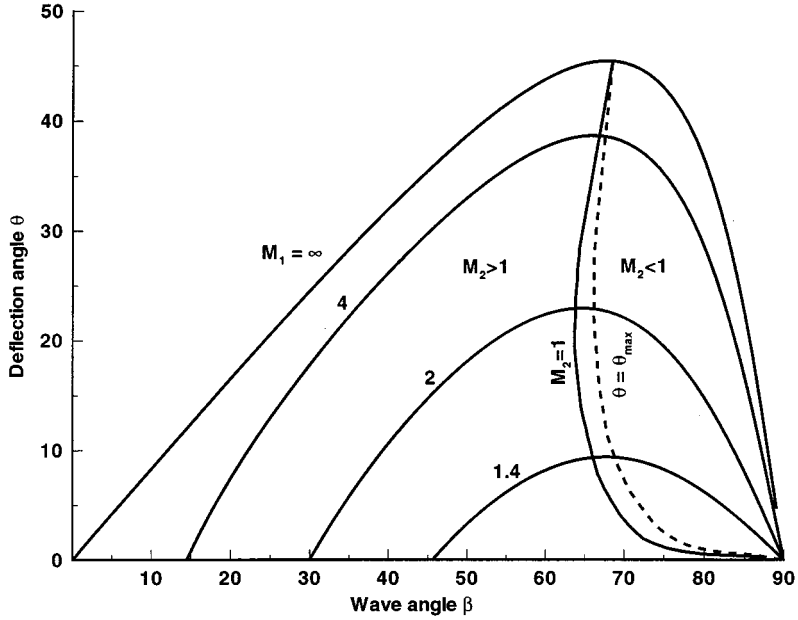


FIGURE 3.7.13 Oblique shock characteristics.

$$\sin^{-1}\frac{1}{M} \le \beta \le \frac{\pi}{2} \tag{3.7.33}$$

Equation 3.7.32 becomes zero at the two limits of β. Figure 3.7.13 shows a plot of θ against β for various values of M_1. For each value of M_1, there is a maximum value of θ. For $\theta < \theta_{max}$, there are two possible solutions having different values of β. The larger value of β gives the stronger shock in which the flow becomes subsonic. A locus of solutions for which $M_2 = 1$ is also shown in the figure. It is seen from the figure that with weak shock solution, the flow remains supersonic except for a small range of θ slightly smaller than θ_{max}.

Let us now consider the limiting case of θ going to zero for the weak shock solution. As θ decreases to zero, β decreases to the limiting value μ, given by

$$M_1^2 \sin^2 \mu - 1 = 0$$

$$\mu = \sin^{-1}\frac{1}{M_1} \tag{3.7.34}$$

For this angle, the oblique shock relations show no jump in flow quantities across the wave or, in other words, there is no disturbance generated in the flow. This angle μ is called the *Mach angle* and the lines at inclination μ are called *Mach lines*.

Thin-Airfoil Theory

For a small deflection angle Δθ, it can be shown that the change in pressure in a flow at Mach M_1 is given approximately by

$$\frac{\Delta p}{p_1} \approx \frac{\gamma M_1^2}{\sqrt{M_1^2 - 1}}\Delta\theta \tag{3.7.35}$$

This expression holds for both compression and expansion. If Δp is measured with respect to the freestream pressure, p_1, and all deflections to the freestream direction, we can write Equation (3.7.35) as

$$\frac{p - p_1}{p_1} = \frac{\gamma M_1^2}{\sqrt{M_1^2 - 1}}\theta \tag{3.7.36}$$

where θ is positive for a compression and negative for expansion. Let us define a pressure coefficient C_p, as

$$C_p = \frac{p - p_1}{q_1}$$

where q_1 is the dynamic pressure and is equal to $\gamma p_1 M_1^2/2$. Equation (3.7.36) then gives

$$C_p = \frac{2\theta}{\sqrt{M_1^2 - 1}} \tag{3.7.37}$$

Equation (3.7.37) states that the pressure coefficient is proportional to the local flow deflection. This relation can be used to develop supersonic thin-airfoil theory. As an example, for a flat plate at angle of attack α_0 (shown in Figure 3.7.14), the pressure coefficients on the upper and lower surfaces are

$$C_p = \frac{2\alpha_0}{\sqrt{M_1^2 - 1}}$$

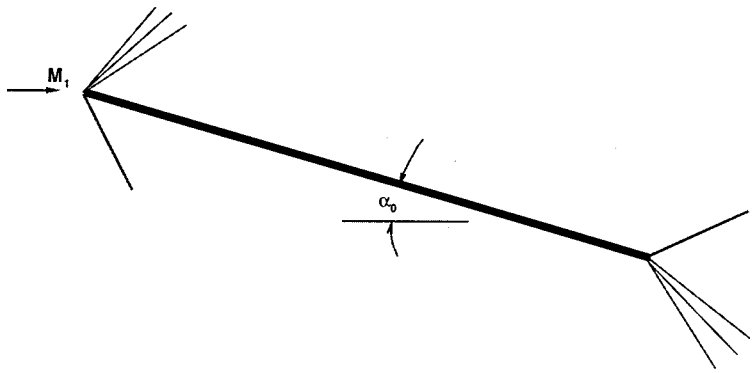


FIGURE 3.7.14 Lifting flat plate.

The lift and drag coefficients can be written as

$$C_L = \frac{(p_L - p_U)c\cos\alpha_0}{q_1 c} = \left(C_{p_L} - C_{p_U}\right)\cos\alpha_0$$

$$C_D = \frac{(p_L - p_U)c\sin\alpha_0}{q_1 c} = \left(C_{p_L} - C_{p_U}\right)\sin\alpha_0$$

where c is the chord length of the plate. Since α_0 is small, we can write

$$C_L = \frac{4\alpha_0}{\sqrt{M_1^2 - 1}}, \qquad C_D = \frac{4\alpha_0^2}{\sqrt{M_1^2 - 1}} \tag{3.7.38}$$

A similar type of expression can be obtained for an arbitrary thin airfoil that has thickness, camber, and angle of attack. Figure 3.7.15 shows such an airfoil. The pressure coefficients on the upper and lower surfaces can be written as

$$C_{p_U} = \frac{2}{\sqrt{M_1^2 - 1}} \frac{dy_U}{dx}, \qquad C_{p_L} = \frac{2}{\sqrt{M_1^2 - 1}} \left(-\frac{dy_L}{dx} \right) \tag{3.7.39}$$

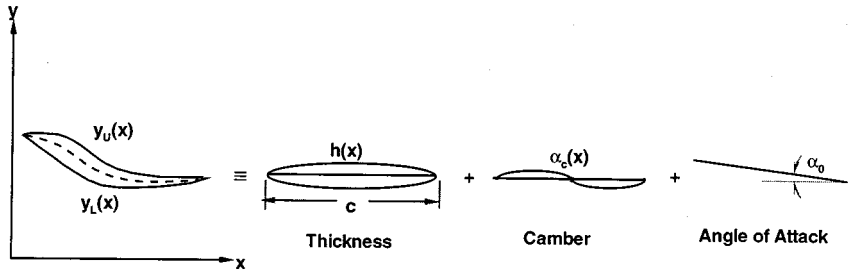


FIGURE 3.7.15 Arbitrary thin airfoil and its components.

For the thin airfoil, the profile may be resolved into three separate components as shown in Figure 3.7.15. The local slope of the airfoil can be obtained by superimposing the local slopes of the three components as

$$\frac{dy_U}{dx} = -\left(\alpha_0 + \alpha_c(x)\right) + \frac{dh}{dx} = -\alpha(x) + \frac{dh}{dx}$$

$$\frac{dy_L}{dx} = -\left(\alpha_0 + \alpha_c(x)\right) - \frac{dh}{dx} = -\alpha(x) - \frac{dh}{dx} \tag{3.7.40}$$

where $\alpha = \alpha_0 + \alpha_c(x)$ is the local total angle of attack of the camber line. The lift and drag for the thin airfoil are given by

$$L = q_1 \int_0^c \left(C_{p_L} - C_{p_U} \right) dx$$

$$D = q_1 \int_0^c \left[C_{p_L} \left(-\frac{dy_L}{dx} \right) + C_{p_U} \left(\frac{dy_U}{dx} \right) \right] dx$$

Let us define an average value of $\alpha(x)$ as

$$\bar{\alpha} = \frac{1}{c} \int_0^c \alpha(x)\, dx$$

Using Equation (3.7.40) and the fact that $\bar{\alpha}_0 = \alpha$ and $\bar{\alpha}_c = 0$ by definition, the lift and drag coefficients for the thin airfoil can be written as

$$C_L = \frac{4\alpha_0}{\sqrt{M_1^2 - 1}}$$

$$C_D = \frac{4}{\sqrt{M_1^2 - 1}}\left[\overline{\left(\frac{dh}{dx}\right)^2} + \overline{\alpha_C^2(x)} + \alpha_0^2\right] \tag{3.7.41}$$

Equations (3.7.41) show that the lift coefficient depends only on the mean angle of attack whereas the drag coefficient is a linear combination of the drag due to thickness, drag due to camber, and drag due to lift (or mean angle of attack).

References

Anderson, J.D. 1982. *Modern Compressible Flow,* McGraw-Hill, New York.

Dennard, J.S. and Spencer, P.B. 1964. Ideal-Gas Tables for Oblique-Shock Flow Parameters in Air at Mach Numbers from 1.05 to 12.0. NASA TN D-2221.

Liepmann, H.W. and Roshko, A. 1966. *Elements of Gas Dynamics,* John Wiley & Sons, New York.

Further Information

As mentioned in the beginning, this section discussed only one- or two-dimensional steady, inviscid compressible flows under perfect gas assumption. Even this discussion was quite brief because of space limitations. For more details on the subject as well as for compressible unsteady viscous flows, readers are referred to Anderson (1982) and Liepmann and Roshko (1966).

3.8 Multiphase Flow

John C. Chen

Introduction

Classic study of fluid mechanics concentrates on the flow of a single homogeneous phase, e.g., water, air, steam. However, many industrially important processes involve simultaneous flow of multiple phases, e.g., gas bubbles in oil, wet steam, dispersed particles in gas or liquid. Examples include vapor–liquid flow in refrigeration systems, steam–water flows in boilers and condensers, vapor–liquid flows in distillation columns, and pneumatic transport of solid particulates. In spite of their importance, multiphase flows are often neglected in standard textbooks. Fundamental understanding and engineering design procedures for multiphase flows are not nearly so well developed as those for single-phase flows. An added complexity is the need to predict the relative concentrations of the different phases in the multiphase flows, a need that doesn't exist for single-phase flows.

Inadequate understanding not withstanding, a significant amount of data have been collected and combinations of theoretical models and empirical correlations are used in engineering calculations. This knowledge base is briefly summarized in this section and references are provided for additional information. While discussions are provided of solid–gas flows and solid–liquid flows, primary emphasis is placed on multiphase flow of gas–liquids since this is the most often encountered class of multiphase flows in industrial applications.

A multiphase flow occurs whenever two or more of the following phases occur simultaneously: gas/vapor, solids, single-liquid phase, multiple (immiscible) liquid phases. Every possible combination has been encountered in some industrial process, the most common being the simultaneous flow of vapor/gas and liquid (as encountered in boilers and condensers). All multiphase flow problems have features which are characteristically different from those found in single-phase problems. First, the relative concentration of different phases is usually a dependent parameter of great importance in multiphase flows, while it is a parameter of no consequence in single-phase flows. Second, the spatial distribution of the various phases in the flow channel strongly affects the flow behavior, again a parameter that is of no concern in single-phase flows. Finally, since the density of various phases can differ by orders of magnitude, the influence of gravitational body force on multiphase flows is of much greater importance than in the case of single-phase flows. In any given flow situation, the possibility exists for the various phases to assume different velocities, leading to the phenomena of slip between phases and consequent interfacial momentum transfer. Of course, the complexity of laminar/turbulent characteristics occurs in multiphase flows as in single-phase flows, with the added complexity of interactions between phases altering the laminar/turbulent flow structures. These complexities increase exponentially with the number of phases encountered in the multiphase problem. Fortunately, a large number of applications occur with just two phase flows, or can be treated as pseudo-two-phase flows.

Two types of analysis are used to deal with two-phase flows. The simpler approach utilizes homogeneous models which assume that the separate phases flow with the same identical local velocity at all points in the fluid. The second approach recognizes the possibility that the two phases can flow at different velocities throughout the fluid, thereby requiring separate conservation equations for mass and momentum for each phase. Brief descriptions of both classes of models are given below.

Fundamentals

Consider n phases in concurrent flow through a duct with cross-sectional area A_c. Fundamental quantities that characterize this flow are

$\dot{m}_i$ = mass flow rate of *i*th phase

u_i = velocity of *i*th phase

α_i = volume fraction of *i*th phase in channel

Basic relationships between these and related parameters are

$$G_i = \text{mass flux of } i\text{th phase} = \frac{\dot{m}_i}{A_c} \tag{3.8.1}$$

$$v_i = \text{superficial velocity of } i\text{th phase} = \frac{G_i}{\rho_i} \tag{3.8.2}$$

$$u_i = \text{actual velocity of } i\text{th phase} = \frac{v_i}{\alpha_i} \tag{3.8.3}$$

$$x_i = \text{flow quality of } i\text{th phase} = \frac{\dot{m}_i}{\sum_i^n m_i} = \frac{G_i}{\sum_{i=1}^n G_i} \tag{3.8.4}$$

$$\alpha_i = \text{volume fraction of } i\text{th phase} = \frac{\left(\frac{x_i}{\rho_i u_i}\right)}{\sum_{i=1}^n \left(\frac{x_i}{\rho_i u_i}\right)} \tag{3.8.5}$$

In most engineering calculations, the above parameters are defined as average quantities across the entire flow area, A_c. It should be noted, however, that details of the multiphase flow could involve local variations across the flow area. In the latter situation, G_i, v_i, and α_i are often defined on a local basis, varying with transverse position across the flow area.

Pressure drop along the flow channel is associated with gravitational body force, acceleration forces, and frictional shear at the channel wall. The total pressure gradient along the flow axis can be represented as

$$\frac{dP}{dz} = \left(\frac{dP}{dz}\right)_g + \left(\frac{dP}{dz}\right)_a + \left(\frac{dP}{dz}\right)_f \tag{3.8.6}$$

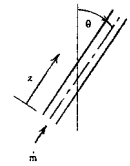
where

$$\left(\frac{dP}{dz}\right)_g = -g\cos\theta \cdot \sum_{i=1}^{n} \alpha_i \rho_i \tag{3.8.7}$$

θ = angle of channel from vertical

and

$$\left(\frac{dP}{dz}\right)_a = -\sum_{i=1}^{n} G_i \frac{du_i}{dz} \tag{3.8.8}$$



$$\left(\frac{dP}{dz}\right)_f = -\frac{\rho u^2}{2D} f \tag{3.8.9}$$

ρ = density of multiphase mixture

$$= \sum_{i=1}^{n} \rho_i \alpha_i \tag{3.8.10}$$

u = an average mixture velocity

$$= \frac{1}{\rho} \sum_{i=1}^{n} G_i \tag{3.8.11}$$

f = equivalent Darcy friction factor for the multiphase flow

In applications, the usual requirement is to determine pressure gradient (dP/dz) and the volume fractions (α_i). The latter quantities are of particular importance since the volume fraction of individual phases affects all three components of the pressure gradient, as indicated in Equations (3.8.7) to (3.8.11). Correlations of various types have been developed for prediction of the volume fractions, all but the simplest of which utilize empirical parameters and functions.

The simplest flow model is known as the homogeneous equilibrium model (HEM), wherein all phases are assumed to be in neutral equilibrium. One consequence of this assumption is that individual phase velocities are equal for all phases everywhere in the flow system:

$$u_i = u \quad \text{for all} \quad i \tag{3.8.12}$$

This assumption permits direct calculation of the volume fractions from known mass qualities:

$$\alpha_i = \frac{x_i}{\rho_i \sum_{i=1}^{n} \left(\frac{x_i}{\rho_i}\right)} \tag{3.8.13}$$

The uniform velocity for all phases is the same as mixture velocity:

$$u = \frac{1}{\rho}\sum_{i=1}^{n} G_i \tag{3.8.14}$$

where

$$\frac{1}{\rho} = \sum_{i=1}^{n}\left(\frac{x_i}{\rho_i}\right) \tag{3.8.15}$$

This homogeneous model permits direct evaluation of all three components of axial pressure gradient, if flow qualities (x_i) are known:

$$\left(\frac{dP}{dz}\right)_g = -\frac{g\cos\theta}{\sum_{i=1}^{n}\left(\frac{x_i}{\rho_i}\right)} \tag{3.8.16}$$

$$\left(\frac{dP}{dz}\right)_a = -\left(\sum_{i=1}^{n} G_i\right)\cdot\frac{du}{dz} \tag{3.8.17}$$

$$\left(\frac{dP}{dz}\right)_f = -\frac{\rho u^2}{2D_f}\cdot f \tag{3.8.18}$$

where u and ρ are given by Equations (3.8.14) and (3.8.15).

Predicting the coefficient of friction (f to clear) remains a problem, even in the homogeneous model. For cases of fully turbulent flows, experience has shown that a value of 0.02 may be used as a first-order approximation for (f to clear). More-accurate estimates require empirical correlations, specific to particular classes of multiphase flows and subcategories of flow regimes.

The following parts of this section consider the more common situations of two-phase flows and describe improved design methodologies specific to individual situations.

Gas–Liquid Two-Phase Flow

The most common case of multiphase flow is two-phase flow of gas and liquid, as encountered in steam generators and refrigeration systems. A great deal has been learned about such flows, including delineation of flow patterns in different flow regimes, methods for estimating volume fractions (gas void fractions), and two-phase pressure drops.

Flow Regimes

A special feature of multiphase flows is their ability to assume different spatial distributions of the phases. These different flow patterns have been classified in flow regimes, which are themselves altered by the direction of flow relative to gravitational acceleration. Figures 3.8.1 and 3.8.2 (Delhaye, 1981) show the flow patterns commonly observed for co-current flow of gas and liquid in vertical and horizontal channels, respectively. For a constant liquid flow rate, the gas phase tends to be distributed as small bubbles at low gas flow rates. Increasing gas flow rate causes agglomeration of bubbles into larger slugs and plugs. Further increasing gas flow rate causes separation of the phases into annular patterns wherein liquid concentrates at the channel wall and gas flows in the central core for vertical ducts. For horizontal ducts, gravitational force tends to drain the liquid annulus toward the bottom of the channel, resulting

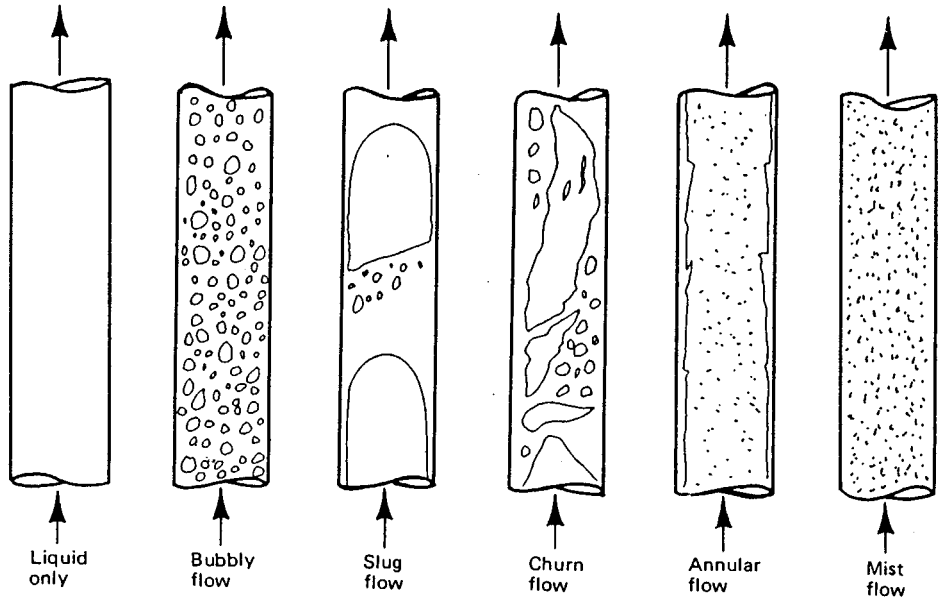


FIGURE 3.8.1 Flow patterns in gas–liquid vertical flow. (From Lahey, R.T., Jr. and Moody, F.I. 1977. *The Thermal Hydraulics of a Boiling Water Nuclear Reactor,* The American Nuclear Society, LaGrange, IL. With permission.)

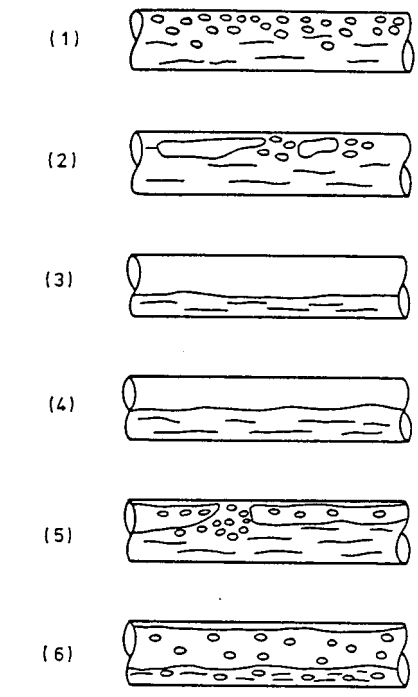


FIGURE 3.8.2 Flow patterns in gas–liquid horizontal flow.

in stratified and stratified wavy flows. This downward segregation of the liquid phase can be overcome by kinetic forces at high flow rates, causing stratified flows to revert to annular flows. At high gas flow rates, more of the liquid tends to be entrained as dispersed drops; in the limit one obtains completely dispersed mist flow.

Flow pattern maps are utilized to predict flow regimes for specific applications. The first generally successful flow map was that of Baker (1954) for horizontal flow, reproduced here in Figure 3.8.3. For vertical flows, the map of Hewitt and Roberts (1969), duplicated in Figure 3.8.4, provides a simple method for determining flow regimes. Parameters used for the axial coordinates of these flow maps are defined as follows:

$$\lambda = \left(\frac{\rho_g \rho_\ell}{\rho_a \rho_w} \right)^{1/2} \tag{3.8.19}$$

$$\psi = \left(\frac{\sigma_w}{\sigma} \right) \left[\left(\frac{\mu_\ell}{\mu_w} \right) \left(\frac{\rho_w}{\rho_\ell} \right)^2 \right]^{1/3} \tag{3.8.20}$$

$$j = \text{volumetric flux}, \frac{G}{\rho} \tag{3.8.21}$$

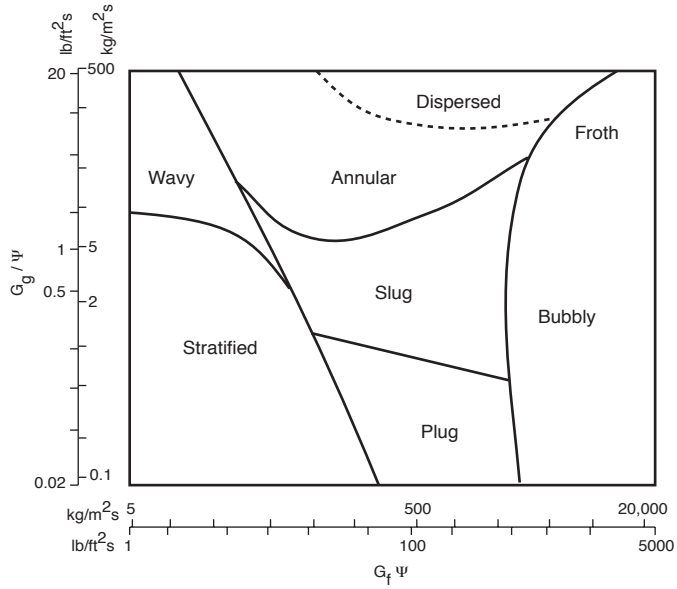


FIGURE 3.8.3 Flow pattern map for horizontal flow (Baker, 1954).

Void Fractions

In applications of gas–liquid flows, the volume fraction of gas (α_g) is commonly called "void fraction" and is of particular interest. The simplest method to estimate void fraction is by the HEM. From Equation (3.8.13), the void fraction can be estimated as

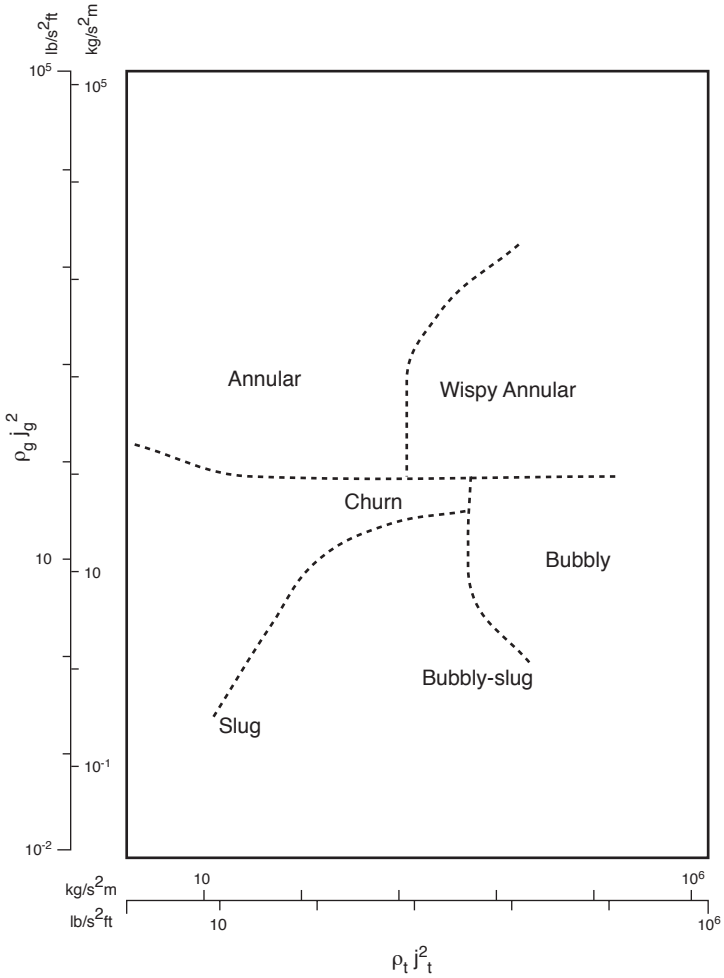


FIGURE 3.8.4 Flow pattern map for vertical flow (Hewitt and Roberts, 1969).

$$\alpha_g = \frac{x_g}{x_g + \left(1 - x_g\right)\dfrac{\rho_g}{\rho_\ell}} \tag{3.8.22}$$

where α_g, x_g, ρ_g, ρ_ℓ are cross-sectional averaged quantities.

In most instances, the homogenous model tends to overestimate the void fraction. Improved estimates are obtained by using separated-phase models which account for the possibility of slip between gas and liquid velocities. A classic separated-phase model is that of Lockhart and Martinelli (1949). The top portion of Figure 3.8.5 reproduces the Lockhart–Martinelli correlation for void fraction (shown as α) as a function of the parameter X which is defined as

$$X = \left[\left(\frac{dP}{dz}\right)_{f\ell} \div \left(\frac{dP}{dz}\right)_{fg}\right]^{1/2} \tag{3.8.23}$$

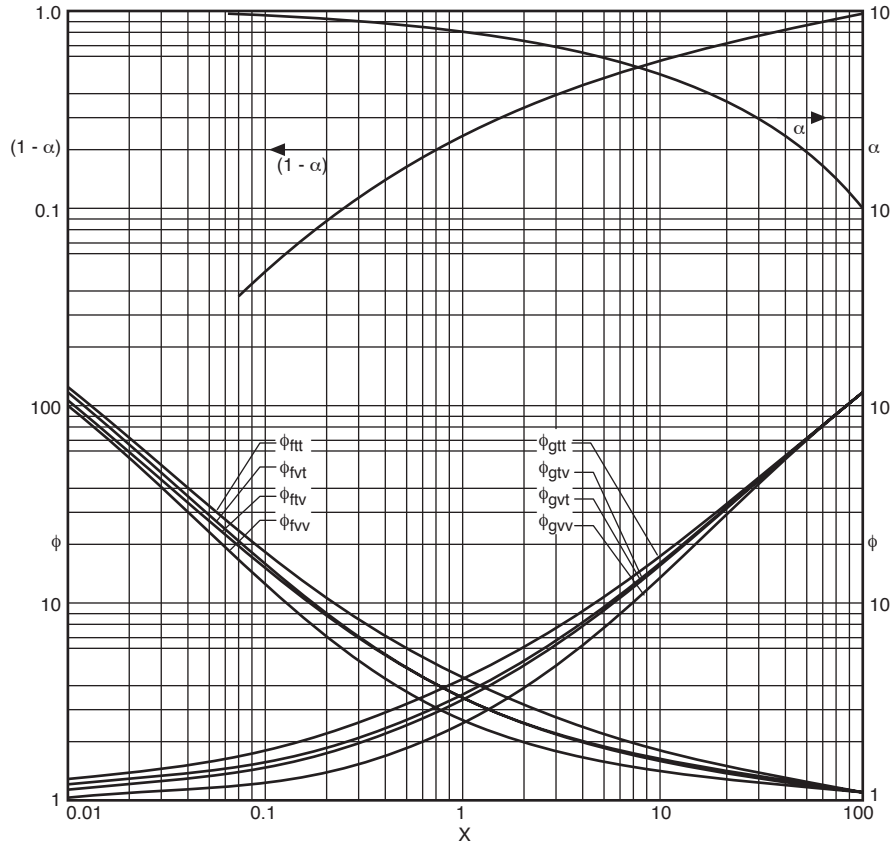


FIGURE 3.8.5 Correlations for void fraction and frictional pressure drop (Lockhart and Martinelli, 1949)

where

$\left(\frac{dP}{dz}\right)_{f\ell}$ = frictional pressure gradient of liquid phase flowing alone in channel

$\left(\frac{dP}{dz}\right)_{fg}$ = frictional pressure gradient of gas phase flowing alone in channel

Often, flow rates are sufficiently high such that each phase if flowing alone in the channel would be turbulent. In this situation the parameter X can be shown to be

$$X_{tt} = \left(\frac{1 - x_g}{x_g}\right)^{0.9} \left(\frac{\rho_g}{\rho_\ell}\right)^{0.5} \left(\frac{\mu_\ell}{\mu_g}\right)^{0.1} \tag{3.8.24}$$

Another type of separated-phase model is the drift-flux formulation of Wallis (1969). This approach focuses attention on relative slip between phases and results in slightly different expressions depending on the flow regime. For co-current upflow in two of the more common regimes, the drift-flux model gives the following relationships between void fraction and flow quality:

Bubbly flow or churn-turbulent flow:

$$\alpha_g = \frac{x_g}{\left(\frac{u_o \rho_g}{G}\right) + C_o\left[x_g + \left(1 - x_g\right)\frac{\rho_g}{\rho_\ell}\right]} \tag{3.8.25}$$

Dispersed drop (mist) flow:

$$x_g = \frac{1 - \left(1 - \alpha_g\right)\left(\frac{u_o \rho_\ell}{G}\alpha_g^2 + 1\right)}{1 - \left(1 - \alpha_g\right)\left(1 - \frac{\rho_\ell}{\rho_g}\right)} \tag{3.8.26}$$

where u_o = terminal rise velocity of bubble, in bubbly flow, or terminal fall velocity of drop in churn-turbulent flow

C_o = an empirical distribution coefficient $\simeq 1.2$

Pressure Drop

Equations 3.8.16 through 3.8.18 permit calculation of two-phase pressure drop by the homogeneous model, if the friction coefficient (f) is known. One useful method for estimating (f) is to treat the entire two-phase flow as if it were all liquid, except flowing at the two-phase mixture velocity. By this approach the frictional component of the two-phase pressure drop becomes

$$\left(\frac{dP}{dz}\right)_f = \left[1 + x_g\left(\frac{\rho_\ell}{\rho_g} - 1\right)\right] \cdot \left(\frac{dP}{dz}\right)_{f\ell G} \tag{3.8.27}$$

where $(dP/dz)_{f\ell G}$ = frictional pressure gradient if entire flow (of total mass flux G) flowed as liquid in the channel.

The equivalent frictional pressure drop for the entire flow as liquid, $(dP/dz)_{f\ell g}$, can be calculated by standard procedures for single-phase flow. In using Equations (3.8.16) through (3.8.18), the void fraction would be calculated with the equivalent homogeneous expression Equation (3.8.13).

A more accurate method to calculate two-phase pressure drop is by the separated-phases model of Lockhart and Martinelli (1949). The bottom half of Figure 3.8.5 shows empirical curves for the Lockhart–Martinelli frictional multiplier, ϕ:

$$\phi_i = \left[\left(\frac{dP}{dz}\right)_f \div \left(\frac{dP}{dz}\right)_{fi}\right]^{1/2} \tag{3.8.28}$$

where (i) denotes either the fluid liquid phase (f) or gas phase (g). The single-phase frictional gradient is based on the ith phase flowing alone in the channel, in either viscous laminar (v) or turbulent (t) modes. The most common case is where each phase flowing alone would be turbulent, whence one could use Figure 3.8.5 to obtain

$$\left(\frac{dP}{dz}\right)_f = \text{frictional pressure gradient for two-phase flow}$$
$$= \phi_{gtt}^2 \cdot \left(\frac{dP}{dz}\right)_{fg} \tag{3.8.29}$$

where $(dP/dz)_{fg}$ is calculated for gas phase flowing alone and $X = X_{tt}$ as given by Equation (3.8.24).

The correlation of Lockhart–Martinelli has been found to be adequate for two-phase flows at low-to-moderate pressures, i.e., with reduced pressures less than 0.3. For applications at higher pressures, the revised models of Martinelli and Nelson (1948) and Thom (1964) are recommended.

Gas–Solid, Liquid–Solid Two-Phase Flows

Two-phase flows can occur with solid particles in gas or liquid. Such flows are found in handling of granular materials and heterogeneous reaction processing. Concurrent flow of solid particulates with a fluid phase can occur with various flow patterns, as summarized below.

Flow Regimes

Consider vertical upflow of a fluid (gas or liquid) with solid particles. Figure 3.8.6 illustrates the major flow regimes that have been identified for such two-phase flows. At low flow rates, the fluid phase percolates between stationary particles; this is termed flow through a fixed bed. At some higher velocity a point is reached when the particles are all suspended by the upward flowing fluid, the drag force between particles and fluid counterbalancing the gravitational force on the particles. This is the point of minimum fluidization, marking the transition from fixed to fluidized beds. Increase of fluid flow rate beyond minimum fluidization causes instabilities in the two-phase mixture, and macroscopic bubbles or channels of fluid are observed in the case of gaseous fluids. In the case of liquid fluids, the two-phase mixture tends to expand, often without discrete bubbles or channels. Further increase of fluid velocity causes transition to turbulent fluidization wherein discrete regions of separated phases (fluid slugs or channels and disperse suspensions of particles) can coexist. Depending on specific operating conditions (e.g., superficial fluid velocity, particle size, particle density, etc.), net transport of solid particles with the flowing fluid can occur at any velocity equal to or greater than that associated with slug flow and turbulent flow. Further increases in fluid velocity increase the net transport of solid particles. This can occur with large-scale clusters of solid particles (as exemplified by the fast fluidization regime) or with dilute dispersions of solid particles (as often utilized in pneumatic conveying). For engineering application of fluid–solid two-phase flows, the important thresholds between flow regimes are marked by the fluid velocity for minimum fluidization, terminal slip, and saltation threshold.

Minimum Fluidization

The transition from flow through packed beds to the fluidization regime is marked by the minimum fluidization velocity of the fluid. On a plot pressure drop vs. superficial fluid velocity, the point of minimum fluidization is marked by a transition from a linearly increasing pressure drop to a relatively constant pressure drop as shown in Figure 3.8.7 for typical data, for two-phase flow of gas with sand particles of 280 μm mean diameter (Chen, 1996). The threshold fluid velocity at minimum fluidization is traditionally derived from the Carman–Kozeny equation,

$$U_{mf} = \frac{\left(\rho_s - \rho_f\right)\left(\phi\, dp\right)^2 g}{150\mu_f} \cdot \frac{\alpha_{mf}^2}{\left(1 - \alpha_{mf}\right)} \tag{3.8.30}$$

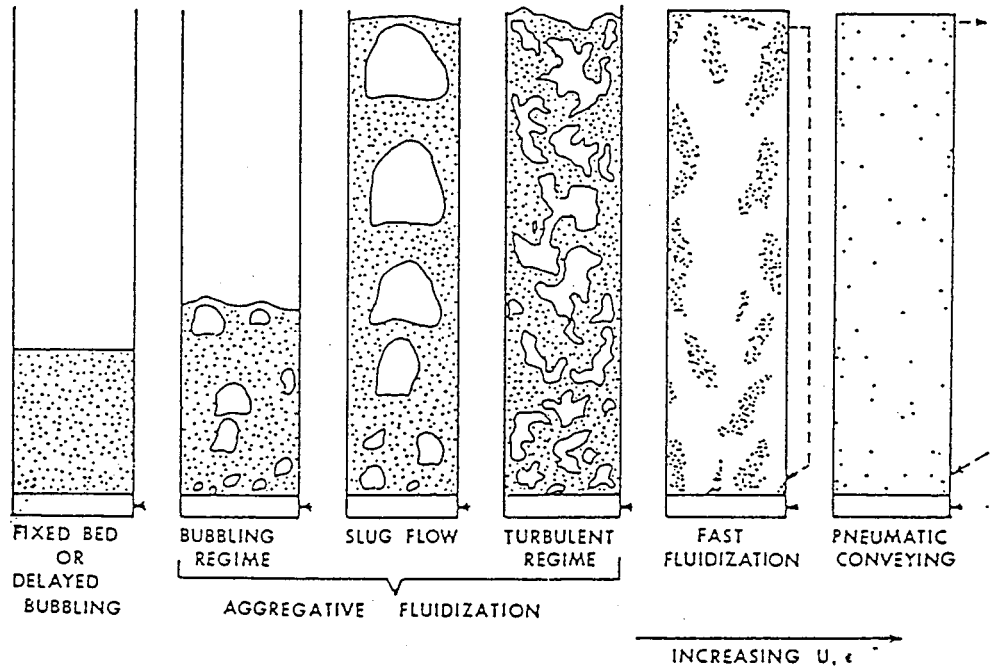


FIGURE 3.8.6 Flow patterns for vertical upflow of solid particles and gas or liquid. (From Chen, J.C. 1994. *Proc. Xth Int. Heat Transfer Conf.,* Brighton, U.K., 1:369–386. With permission.)

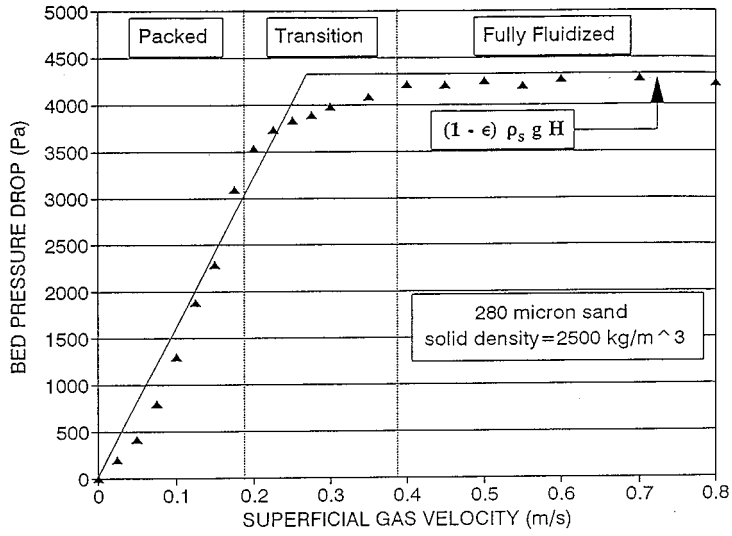


FIGURE 3.8.7 Transition at minimum fluidization. (From Chen, J.C. 1996. In *Annual Review of Heat Transfer,* Vol. VII, Begal House, Washington, D.C. With permission.)

where ϕ = sphericity of particles (unity for spherical particles)

α_{mf} = volumetric fraction of fluid at minimum fluidization

Small, light particles have minimum fluidization voidage (α_{mf}) of the order 0.6, while larger particles such as sand have values closer to 0.4.

An alternative correlation for estimating the point of minimum fluidization is that of Wen and Yu (1966):

$$\frac{U_{mf} d_p \rho_f}{\mu_f} = \left(33.7 + 0.041 Ga\right)^{0.5} - 33.7 \tag{3.8.31}$$

where $Ga = \rho_f d_p^3 (\rho_s - \rho_f) g / \mu_f^2$.

When the fluid velocity exceeds U_{mf}, the two-phase mixture exists in the fluidized state in which the pressure gradient is essentially balanced by the gravitational force on the two-phase mixture:

$$\frac{dP}{dz} = g\left[\alpha_s \rho_s + \alpha_f \rho_f\right] \tag{3.8.32}$$

This fluidized state exists until the fluid velocity reaches a significant fraction of the terminal slip velocity, beyond which significant entrainment and transport of the solid particles occur.

Terminal Slip Velocity

For an isolated single particle the maximum velocity relative to an upflowing fluid is the terminal slip velocity. At this condition, the interfacial drag of the fluid on the particle exactly balances the gravitational body force on the particle:

$$U_t = \left(U_f - U_s\right)_t = \left[\frac{4 d_p \left(\rho_s - \rho_f\right)}{3 \rho_f} \cdot \frac{1}{C_D}\right]^{1/2} \tag{3.8.33}$$

where C_D = coefficient of drag on the particle.

The coefficient of drag on the particle (C_D) depends on the particle Reynolds number:

$$\mathrm{Re}_p = \frac{\rho_f d_p \left(U_f - U_s\right)}{\mu_f} \tag{3.8.34}$$

The following expressions may be used to estimate C_D as appropriate:

$$\begin{aligned} C_D &= \frac{32}{\mathrm{Re}_p}, && \mathrm{Re}_p \le 1 \\ C_D &= \frac{18.5}{\mathrm{Re}_p^{0.67}}, && 1 \le \mathrm{Re}_p \le 10^3 \end{aligned} \tag{3.8.35}$$

Pneumatic Conveying

A desirable mode of pneumatic conveying is two-phase flow with solid particles dispersed in the concurrent flowing fluid. Such dispersed flows can be obtained if the fluid velocity is sufficiently high. For both horizontal and vertical flows, there are minimum fluid velocities below which saltation of the solid particles due to gravitational force occurs, leading to settling of the solid particles in horizontal channels and choking of the particles in vertical channels. Figures 3.8.8 and 3.8.9 for Zenz and Othmer (1960) show these different regimes of pneumatic conveying for horizontal and vertical transport, respectively. Figure 3.8.8 shows that for a given rate of solids flow (W) there is a minimum superficial fluid velocity below which solid particles tend to settle into a dense layer at the bottom of the horizontal channels. Above this saltation threshold, fully dispersed two-phase flow is obtained. In the case of

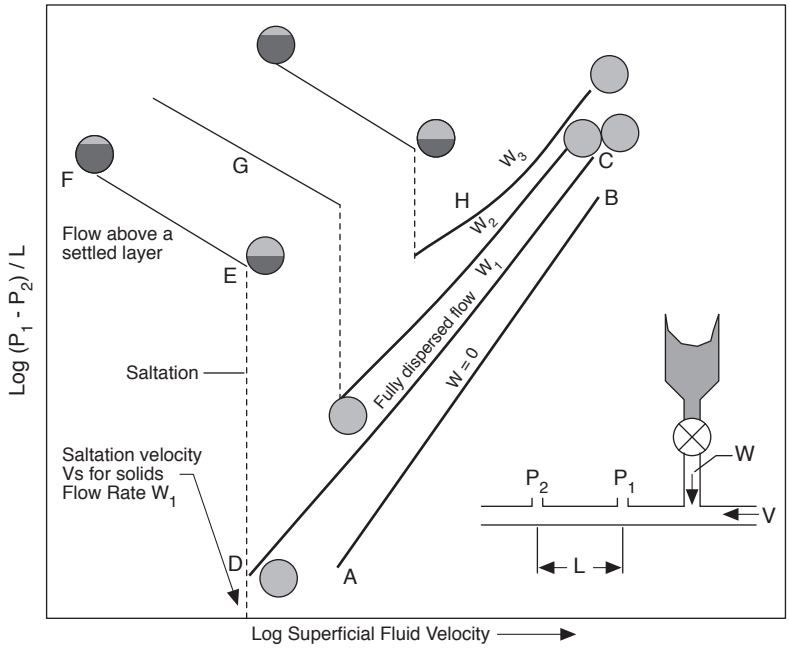


FIGURE 3.8.8 Flow characteristics in horizontal pneumatic conveying.

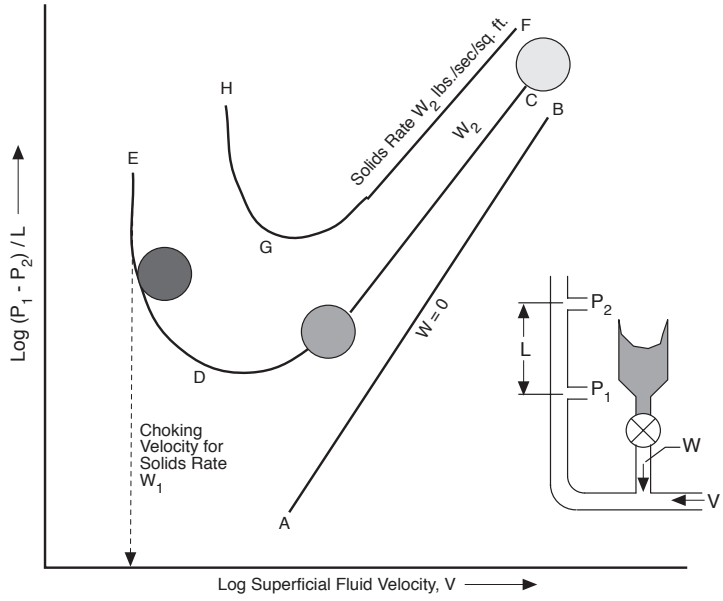


FIGURE 3.8.9 Flow characteristics in vertical pneumatic conveying.

vertical transport illustrated in Figure 3.8.9, there is a minimum fluid velocity below which solid particles tend to detrain from the two-phase suspension. This choking limit varies not only with particle properties but also with the actual rate of particle flow. Well-designed transport systems must operate with superficial fluid velocities greater than these limiting saltation and choking velocities.

Zenz and Othmer (1960) recommend the empirical correlations represented in Figure 3.8.10 estimating limiting superficial fluid velocities at incipient saltation or choking, for liquid or gas transport of uniformly sized particles. Note that these correlations are applicable for either horizontal or vertical concurrent flow. Figure 3.8.10 is duplicated from the original source and is based on parameters in engineering units, as noted in the figure. To operate successfully in dispersed pneumatic conveying of solid particles, the superficial fluid velocity must exceed that determined from the empirical correlations of Figure 3.8.10.

Nomenclature

A_c	cross-sectional flow area of channel
C_o	Wallis' distribution coefficient
d_p	diameter of solid particles
f_D	Darcy friction factor
G	mass flow flux, kg/m^2 · sec
j	volumetric flow flux, m/sec
$\dot{m}$	mass flow rate, kg/sec
P	pressure, N/m^2
u	velocity in axial flow direction, m/sec
v	superficial velocity in axial flow direction, m/sec
x	mass flow quality
z	axial coordinate

Greek Letters

α	volume fraction
λ	parameter in Baker flow map
ϕ	sphericity of solid particles
ϕ_i	frictional multiphase for pressure drag, Equation (3.8.28)
ψ	parameter in Baker flow map
σ	surface tension
θ	angle from vertical

Subscripts

a	air
f	fluid phase
g	gas phase
l	liquid phase
mf	minimum fluidization
p	particle
s	solid phase
t	terminal slip
w	water

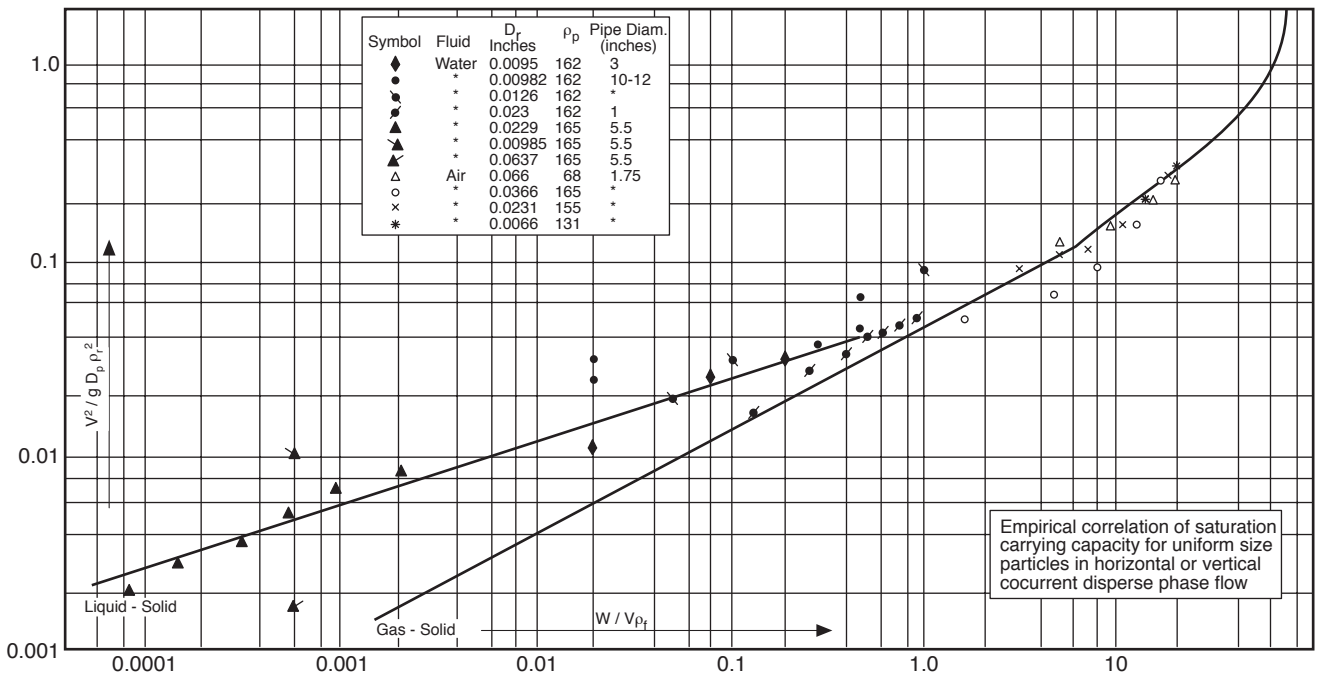


FIGURE 3.8.10 Correlations for limiting velocities in pneumatic conveying. (

References

Baker, O. 1954. Design of pipelines for simultaneous flow of oil and gas, *Oil Gas J.*.

Chen, J.C. 1994. Two-phase flow with and without phase changes: suspension flows. Keynote lecture, *Proc. Xth Int. Heat Transfer Conf.,* Brighton, U.K., 1:369–386.

Chen, J.C. 1996. Heat transfer to immersed surfaces in bubbling fluidized beds, in *Annual Review of Heat Transfer*, Vol. VII, Bengel House, Washington, D.C.

Collier, J.G. 1972. *Convective Boiling and Condensation,* McGraw-Hill, London.

Delhaye, J.M. 1981. Two-phase flow patterns, in *Two-Phase Flow and Heat Transfer,* A.E. Bergles, J.G. Collier, J.M. Delhaye, G.F. Newitt, and F. Mayinger, Eds., Hemisphere Publishing, McGraw-Hill, New York.

Hewitt, G.F. and Roberts, D.N. 1969. Studies of Two-Phase Flow Patterns by Simultaneous X-Ray and Flash Photography, Report AERE-M 2159.

Lahey, R.T., Jr. and Moody, F.I. 1977. *The Thermal Hydraulics of a Boiling Water Nuclear Reactor,* The American Nuclear Society, La Grange, IL.

Lockhart, R.W. and Martinelli, R.C. 1949. Proposed correlation of data for isothermal two-phase two-component flow in pipes, *Chem. Eng. Progr.,* 45:39.

Martinelli, R.C. and Nelson, D.B. 1984. Prediction of pressure drop during forced-circulation boiling of water, *Trans. ASME,* 70:695–702.

Thom, J.R.S. 1964. Prediction of pressure drop during forced circulation boiling of water, *Int. J. Heat Mass Transfer,* 7:709–724.

Wallis, G.B. 1969. *One-Dimensional Two-Phase Flow,* McGraw-Hill, New York.

Wen, C.Y. and Yu, Y.H. 1966. A generalized method of predicting the minimum fluidization velocity, *AIChE J.,* 12:610–612.

Zenz, F.A. and Othmer, D.F. 1960. *Fluidization and Fluid-Particle Systems,* Reinhold, New York.

3.9 New-Newtonian Flows

Thomas F. Irvine Jr. and Massimo Capobianchi

Introduction

An important class of fluids exists which differ from Newtonian fluids in that the relationship between the shear stress and the flow field is more complicated. Such fluids are called non-Newtonian or rheological fluids. Examples include various suspensions such as coal–water or coal–oil slurries, food products, inks, glues, soaps, polymer solutions, etc.

An interesting characteristic of rheological fluids is their large "apparent viscosities". This results in laminar flow situations in many applications, and consequently the engineering literature is concentrated on laminar rather than turbulent flows. It should also be mentioned that knowledge of non-Newtonian fluid mechanics and heat transfer is still in an early stage and many aspects of the field remain to be clarified.

In the following sections, we will discuss the definition and classification of non-Newtonian fluids, the special problems of thermophysical properties, and the prediction of pressure drops in both laminar and turbulent flow in ducts of various cross-sectional shapes for different classes of non-Newtonian fluids.

Classification of Non-Newtonian Fluids

It is useful to first define a Newtonian fluid since all other fluids are non-Newtonian. Newtonian fluids possess a property called viscosity and follow a law analogous to the Hookian relation between the stress applied to a solid and its strain. For a one-dimensional Newtonian fluid flow, the shear stress at a point is proportional to the rate of strain (called in the literature the *shear rate*) which is the velocity gradient at that point. The constant of proportionality is the dynamic viscosity, i.e.,

$$\tau_{y,x} = \mu \frac{du}{dy} = \mu \dot{\gamma} \tag{3.9.1}$$

where x refers to the direction of the shear stress y the direction of the velocity gradient, and $\dot{\gamma}$ is the shear rate. The important characteristic of a Newtonian fluid is that the dynamic viscosity is independent of the shear rate.

Equation (3.9.1) is called a constitutive equation, and if $\tau_{x,y}$ is plotted against $\dot{\gamma}$, the result is a linear relation whose slope is the dynamic viscosity. Such a graph is called a *flow curve* and is a convenient way to illustrate the viscous properties of various types of fluids.

Fluids which do not obey Equation (3.9.1) are called non-Newtonian. Their classifications are illustrated in Figure 3.9.1 where they are separated into various categories of purely viscous time-independent or time-dependent fluids and viscoelastic fluids. Viscoelastic fluids, which from their name possess both viscous and elastic properties (as well as memory), have received considerable attention because of their ability to reduce both drag and heat transfer in channel flows. They will be discussed in a later subsection.

Purely viscous time-dependent fluids are those in which the shear stress in a function only of the shear rate but in a more complicated manner than that described in Equation (3.9.1). Figure 3.9.2 illustrates the characteristics of purely viscous time-independent fluids. In the figure, (a) and (b) are fluids where the shear stress depends only on the shear rate but in a nonlinear way. Fluid (a) is called pseudoplastic (or shear thinning), and fluid (b) is called dilatant (or shear thickening). Curve (c) is one which has an initial yield stress after which it acts as a Newtonian fluid, called Buckingham plastic, and curve (d), called Hershel-Buckley, also has a yield stress after which it becomes pseudoplastic. Curve (e) depicts a Newtonian fluid.

Figure 3.9.3 shows flow curves for two common classes of purely viscous time-dependent non-Newtonian fluids. It is seen that such fluids have a hysteresis loop or memory whose shape depends

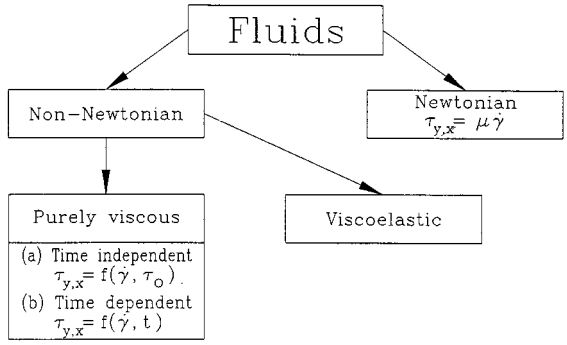


FIGURE 3.9.1 Classification of fluids.

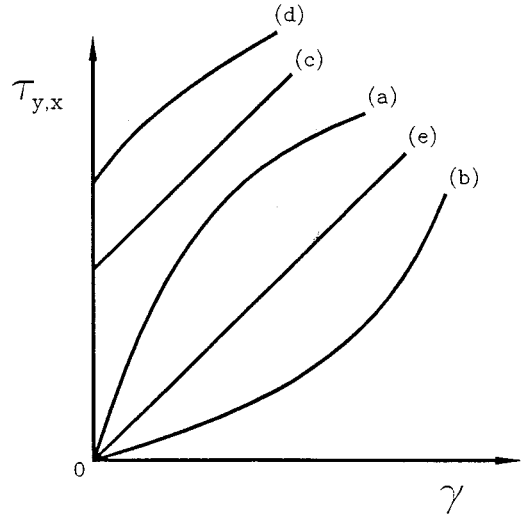


FIGURE 3.9.2 Flow curves of purely viscous, time-independent fluids: (a) pseudoplastic; (b) dilatant; (c) Bingham plastic; (d) Hershel–Buckley; (e) Newtonian.

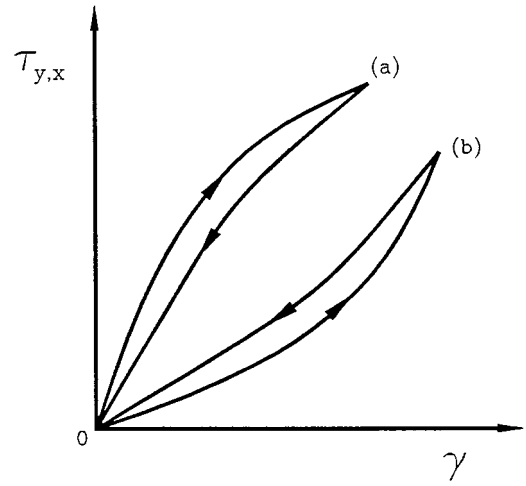


FIGURE 3.9.3 Flow curves for purely viscous, time-dependent fluids: (a) thixotropic; (b) rheopectic.

upon the time-dependent rate at which the shear stress is applied. Curve (a) illustrates a pseudoplastic time-dependent fluid and curve (b) a dilatant time-dependent fluid. They are called, respectively, thixotropic and rheopectic fluids and are complicated by the fact that their flow curves are difficult to characterize for any particular application.

Apparent Viscosity

Although non-Newtonian fluids do not have the property of viscosity, in the Newtonian fluid sense, it is convenient to define an apparent viscosity which is the ratio of the local shear stress to the shear rate at that point.

$$\mu_a = \frac{\tau}{\dot{\gamma}} \tag{3.9.2}$$

The apparent viscosity is not a true property for non-Newtonian fluids because its value depends upon the flow field, or shear rate. Nevertheless, it is a useful quantity and flow curves are often constructed with the apparent viscosity as the ordinate and shear rate as the abscissa. Such a flow curve will be illustrated in a later subsection.

Constitutive Equations

A constitutive equation is one that expresses the relation between the shear stress or apparent viscosity and the shear rate through the rheological properties of the fluid. For example, Equation (3.9.1) is the constitutive equation for a Newtonian fluid.

Many constitutive equations have been developed for non-Newtonian fluids with some of them having as many as five rheological properties. For engineering purposes, simpler equations are normally satisfactory and two of the most popular will be considered here.

Since many of the non-Newtonian fluids in engineering applications are pseudoplastic, such fluids will be used in the following to illustrate typical flow curves and constitutive equations. Figure 3.9.4 is a qualitative flow curve for a typical pseudoplastic fluid plotted with logarithmic coordinates. It is seen in the figure that at low shear rates, region (a), the fluid is Newtonian with a constant apparent viscosity of μ_o (called the *zero shear rate viscosity*). At higher shear rates, region (b), the apparent viscosity begins to decrease until it becomes a straight line, region (c). This region (c) is called the power law region and is an important region in fluid mechanics and heat transfer. At higher shear rates than the power law region, there is another transition region (d) until again the fluid becomes Newtonian in region (e). As discussed below, regions (a), (b), and (c) are where most of the engineering applications occur.

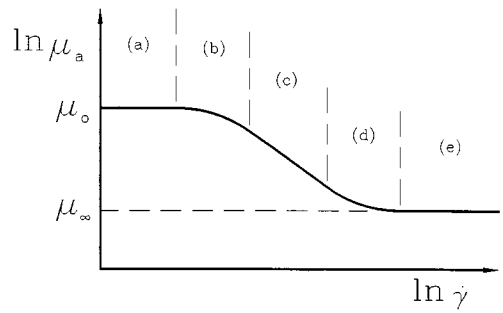


FIGURE 3.9.4 Illustrative flow curve for a pseudoplastic fluid (a) Newtonian region; (b) transition region I: (c) power law region; (d) transition region II; (e) high-shear-rate Newtonian region.

Power Law Constitutive Equation

Region (c) in Figure 3.9.4, which was defined above as the power law region, has a simple constitutive equation:

$$\tau = K\dot{\gamma}^{n} \tag{3.9.3}$$

or, from Equation (3.9.2):

$$\mu_a = K\dot{\gamma}^{n-1} \tag{3.9.4}$$

Here, K is called the fluid consistency and n the flow index. Note that if $n = 1$, the fluid becomes Newtonian and K becomes the dynamic viscosity. Because of its simplicity, the power law constitutive equation has been most often used in rheological studies, but at times it is inappropriate because it has several inherent flaws and anomalies. For example, if one considers the flow of a pseudoplastic fluid ($n < 1$) through a circular duct, because of symmetry at the center of the duct the shear rate (velocity gradient) becomes zero and thus the apparent viscosity from Equation (3.9.4) becomes infinite. This poses conceptual difficulties especially when performing numerical analyses on such systems. Another difficulty arises when the flow field under consideration is not operating in region (c) of Figure 3.9.4 but may have shear rates in region (a) and (b). In this case, the power law equation is not applicable and a more general constitutive equation is needed.

Modified Power Law Constitutive Equation

A generalization of the power law equation which extends the shear rate range to regions (a) and (b) is given by

$$\mu_a = \frac{\mu_o}{1 + \dfrac{\mu_o}{K}\dot{\gamma}^{1-n}} \tag{3.9.5}$$

Examination of Equation (3.9.5) reveals that at low shear rates, the second term in the denominator becomes small compared with unity and the apparent viscosity becomes a constant equal to μ_O. This represents the Newtonian region in Figure 3.9.4. On the other hand, as the second term in the denominator becomes large compared with unity, Equation (3.9.5) becomes Equation (3.9.4) and represents region (c), the power law region. When both denominator terms must be considered, Equation (3.9.5) represents region (b) in Figure 3.9.4.

An important advantage of the modified power law equation is that it retains the rheological properties K and n of the power law model plus the additional property μ_o. Thus, as will be shown later, in the flow and heat transfer equations, the same dimensionless groups as in the power law model will appear plus an additional dimensionless parameter which describes in which of the regions (a), (b), or (c) a particular system is operating. Also, solutions using the modified power law model will have Newtonian and power law solutions as asymptotes.

Equation (3.9.5) describes the flow curve for a pseudoplastic fluid ($n < 1$). For a dilatant fluid, ($n > 1$), an appropriate modified power law model is given by

$$\mu_a = \mu_o\left[1 + \frac{K}{\mu_o}\dot{\gamma}^{n-1}\right] \tag{3.9.6}$$

Many other constitutive equations have been proposed in the literature (Skelland, 1967; Cho and Hartnett, 1982; Irvine and Karni, 1987), but the ones discussed above are sufficient for a large number of engineering applications and agree well with the experimental determinations of rheological properties.

Rheological Property Measurements

For non-Newtonian fluids, specifying the appropriate rheological properties for a particular fluid is formidable because such fluids are usually not pure substances but various kinds of mixtures. This means that the properties are not available in handbooks or other reference materials but must be measured for each particular application. A discussion of the various instruments for measuring rheological properties is outside the scope of the present section, but a number of sources are available which describe different rheological property measurement techniques and instruments: Skelland (1967), Whorlow (1980), Irvine and Karni (1987), and Darby (1988). Figure 3.9.5 is an illustration of experimental flow curves measured with a falling needle viscometer and a square duct viscometer for polymer solutions of different concentrations. Also known in the figure as solid lines is the modified power law equation used to represent the experimental data. It is seen that Equation (3.9.5) fits the experimental data within ±2%. Table 3.9.1 lists the rheological properties used in the modified power law equations in Figure 3.9.5. It must be emphasized that a proper knowledge of these properties is vital to the prediction of fluid mechanics and heat transfer phenomena in rheological fluids.

TABLE 3.9.1 Rheological Properties Used in the Modified Power Law Equations in Figure 3.9.5 for Three Polymer Solutions of CMC-7H4

CMC	K (N · sec^n/m^2)	n	μ_o (N · sec/m^2)n
5000 wppm	2.9040	0.3896	0.21488
2500 wppm	1.0261	0.4791	0.06454
1500 wppm	0.5745	0.5204	0.03673

Source: Park, S. et al., *Proc. Third World Conf. Heat Transfer, Fluid Mechanics, and Thermodynamics,* Vol. 1, Elsevier, New York, 1993, 900–908.

Fully Developed Laminar Pressure Drops for Time-Independent Non-Newtonian Fluids

Modified Power Law Fluids

This important subject will be considered by first discussing modified power law fluids. The reason is that such solutions include both friction factor–Reynolds number relations and a shear rate parameter. The latter allows the designer to determine the shear rate region in which his system is operating and thus the appropriate solution to be used, i.e., regions (a), (b), or (c) in Figure 3.9.4.

For laminar fully developed flow of a modified power law fluid in a circular duct, the product of the friction factor and a certain Reynolds number is a constant depending on the flow index, n, and the shear rate parameter, β.

$$f_D \cdot \mathrm{Re}_m = \text{constant}(n,\beta) \tag{3.9.7}$$

where f_D is the Darcy friction factor and Re_m the modified power law Reynolds number, i.e.,

$$f_D = \frac{2\dfrac{\Delta p}{L} D_H}{\rho \bar{u}^2} \quad \text{(Darcy friction factor)}^*$$

* It should be noted that the Fanning friction factor is also used in the technical literature. The Fanning friction factor is $^1/_4$ of the Darcy friction factor, and will be characterized by the symbol f_F.

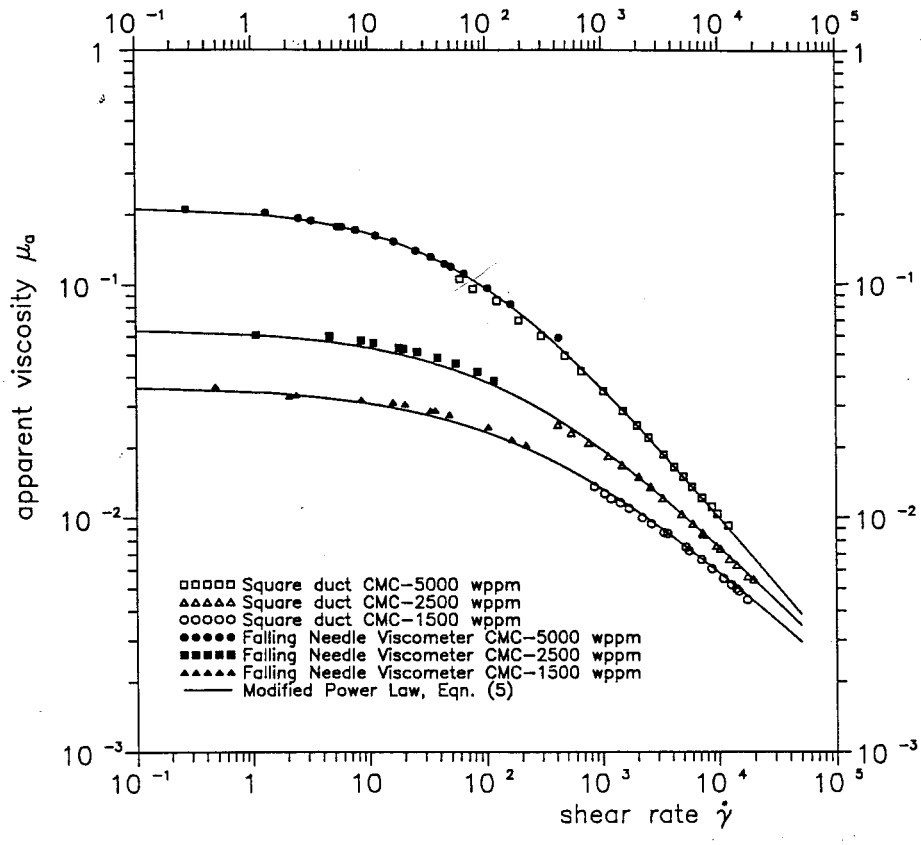


FIGURE 3.9.5 Experimental measurements of apparent viscosity vs. shear rate for polymer solutions (CMC-7H4) at different concentrations. (From Park, S. et al., in *Proc. Third World Conf. Heat Transfer, Fluid Mechanics, and Thermodynamics*, Vol. 1, Elsevier, New York, 1993, 900–908.).

$$\mathrm{Re}_m = \frac{\rho \bar{u} D_H}{\mu^*}$$

$$\mu^* = \frac{\mu_o}{1+\beta}$$

$$\beta = \frac{\mu_o}{K}\left(\frac{\bar{u}}{D_H}\right)^{1-n}$$

where β is the shear rate parameter mentioned previously which can be calculated by the designer for a certain operating duct ($\bar{u}$ and d) and a certain pseudoplastic fluid (μ_o, K, n). The solution for a circular tube has been calculated by Brewster and Irvine (1987) and the results are shown in Figure 3.9.6 and in Table 3.9.2. Referring to 3.9.6, we can see that when the $\log_{10}\beta$ is less than approximately –2, the duct is operating in region (a) of Figure 3.9.4 which is the Newtonian region and therefore classical Newtonian solutions can be used. Note that in the Newtonian region, Re_m reverts to the Newtonian Reynolds number given by

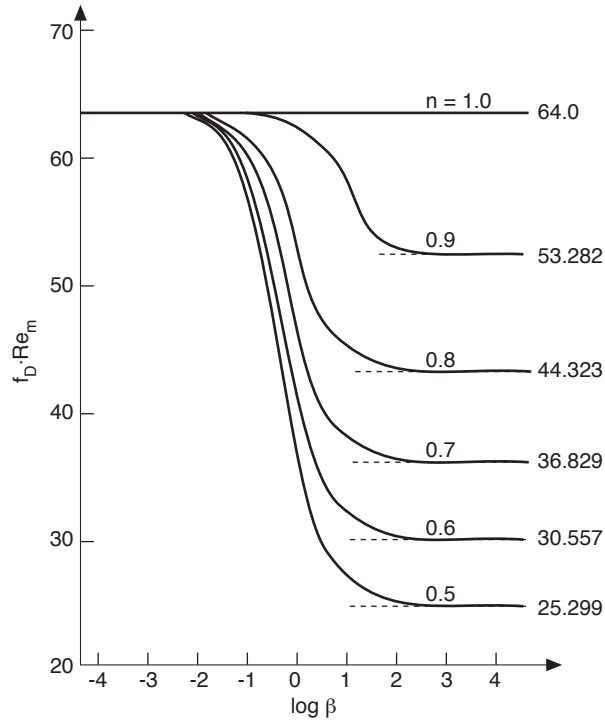


FIGURE 3.9.6 Product of friction factor and modified Reynolds number vs. $\log_{10} \beta$ for a circular duct.

.

TABLE 3.9.2 Summary of Computed Values of $f_D \cdot Re_m$ for Various Values of n and β for a Circular Duct

	$f_D \cdot Re_m$ for Flow Index: n =					
β	**1.0**	**0.9**	**0.8**	**0.7**	**0.6**	**0.5**
10^{-5}	64.000	64.000	64.000	64.000	63.999	63.999
10^{-4}	64.000	63.999	63.997	63.995	63.993	63.990
10^{-3}	64.000	63.987	63.972	63.953	63.930	63.903
10^{-2}	64.000	63.873	63.720	63.537	63.318	63.055
10^{-1}	64.000	62.851	61.519	59.987	58.237	56.243
10^{0}	64.000	58.152	52.377	46.761	41.384	36.299
10^{1}	64.000	54.106	45.597	38.308	32.082	26.771
10^{2}	64.000	53.371	44.458	36.985	30.716	25.451
10^{3}	64.000	53.291	44.336	36.845	30.573	25.314
10^{4}	64.000	53.283	44.324	36.831	30.559	25.300
10^{5}	64.000	53.282	44.323	36.830	30.557	25.299
Exact solution	64.000	53.282	44.323	36.829	30.557	25.298

Source: Brewster, R.A. and Irvine, T.F., Jr., *Wärme und Stoffübertragung*, 21, 83–86, 1987. With permission.

$$\mathrm{Re}_N = \frac{\rho \bar{u} D_H}{\mu_o} \tag{3.9.8}$$

When the value of $\log_{10} \beta$ is approximately in the range $-2 \le \log_{10} \beta \le 2$, the duct is operating in the transition region (b) of Figure 3.9.4 and the values of $f_D \cdot \mathrm{Re}_m$ must be obtained from Figure 3.9.6 or from Table 3.9.2.

When $\log_{10} \beta$ is greater than approximately 2, the duct is operating in the power law region (c) of Figure 3.9.4 and power law friction factor Reynolds number relations can be used. They are also indicated in Figure 3.9.6 and Table 3.9.2. In this region, Re_m becomes the power law Reynolds number given by

$$\mathrm{Re}_g = \frac{\rho \bar{u}^{2-n} D_H^n}{K} \tag{3.9.9}$$

For convenience, Brewster and Irvine (1987) have presented a correlation equation which agrees within 0.1% with the results tabulated in Table 3.9.2.

$$f_D \cdot \mathrm{Re}_m = \frac{1+\beta}{\dfrac{1}{64} + \dfrac{\beta}{2^{3n+3}\left(\dfrac{3n+1}{4n}\right)^n}} \tag{3.9.10}$$

Thus, Equation (3.9.10) contains all of the information required to calculate the circular tube laminar fully developed pressure drop for a pseudoplastic fluid depending upon the shear rate region(s) under consideration, i.e., regions (a), (b), or (c) of Figure 3.9.4. Note that in scaling such non-Newtonian systems, both Re_m and β must be held constant. Modified power law solutions have been reported for two other duct shapes. Park et al. (1993) have presented the friction factor–Reynolds number relations for rectangular ducts and Capobianchi and Irvine (1992) for concentric annular ducts.

Power Law Fluids

Since the power law region of modified power law fluids ($\log_{10} \beta \ge 2$) is often encountered, the friction factor–Reynolds number relations will be discussed in detail in this subsection.

An analysis of power law fluids which is most useful has been presented by Kozicki et al. (1967). Although the method is approximate, its overall accuracy (±5%) is usually sufficient for many engineering calculations. His expression for the friction factor–Reynolds number product is given by

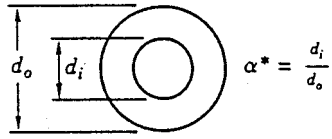
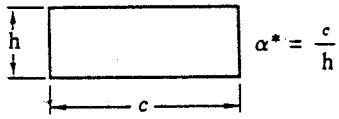
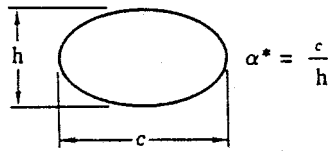
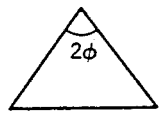
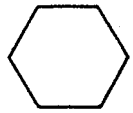
$$f_D \cdot \mathrm{Re}^* = 2^{6n} \tag{3.9.11}$$

where

$$\mathrm{Re}^* = \text{Kozicki Reynolds number}, \quad \mathrm{Re}^* = \frac{\mathrm{Re}_g}{\left[\dfrac{a+bn}{n}\right]^n 8^{n-1}} \tag{3.9.12}$$

and a and b are geometric constants which depend on the cross-sectional shape of the duct. For example, for a circular duct, $a = 0.25$ and $b = 0.75$. Values of a and b for other duct shapes are tabulated in Table 3.9.3. For additional duct shapes in both closed and open channel flows, Kozicki et al. (1967) may be consulted.

TABLE 3.9.3 Constants *a* and *b* for Various Duct Geometrics Used in the Method Due to Kozicki et al. (1967)

Geometry	α^*	*a*	*b*
Concentric annuli 	0.1	0.4455	0.9510
	0.2	0.4693	0.9739
	0.3	0.4817	0.9847
	0.4	0.4890	0.9911
	0.5	0.4935	0.9946
	0.6	0.4965	0.9972
	0.7	0.4983	0.9987
	0.8	0.4992	0.9994
	0.9	0.4997	1.0000
	1.0[a]	0.5000	1.0000
Rectangular 	0.0	0.5000	1.0000
	0.25	0.3212	0.8482
	0.50	0.2440	0.7276
	0.75	0.2178	0.6866
	1.00	0.2121	0.8766
Elliptical 	0.00	0.3084	0.9253
	0.10	0.3018	0.9053
	0.20	0.2907	0.8720
	0.30	0.2796	0.8389
	0.40	0.2702	0.8107
	0.50	0.2629	0.7886
	0.60	0.2575	0.7725
	0.70	0.2538	0.7614
	0.80	0.2515	0.7546
	0.90	0.2504	0.7510
	1.00[b]	0.2500	0.7500
	2ϕ (deg)		
Isosceles triangular 	10	0.1547	0.6278
	20	0.1693	0.6332
	40	0.1840	0.6422
	60	0.1875	0.6462
	80	0.1849	0.6438
	90	0.1830	0.6395
	N		
Regular polygon (*N* sides) 	4	0.2121	0.6771
	5	0.2245	0.6966
	6	0.2316	0.7092
	8	0.2391	0.7241

[a] Parallel plates.

[b] Circle.

Source: Irvine, T.F., Jr. and Karni, J., in *Handbook of Single Phase Convective Heat Transfer,* John Wiley and Sons, New York, 1987, pp 20-1–20-57.

Fully Developed Turbulent Flow Pressure Drops

In a number of engineering design calculations for turbulent flow, the shear rate range falls in region (c) of Figure 3.9.4. Thus, power law relations are appropriate for such pressure drop calculations.

Hartnett and Kostic (1990) have investigated the various correlations which have appeared in the literature for circular tubes and have concluded that for a circular tube the relation proposed by Dodge and Metzner (1959) is the most reliable for pseudoplastic fluids. It is given by

$$\frac{1}{f_F^{1/2}} = \frac{4.0}{n^{0.75}} \cdot \log_{10}\left[\mathrm{Re}'_g \left(f_F\right)^{1-(1/2n)}\right] - \frac{0.40}{n^{1.2}} \tag{3.9.13}$$

where f_F is the Fanning friction factor and

$$\mathrm{Re}'_g = \mathrm{Re}_g \left[\frac{8^{1-n}}{\left[\frac{3n+1}{4n}\right]^n}\right] \tag{3.9.14}$$

Figure 3.9.7 is a graphical representation of Equation (3.9.13) which indicates the Dodge and Metzner experimental regions by solid lines, and by dashed lines where the data are extrapolated outside of their experiments.

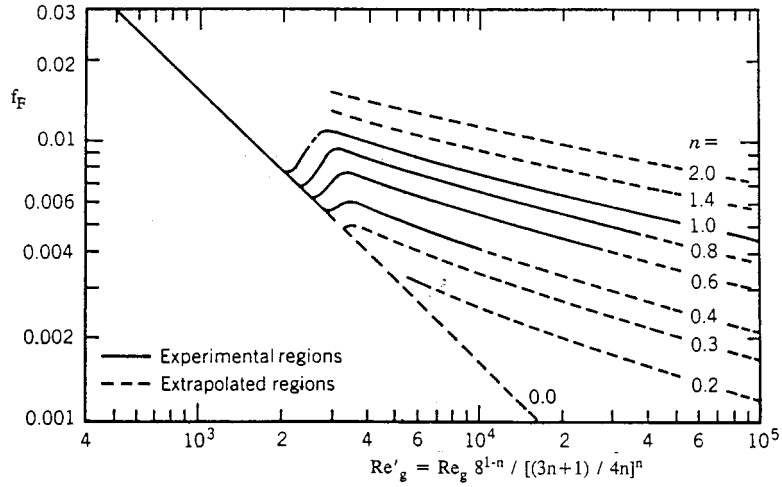


FIGURE 3.9.7 Dodge and Metzner relation between Fanning friction factor and Re'_g. (From Dodge, D.W. and Metzner, A.B., *AIChE J.*, 5, 189–204, 1959.)

For noncircular ducts in turbulent fully developed flow, only a limited amount of experimental data are available. Kostic and Hartnett (1984) suggest the correlation:

$$\frac{1}{f_F^{1/2}} = \frac{4}{n^{0.75}} \cdot \log_{10}\left[\mathrm{Re}^* \left(f_F\right)^{1-(1/2n)}\right] - \frac{0.40}{n^{0.5}} \tag{3.9.15}$$

where f_F is again the Fanning friction factor and Re^* is the Kozicki Reynolds number:

$$\mathrm{Re}^* = \frac{\mathrm{Re}_g}{\left[\frac{a+bn}{n}\right]^n 8^{n-1}} \tag{3.9.16}$$

and a and b are geometric constants given in Table 3.9.3.

Viscoelastic Fluids

Fully Developed Turbulent Flow Pressure Drops

Viscoelastic fluids are of interest in engineering applications because of reductions of pressure drop and heat transfer which occur in turbulent channel flows. Such fluids can be prepared by dissolving small amounts of high-molecular-weight polymers, e.g., polyacrylamide, polyethylene oxide (Polyox), etc., in water. Concentrations as low as 5 parts per million by weight (wppm) result in significant pressure drop reductions. Figure 3.9.8 from Cho and Hartnett (1982) illustrates the reduction in friction factors for Polyox solutions in a small-diameter capillary tube. It is seen that at zero polymer concentration the data agree with the Blasius equation for Newtonian turbulent flow. With the addition of only 7 wppm of Polyox, there is a significant pressure drop reduction and for concentrations of 70 wppm and greater all the data fall on the Virk line which is the maximum drag-reduction asymptote. The correlations for the Blasius and Virk lines as reported by Cho and Hartnett (1982) are

$$f_F = \frac{0.079}{\mathrm{Re}^{1/4}} \quad \text{(Blasius)} \tag{3.9.17}$$

$$f_F = 0.20\,\mathrm{Re}_a^{-0.48} \quad \text{(Virk)} \tag{3.9.18}$$

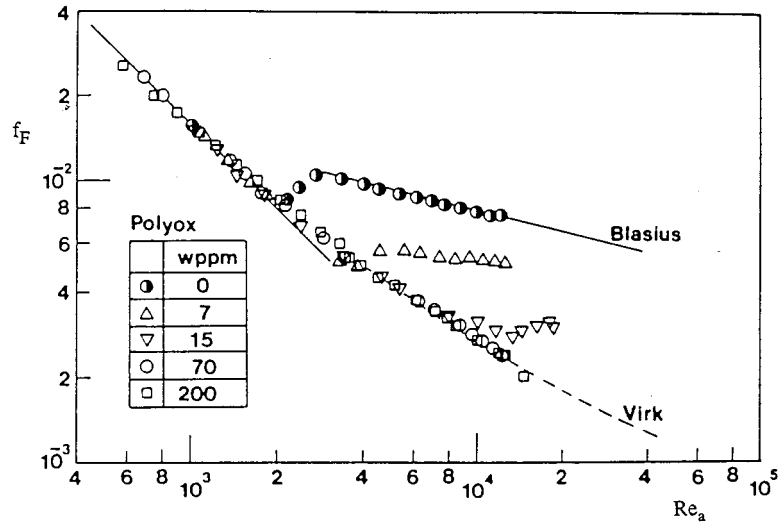


FIGURE 3.9.8 Reduction in friction factors for polyethylene oxide (Polyox) solutions in a small-diameter capillary tube. (From Cho, Y.I. and Harnett, J.P., *Adv. Heat Transfer,* 15, 59–141, 1982. With permission.)

At the present time, no generally accepted method exists to predict the drag reduction between the Blasius and Virk lines. Kwack and Hartnett (1983) have proposed that the amount of drag reduction between those two correlations is a function of the Weissenberg number, defined as

$$w_s = \frac{\lambda \bar{u}}{D_H} \tag{3.9.19}$$

where λ = characteristic time of the viscoelastic fluid. They present correlations which allow the friction factor to be estimated at several Reynolds numbers between the Blasius and Virk lines.

Fully Developed Laminar Flow Pressure Drops

The above discussion on viscoelastic fluids has only considered fully developed turbulent flows. Laminar fully developed flows can be considered as nonviscoelastic but purely viscous non-Newtonian. Therefore, the method of Kozicki et al. (1967) may be applied to such situations once the appropriate rheological properties have been determined.

Nomenclature

a = duct shape geometric constant
b = duct shape geometric constant
c = duct width (see Table 3.9.3) (m)
d_i = concentric annuli inner diameter (see Table 3.9.3) (m)
d_o = concentric annuli outer diameter (see Table 3.9.3) (m)
f_D = Darcy friction factor
f_F = Fanning friction factor
h = duct height (see Table 3.9.3) (m)
K = fluid consistency (Ns^n/m^2)
n = flow index
N = number of sides in polygon (see Table 3.9.3)
Re_g = generalized Reynolds number,

$$\mathrm{Re}_g = \frac{\rho \bar{u}^{2-n} D_H^n}{K}$$

Re_m = modified power law Reynolds number,

$$\mathrm{Re}_m = \frac{\rho \bar{u} D_H}{\mu^*}$$

Re_N = modified power law Reynolds number Newtonian asymptote,

$$\mathrm{Re}_N = \frac{\rho \bar{u} D_H}{\mu_o}$$

Re_a = apparent Reynolds number

$$\mathrm{Re}_a = \frac{\mathrm{Re}_g}{\left(\frac{3n+1}{4n}\right)^{n-1} 8^{n-1}}$$

Re^* = Kozicki Reynolds number

$$\mathrm{Re}^* = \frac{\rho \bar{u}^{2-n} D_H^n}{K\left[\frac{a+bn}{n}\right]^n 8^{n-1}}$$

Re'_g = Metzner Reynolds number

$$\mathrm{Re}'_g = \mathrm{Re}_g \left[\frac{8^{1-n}}{\left[\frac{3n+1}{4n} \right]^n} \right]$$

$\bar{u}$ = average streamwise velocity (m/sec)
t = time (sec)
w_s = Weissenberg number
x = direction of shear stress (m)
y = direction of velocity gradient (m)

Greek

α^* = duct aspect ratio in Table 3.9.3
β = shear rate parameter

$$\beta = \frac{\mu_o}{K} \left(\frac{\bar{u}}{D_H} \right)^{1-n}$$

$\dot{\gamma}$ = shear rate (L/sec)
ΔP = presure drop (N/m^2)
λ = characteristic time of viscoelastic fluid (sec)
μ_a = apparent viscosity (N · sec/m^2)
μ_o = zero shear rate viscosity (N · sec/m^2)
μ_∞ = high shear rate viscosity (N · sec/m^2)
μ^* = reference viscosity

$$\mu^* = \frac{\mu_o}{1+\beta} \quad \left(\mathrm{N \cdot sec/m^2} \right)$$

τ_o = yield stress (N/m^2)
$\tau_{y,x}$ = shear stress (N/m^2)
ϕ = half apex angle (see Table 3.9.3) (°)

References

Brewster, A.A. and Irvine, T.F. Jr. 1987. Similtude considerations in laminar flow of power law fluids in circular ducts, *Wärme und Stoffübertagung,* 21:83–86.

Capobianchi, M. and Irvine, T.F. Jr. 1992. Predictions of pressure drop and heat transfer in concentric annular ducts with modified power law fluids, *Wärme und Stoffübertagung,* 27:209–215.

Cho, Y.I. and Hartnett, J.P. 1982. Non-Newtonian fluids in circular pipe flow, in *Adv. Heat Transfer,* 15:59–141.

Darby, R. 1988. Laminar and turbulent pipe flows of non-Newtonian fluids, in *Encyclopedia of Fluid Mechanics,* Vol. 7, Gulf Publishing, Houston, 7:20–53.

Dodge, D.W. and Metzner, A.B. 1959. Turbulent flow of non-Newtonian systems, *AIChE J.,* 5:189–204.

Harnett, J.P. and Kostic, M. 1990. Turbulent Friction Factor Correlations for Power Law Fluids in Circular and Non-Circular Channels, *Int. Comm. Heat and Mass Transfer,* 17:59–65.

Irvine, T.F. Jr. and Karni, J. 1987. Non-Newtonian fluid flow and heat transfer, in *Handbook of Single Phase Convective Heat Transfer,* pp. 20-1–20-57, John Wiley and Sons, New York.

Kostic, M. and Hartnett, J.P. 1984. Predicting turbulent friction factors of non-Newtonian fluids in non-circular ducts, *Int. Comm. Heat and Mass Transfer,* 11:345–352.

Kozicki, W., Chou, C.H., and Tiu, C. 1967. Non-Newtonian flow in ducts of arbitrary cross-sectional shape, *Can. J. Chem. Eng.,* 45:127–134.

Kwack, E.Y. and Hartnett, J.P. 1983. Empirical correlations of turbulent friction factors and heat transfer coefficients for viscoelastic fluids, *Int. Comm. Heat and Mass Transfer,* 10:451–461.

Park, S., Irvine, T.F. Jr., and Capobianchi, M. 1993. Experimental and numerical study of friction factor for a modified power law fluid in a rectangular duct, *Proc. Third World Conf. Heat Transfer, Fluid Mechanics, and Thermodynamics,* Vol. 1, Elsevier, New York, 1:900–908.

Skelland, A.H.P. 1967. *Non-Newtonian Flow and Heat Transfer,* John Wiley and Sons, New York.

Whorlow, R.W. 1980. *Rheological Techniques,* Halsted Press, New York.

Further Information

It is not possible to include all of the interesting non-Newtonian topics in a section of this scope. Other items which may be of interest and importance are listed below along with appropriate references: hydrodynamic and thermal entrance lengths, Cho and Hartnett (1982); non-Newtonian flow over external surfaces, Irvine and Karni (1987); chemical, solute, and degradation effects in viscoelastic fluids, Cho and Harnett (1982); general references, Skelland (1967), Whorlow (1980), and Darby (1988).

3.10 Tribology, Lubrication, and Bearing Design

Francis E. Kennedy, E. Richard Booser, and Donald F. Wilcock

Introduction

Tribology, the science and technology of contacting surfaces involving friction, wear, and lubrication, is extremely important in nearly all mechanical components. A major focus of the field is on friction, its consequences, especially wear and its reduction through lubrication and material surface engineering. The improper solution of tribological problems is responsible for huge economic losses in our society, including shortened component lives, excessive equipment downtime, and large expenditures of energy. It is particularly important that engineers use appropriate means to reduce friction and wear in mechanical systems through the proper selection of bearings, lubricants, and materials for all contacting surfaces. The aim of this section is to assist in that endeavor.

Sliding Friction and its Consequences

Coefficient of Friction

If two stationary contacting bodies are held together by a normal force W and a tangential force is applied to one of them, the tangential force can be increased until it reaches a magnitude sufficient to initiate sliding. The ratio of the friction force at incipient sliding to the normal force is known as the static coefficient of friction, f_s. After sliding begins, the friction force always acts in the direction opposing motion and the ratio between that friction force and the applied normal force is the kinetic coefficient of friction, f_k.

Generally, f_k is slightly smaller than f_s and both coefficients are independent of the size or shape of the contacting surfaces. Both coefficients are very much dependent on the materials and cleanliness of the two contacting surfaces. For ordinary metallic surfaces, the friction coefficient is not very sensitive to surface roughness. For ultrasmooth or very rough surfaces, however, the friction coefficient can be larger. Typical friction coefficient values are given in Table 3.10.1. Generally, friction coefficients are greatest when the two surfaces are identical metals, slightly lower with dissimilar but mutually soluble metals, still lower for metal against nonmetal, and lowest for dissimilar nonmetals.

TABLE 3.10.1 Some Typical Friction Coefficients[a]

Material Pair	Static Friction Coefficient f_s		Kinetic Friction Coefficient f_k	
	In Air	In Vacuo	In Air, Dry	Oiled
Mild steel vs. mild steel	0.75	—	0.57	0.16
Mild steel vs. copper	0.53	0.5 (oxidized) 2.0 (clean)	0.36	0.18
Copper vs. copper	1.3	21.0	0.8	0.1
Tungsten carbide vs. copper	0.35	—	0.4	—
Tungsten carbide vs. tungsten carbide	0.2	0.4	0.15	—
Mild steel vs. polytetrafluoroethylene	0.04	—	0.05	0.04

[a] The friction coefficient values listed in this table were compiled from several of the references listed at the end of this section.

The kinetic coefficient of friction, f_k, for metallic or ceramic surfaces is relatively independent of sliding velocity at low and moderate velocities, although there is often a slight decrease in f_k at higher velocities. With polymers and soft metals there may be an increase in the friction coefficient with increasing velocity until a peak is reached, after which friction may decrease with further increases in velocity or temperature. The decrease in kinetic friction coefficient with increasing velocity, which may become especially pronounced at higher sliding velocities, can be responsible for friction-induced

vibrations (stick–slip oscillations) of the sliding systems. Such vibrations are an important design consideration for clutches and braking systems, and can also be important in the accurate control and positioning of robotic mechanisms and precision manufacturing systems.

Wear

Wear is the unwanted removal of material from solid surfaces by mechanical means; it is one of the leading reasons for the failure and replacement of manufactured products. It has been estimated that the costs of wear, which include repair and replacement, along with equipment downtime, constitute up to 6% of the U.S. gross national product (Rabinowicz, 1995). Wear can be classified into four primary types: sliding wear, abrasion, erosion, and corrosive wear. Owing to its importance, wear and its control have been the subject of several handbooks (Peterson and Winer, 1980; Blau, 1992), which the interested reader may consult for further information.

Types of Wear. Sliding wear occurs to some degree whenever solid surfaces are in sliding contact. There are two predominant sliding wear mechanisms, adhesion and surface fatigue. *Adhesive wear* is caused by strong adhesive forces between the two surfaces within the real area of contact. It results in the removal of small particles from at least one of the surfaces, usually the softer one. These particles can then transfer to the other surface or mix with other material from both surfaces before being expelled as loose wear debris. Adhesive wear can be particularly severe for surfaces which have a strong affinity for each other, such as those made from identical metals. *Surface fatigue wear* occurs when repeated sliding or rolling/sliding over a wear track results in the initiation of surface or subsurface cracks, and the propagation of those cracks produces wear particles in ductile materials by a process that has been called delamination. With brittle materials, sliding wear often occurs by a *surface fracture* process.

After an initial transition or "running-in" period, sliding wear tends to reach a steady state rate which is approximated by the following Archard (or Holm/Archard) wear equation:

$$V = K * W * s/H \tag{3.10.1}$$

where V = volume of worn material, K = dimensionless wear coefficient, s = sliding distance, W = normal load between the surfaces, and H = hardness of the softer of the two contacting surfaces.

The dimensionless wear coefficient gives an indication of the tendency of a given material combination to wear; relative wear coefficient values are given in Figure 3.10.1. In general, wear coefficients are highest for identical metals sliding without lubrication, and wear is decreased by adding a lubricant and by having material pairs which are dissimilar.

Abrasive wear occurs when a hard, rough surface slides against a softer surface (*two-body abrasion*) or when hard particles slide between softer surfaces (*three-body abrasion*). This process usually results in material removal by plowing or chip formation, especially when the abraded surface is metallic; surface fracture can occur during abrasion of brittle surfaces. In fact, abrasion mechanisms are similar to those of grinding and lapping, which could be considered as intentional abrasion. Consideration of the cutting and plowing processes shows that abrasive wear obeys the same equation (3.10.1) as sliding wear (Archard, 1980; Rabinowicz, 1995). Typical wear coefficients for abrasive wear are given in Figure 3.10.1. Since the relative size, hardness, and sharpness of the abrading particles, or surface asperities, also affect abrasive wear rates, the wear coefficients for abrasion must include recognition of those factors (Rabinowicz, 1995).

Erosion occurs when solid particles or liquid droplets impinge on a solid surface. When impingement is on a ductile metallic surface, the wear process is similar to that caused by abrasion, and is dominated by plastic deformation. Brittle surfaces, on the other hand, tend to erode by surface fracture mechanisms. The material removal rate is dependent on the angle of attack of the particles, with erosion reaching a peak at low angles (about 20°) for ductile surfaces and at high angles (90°) for brittle materials. In either case, the wear rate is proportional to the mass rate of flow of the particles and to their kinetic energy; it is inversely proportional to the hardness of the surface and the energy-absorbing potential (or toughness)

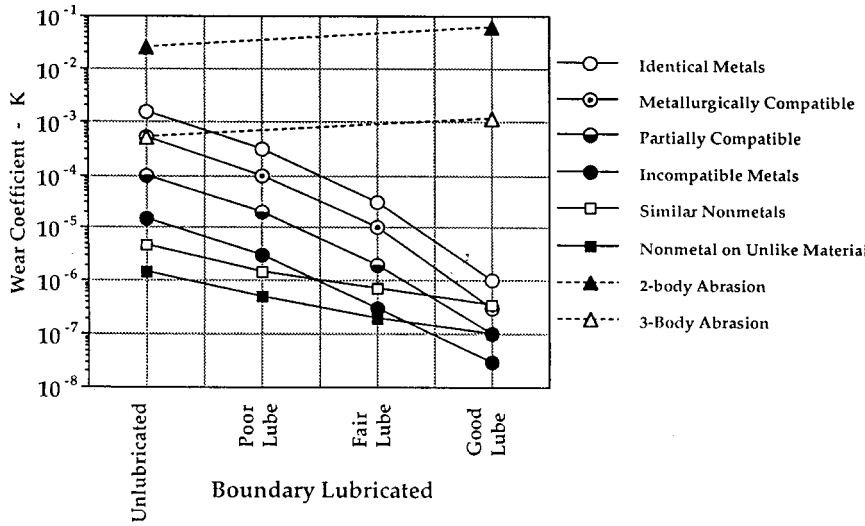


FIGURE 3.10.1 Typical values of wear coefficient for sliding and abrasive wear. (Modified from Rabinowicz, 1980, 1995.)

of the impinged surface (Schmitt, 1980). Although erosion is usually detrimental, it can be used beneficially in such material removal processes as sandblasting and abrasive water jet machining.

Corrosive wear results from a combination of chemical and mechanical action. It involves the synergistic effects of chemical attack (corrosion) of the surface, followed by removal of the corrosion products by a wear mechanism to expose the metallic surface, and then repetition of those processes. Since many corrosion products act to protect the surfaces from further attack, the removal of those films by wear acts to accelerate the rate of material removal. Corrosive wear can become particularly damaging when it acts in a low-amplitude oscillatory contact, which may be vibration induced, in which case it is called *fretting corrosion*.

Means for Wear Reduction.
The following actions can be taken to limit sliding wear:

- Insure that the sliding surfaces are well lubricated. This can best be accomplished by a liquid lubricant (see sub-section on effect of lubrication on friction and wear), but grease, or solid lubricants such as graphite or molybdenum disulfide, can sometimes be effective when liquid lubricants cannot be used.
- Choose dissimilar materials for sliding pairs.
- Use hardened surfaces.
- Add wear-resistant coatings to the contacting surfaces (see the following subsection).
- Reduce normal loads acting on the contact.
- Reduce surface temperatures. This is particularly important for polymer surfaces.

To reduce abrasive wear:

- Use hardened surfaces.
- Add a hard surface coating.
- Reduce the roughness of hard surfaces that are in contact with softer surfaces.
- Provide for the removal of abrasive particles from contacting surfaces. This can be done by flushing surfaces with liquid and/or filtering liquid coolants and lubricants.
- Reduce the size of abrasive particles.

To reduce erosion:

- Modify the angle of impingement of solid particles or liquid droplets.
- Provide for the removal of solid particles from the stream of fluid.
- Use hardened surfaces.
- Use tough materials for surfaces.
- Add protective coating to surfaces.

Surface Engineering for Friction and Wear Reduction

Surface treatments have long been an important remedy for wear problems, and that importance has grown in recent years with the introduction of new techniques to harden surfaces or apply hard surface coatings. Available processes and characteristics for treating a steel substrate are listed in Table 3.10.2.

Thermal transformation hardening processes are used to harden ferrous (primarily steel) surfaces by heating the surface rapidly, transforming it to austenite, and then quenching it to form martensite. The source of heat can be one of the following: an oxyacetylene or oxypropane flame (*flame hardening*), eddy currents induced by a high-frequency electric field (*induction hardening*), a beam from a high-power laser (*laser hardening*), or a focused electron beam (*electron beam hardening*). The depth and uniformity of the hard layer depend on the rate and method of heating. These processes are characterized by a short process time and all except electron beam hardening (which requires a moderate vacuum) can be done in air.

TABLE 3.10.2 Characteristics of Surface Treatment Processes for Steel

	Coating or Treated Layer		
Process	**Hardness (*HV*)**	**Thickness (μm)**	**Substrate Temperature (°C)**
Surface hardening			
Flame or induction hardening	500–700	250–6000	800–1000
Laser or electron beam hardening	500–700	200–1000	950–1050
Carburizing	650–900	50–1500	800–950
Carbonitriding	650–900	25–500	800–900
Nitriding	700–1200	10–200	500–600
Boronizing	1400–1600	50–100	900–1100
Coating			
Chrome plating	850–1250	1–500	25–100
Electroless nickel	500–700	0.1–500	25–100
Hardfacing	800–2000	500–50000	1300–1400
Thermal spraying	400–2000	50–1500	<250
Physical vapor deposition	100–3000	0.05–10	100–300
Chemical vapor deposition	1000–3000	0.5–100	150–2200
Plasma-assisted chemical vapor deposition	1000–5000	0.5–10	<300
Ion implantation	750–1250	0.01–0.25	<200

Thermal diffusion processes involve the diffusion of atoms into surfaces to create a hard layer. In the most widely used of these processes, *carburizing* (or case hardening), carbon diffuses into a low-carbon steel surface to produce a hard, carbon-rich case. The hardness and thickness of the case depend on the temperature, exposure time, and source of carbon (either a hydrocarbon gas, a salt bath, or a packed bed of carbon). *Carbonitriding* is a process similar to carburizing which involves the simultaneous diffusion of carbon and nitrogen atoms into carbon steel surfaces. In the *nitriding* process, nitrogen atoms diffuse into the surface of a steel which contains nitride-forming elements (such as Al, Cr, Mo, V, W, or Ti) and form fine precipitates of nitride compounds in a near-surface layer. The hardness of the surface layer depends on the types of nitrides formed. The source of nitrogen can be a hot gas (usually ammonia) or a plasma. *Nitrocarburizing* and *boronizing* are related processes in which nitrogen or boron atoms diffuse

into steel surfaces and react with the iron to form a hard layer of iron carbonitride or iron boride, respectively.

Thin, hard metallic coatings can be very effective in friction and wear reduction and can be applied most effectively by *electroplating processes* (Weil and Sheppard, 1992). The most common of such coatings are *chromium*, which is plated from a chromic acid bath, and *electroless nickel*, which is deposited without electric current from a solution containing nickel ions and a reducing agent. Chromium coatings generally consist of fine-grained chromium with oxide inclusions, while electroless nickel coatings contain up to 10% of either phosphorus or boron, depending on the reducing agent used.

Thermal spray processes (Kushner and Novinski, 1992) enable a large variety of coating materials, including metals, ceramics and polymers, to be deposited rapidly on a wide range of substrates. Four different thermal spray processes are commercially available: *oxyfuel* (or flame) spraying of metallic wire or metallic or ceramic powder, *electric arc* spraying of metallic wire, *plasma arc* spraying of powder (metallic or ceramic), and *high-velocity oxyfuel* (or detonation gun) powder spray. In each thermal spray process the coating material, in either wire or powder form, is heated to a molten or plastic state, and the heated particles are propelled toward the surface to be coated where they adhere and rapidly solidify to form a coating. The hardness of the coating depends on both the sprayed material and the process parameters.

Weld hardfacing processes (Crook and Farmer, 1992) involve the application of a wear-resistant material to the surface of a part by means of a weld overlay. Weld overlay materials include ferrous alloys (such as martensitic air-hardening steel or high-chromium cast iron), nonferrous alloys (primarily cobalt- or nickel-based alloys containing hard carbide, boride, or intermetallic particles), and cemented carbides (usually tungsten carbide/cobalt cermets). In each case the surface being coated is heated to the same temperature as the molten weld layer, thus posing a limitation to the process. Weld hardfacing is best used when abrasion or sliding wear cannot be avoided (as with earthmoving or mining equipment) and the goal is to limit the wear rate.

Vapor deposition processes for wear-resistant coatings include *physical vapor deposition* (PVD), *chemical vapor deposition* (CVD), and several variants of those basic processes (Bhushan and Gupta, 1991). Each of the processes consists of three steps: (1) creation of a vapor phase of the coating material, (2) transportation of the vapor from source to substrate, and (3) condensation of the vapor phase on the substrate and growth of a thin solid film. In PVD processes the vapor is produced by either evaporation (by heating of the coating source) or sputtering (in which coating material is dislodged and ejected from the source as a result of bombardment by energetic particles). In some PVD processes the vapor becomes ionized or reacts with a gas or plasma en route to the substrate, thus modifying the structure or composition of the deposited film. In CVD processes a gas composed of a volatile component of the coating material is activated either thermally or by other means in the vicinity of the substrate, and it reacts to form a solid deposit on the surface of the hot substrate.

Both PVD and CVD methods can be used to produce a wide variety of coatings, including metals, alloys, and refractory compounds. Among the most popular vapor-deposited hard coatings for wear resistance are titanium nitride and titanium carbide. Deposition rates are relatively low compared with some other coating processes, ranging from <0.1 μm/min for some ion beam–sputtering or ion-plating processes, up to 25 μm/min or more for activated reactive evaporation or CVD processes. Most PVD processes are done in a vacuum, while CVD processes are done in a reaction chamber which may be at atmospheric pressure. *Plasma-assisted chemical vapor deposition* (PACVD) is a hybrid process in which the constituents of the vapor phase react to form a solid film when assisted by a glow discharge plasma. The advantages of PACVD over other CVD processes include lower substrate temperatures, higher deposition rates, and a wider variety of coating possibilities.

Ion implantation (Fenske, 1992) is a process in which charged particles are created in an ion source, accelerated toward the surface at high velocity, and then injected into the substrate surface. The most commonly implanted ions for surface engineering are nitrogen, carbon, boron, and titanium, although virtually any element could be implanted. The microstructure of the near-surface region is changed by

the presence of the implanted ions and the result can be high near-surface hardness and wear resistance. The affected layer is very thin (<1 μm).

Effect of Lubrication on Friction and Wear

Whenever lubricated surfaces slide together at low sliding speeds or with a high applied normal load, the lubricant may not separate the two solid surfaces completely. However, the lubricant can still significantly reduce the friction coefficient by reducing the shear strength of adhesive junctions between the two surfaces. In this so-called boundary lubrication regime, the effectiveness of the lubricant can be improved if the lubricant molecules adhere well to the solid surfaces. This is best accomplished by introducing a lubricant or additive that forms a surface film through adsorption, chemisorption, or chemical reaction with the surface. The ensuing reduced shear strength of the surface film can lower the friction coefficient by as much as an order of magnitude from the dry friction value.

When a good supply of a viscous lubricant is available, the separation between the surfaces will increase as the sliding speed increases or the normal load decreases. As the separation increases, the amount of solid/solid contact between the surfaces will decrease, as will the friction coefficient and wear rate. In this "mixed friction" regime, friction is determined by the amount of plowing deformation on the softer surface by the harder surface asperities and by adhesion within the solid/solid contacts. When the surfaces become completely separated by a self-acting or externally pressurized lubricant film, the lubricating regime is hydrodynamic, wear is reduced to nearly zero, and friction reaches a low value governed by viscous shear of the lubricant. Friction coefficients in such cases can be 0.001 or lower, depending on the surface velocities and the lubricant viscosity. This is the case for most journal or thrust bearings (see subsection on fluid film bearings).

Bearings for Friction Reduction

Most mechanical systems contain moving components, such as shafts, which must be supported and held in position by stationary members. This is best done by appropriate design or selection of bearings to be used wherever the moving member is to be supported. Most bearings may be classified as either fluid film bearings, dry or semilubricated bearings, or rolling element bearings.

Fluid film bearings (see subsection below) have a conformal geometry, with a thin film of fluid separating the two surfaces. The fluid lubricant could be a liquid, such as oil, or a gas, such as air. Fluid film bearings are commonly used to support rotating cylindrical shafts, and the load on such a bearing could be either radial, in which case the bearing is called a journal bearing, or axial, for a thrust bearing. In most cases the fluid film is generated by the motion within the bearing itself, so the bearing is called self-acting or hydrodynamic. Whether or not a self-acting bearing can develop a fluid film sufficient to separate and support the two surfaces is determined by magnitude of the quantity $\mu U/W$, where μ is the (absolute) fluid viscosity, U is the relative sliding velocity, and W is the normal load. If that quantity is too small, the fluid film will be too thin and high friction will occur. This can be a problem during start-up of equipment when sliding velocities are low. That problem can be overcome by pressurizing the fluid film from an external pressure source to create a hydrostatic bearing. Whether the fluid film is externally pressurized (hydrostatic) or self-acting (hydrodynamic), separation of the solid surfaces allows wear to be essentially eliminated and friction to be very low, even when very large loads are carried by the pressurized lubricant.

Dry and semilubricated bearings (see subsection below) have conformal surfaces which are in direct contact with each other. This category includes bearings which run dry (without liquid lubrication) or those which have been impregnated with a lubricant. Dry bearings are made of a material such as a polymer or carbon-graphite which has a low friction coefficient, and they are generally used in low-load and low-speed applications. Semilubricated bearings are made of a porous material, usually metal, and are impregnated with a lubricant which resides within the pores. The lubricant, which could be oil or grease, cannot provide a complete fluid film, but usually acts as a boundary lubricant. Semilubricated bearings can carry greater loads at greater speeds than dry bearings, but not as high as either fluid film or rolling element bearings. The failure mechanism for both dry and semilubricated bearings is wear.

Rolling element bearings (see subsection below) have the advantage that rolling friction is lower than sliding friction. These bearings include rolling elements, either balls or rollers, between hardened and ground rings or plates. Their main advantage over fluid film bearings is that they have low friction both during start-up and at operating velocities, although the friction can be higher than that of fluid film bearings during steady state operation. Ball and roller bearings are most commonly lubricated by either oil or grease. In either case the lubricating film at the concentrated contacts between rolling elements and rings is very thin and the pressures in the film are very high; this is the condition known as elastohydrodynamic lubrication. Rolling element bearings fail by a number of mechanisms, often stemming from improper installation or use or from poor lubrication, but the overriding failure mechanism is rolling contact fatigue.

Each type of bearing has advantages and disadvantages, and these are summarized in Table 3.10.3. The Engineering Sciences Data Unit (ESDU) (1965; 1967) has developed some general guides to the selection of bearing type for different load and speed conditions, and those guides for journal and thrust bearing selection are given in Figures 3.10.2 and 3.10.3.

TABLE 3.10.3 Bearing Characteristics

	Fluid Film Bearings	Dry Bearings	Semilubricated	Rolling Element Bearings
Start-up friction coefficient	0.25	0.15	0.10	0.002
Running friction coefficient	0.001	0.10	0.05	0.001
Velocity limit	High	Low	Low	Medium
Load limit	High	Low	Low	High
Life limit	Unlimited	Wear	Wear	Fatigue
Lubrication requirements	High	None	Low/None	Low
High temperature limit	Lubricant	Material	Lubricant	Lubricant
Low temperature limit	Lubricant	None	None	Lubricant
Vacuum	Not applicable	Good	Lubricant	Lubricant
Damping capacity	High	Low	Low	Low
Noise	Low	Medium	Medium	High
Dirt/dust	Need Seals	Good	Fair	Need seals
Radial space requirement	Small	Small	Small	Large
Cost	High	Low	Low	Medium

Lubricant Properties

Petroleum Oils

The vast majority of lubricants in use today are mineral oils which are obtained through the distillation of crude petroleum. Mineral oils are composed primarily of three types of hydrocarbon structures: paraffinic, aromatic, and alicyclic (naphthenic). The molecular weights of the hydrocarbons range from about 250 for low-viscosity grades, up to nearly 1000 for more-viscous lubricants.

Mineral oils by themselves do not have all of the properties required of modern lubricants. For that reason, almost all current lubricants are fortified with a chemical additive package which consists of some of the following:

Oxidation inhibitors limit oxidation of hydrocarbon molecules by interrupting the hydroperoxide chain reaction.

Rust inhibitors are surface-active additives that preferentially adsorb on iron or steel surfaces and prevent their corrosion by moisture.

Antiwear and extreme pressure agents form low shear strength films on metallic surfaces which limit friction and wear, particularly in concentrated contacts.

Friction modifiers form adsorbed or chemisorbed surface films which are effective in reducing friction of bearings during low-speed operation (boundary lubrication regime).

Detergents and dispersants reduce deposits of oil-insoluble compounds (e.g., sludge) in internal combustion engines.

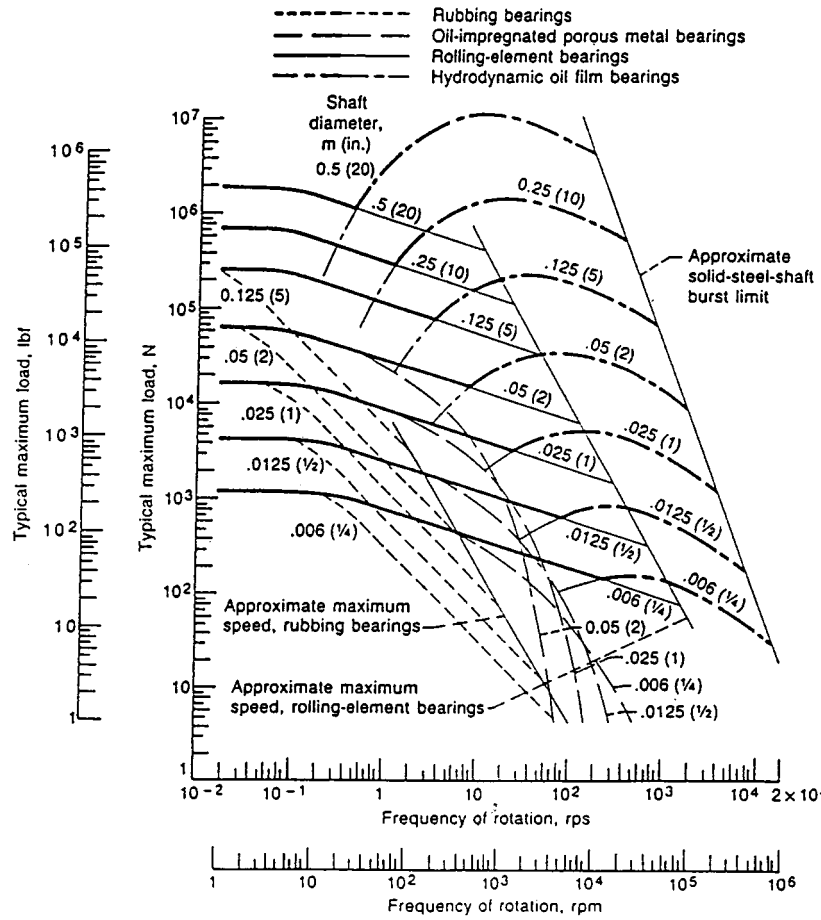


FIGURE 3.10.2 General guide to journal bearing–type selection. Except for rolling element bearings, curves are drawn for bearings with width/diameter = 1. A medium-viscosity mineral oil is assumed for hydrodynamic bearings. From ESDU, *General Guide to the Choice of Journal Bearing Type*, Item 67073, Institution of Mechanical Engineers, London, 1965. With permission.

Pour-point depressants lower the temperature at which petroleum oils become immobilized by crystallized wax.

Foam inhibitors are silicone polymers which enhance the separation of air bubbles from the oil.

Viscosity-index improvers are long-chain polymer molecules which reduce the effect of temperature on viscosity. They are used in multigrade lubricants.

Properties of Petroleum Oils

The lubricating oil property which is of most significance to bearing performance is viscosity. The absolute viscosity, designated as μ, could be given in SI units as pascal second (Pa · sec = N · sec/m^2) or centipoise (1 cP = 0.001 Pa · sec) or in English units as lb · sec/in^2 (or reyn). Kinematic viscosity, designated here as ν, is defined as absolute viscosity divided by density. It is given in SI units as m^2/sec or centistokes (1 cSt = 10^{-6} m^2/sec) and in English units as in.^{2}/sec.

Viscosity data in Table 3.10.4 are representative of typical petroleum "turbine" and "hydraulic" oils which are widely used in industry and closely correspond to properties of most other commercially available petroleum oils. Table 3.10.5 gives equivalent viscosity grades for common automotive (SAE), gear (SAE and AGMA), and reciprocating aircraft engine (SAE) oils (Booser, 1995). Equivalent ISO viscosity grades are listed for the single-graded SAE automotive oils such as SAE 10W and SAE 30.

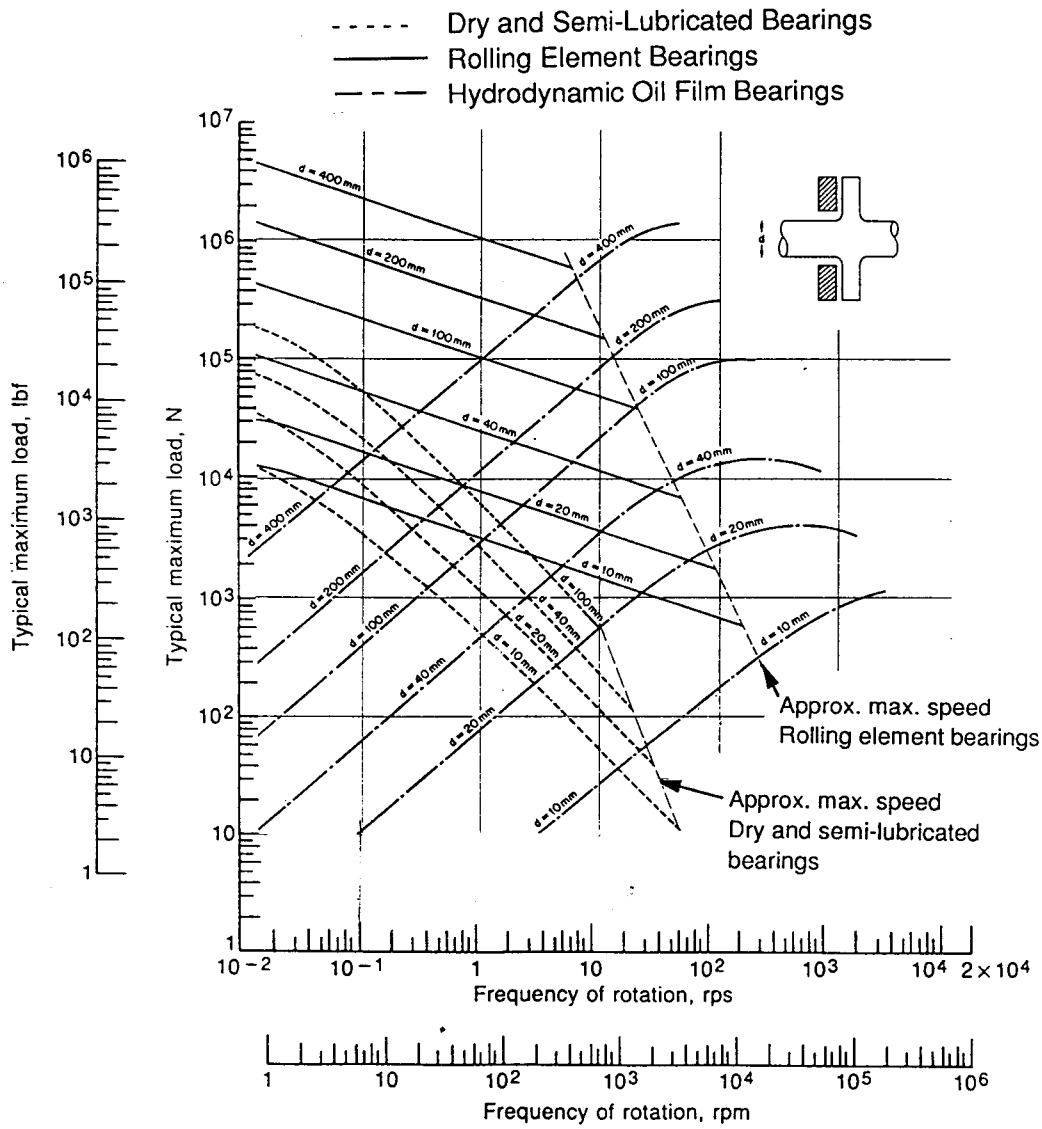


FIGURE 3.10.3 General guide to thrust bearing type selection. Except for rolling element bearings, curves are drawn for a ratio of inside diameter to outside diameter equal to 2 and for a nominal life of 10,000 hr. A medium-viscosity mineral oil is assumed for hydrodynamic bearings. (Based on ESDU, *General Guide to the Choice of Journal Bearing Type*, Item 67073, Institution of Mechanical Engineers, London, 1965, and Neale, M.J., *Bearings*, Butterworth-Heinemann, Oxford, 1993.)

For multigrade oils such as SAE 10W–30, however, the added viscosity-index improvers provide unique viscosity–temperature characteristics. Typical properties of a number of these multigrade SAE oils are included in Table 3.10.4.

ISO viscosity grade 32 and the equivalent SAE 10W are most widely used industrially. Lower-viscosity oils often introduce evaporation and leakage problems, along with diminished load capacity. Higher viscosity may lead to high temperature rise, unnecessary power loss, and start-up problems at low temperature. For low-speed machines, however, higher-viscosity oils ranging up to ISO 150, SAE 40 and sometimes higher are often used to obtain higher load capacity.

TABLE 3.10.4 Representative Oil Properties

	Viscosity				Density	
	Centistokes		10^{-6} reyns(lb·sec/in^2)		gm/cc	lb/in^3
	40°C	100°C	104°F	212°F	40°C	104°F
ISO Grade (Equivalent SAE)						
32 (10W)	32.0	5.36	3.98	0.64	0.857	0.0310
46 (20)	46.0	6.76	5.74	0.81	0.861	0.0311
68 (20W)	68.0	8.73	8.53	1.05	0.865	0.0313
100 (30)	100.0	11.4	12.60	1.38	0.869	0.0314
150 (40)	150.0	15.0	18.97	1.82	0.872	0.0315
220 (50)	220.0	19.4	27.91	2.36	0.875	0.0316
SAE Multigrade						
5W-30	64.2	11.0	8.15	0.99	0.860	0.0311
10W-30	69.0	11.0	8.81	1.08	0.865	0.0312
10W-40	93.5	14.3	11.9	1.45	0.865	0.0312
20W-50	165.5	18.7	21.3	2.74	0.872	0.0315

TABLE 3.10.5 Equivalent Viscosity Grades for Industrial Lubricants

ISO-VG	Viscosity, cSt (at 40°C)		SAE Crankcase	SAE Aircraft	SAE Gear	AGMA Gear Lube Grades	
Grade	Mimimum	Maximum	Oil Grades[a]	Oil Grades[a]	Lube Grades[a]	Regular	EP
2	1.98	2.42	—	—	—	—	—
3	2.88	3.52	—	—	—	—	—
5	4.14	5.06	—	—	—	—	—
7	6.12	7.48	—	—	—	—	—
10	9.00	11.0	—	—	—	—	—
15	13.5	16.5	—	—	—	—	—
22	19.8	24.2	5W	—	—	—	—
32	28.8	35.2	10W	—	—	—	—
46	41.4	50.6	15W	—	75W	1	
68	61.2	74.8	20W	—	—	2	2 EP
100	90.0	110	30	65	80W-90	3	3 EP
150	135	165	40	80	—	4	4 EP
220	198	242	50	100	90	5	5 EP
320	288	352	60	120	—	6	6 EP
460	414	506	—	—	85W-140	7 comp	7 EP
680	612	748	—	—	—	8 comp	8 EP
1000	900	1100	—	—	—	8A comp	8A EP
1500	1350	1650	—	—	250	—	—

[a] Comparisons are nominal since SAE grades are not specified at 40°C viscosity; VI of lubes could change some of the comparisons.

Oil viscosity decreases significantly with increasing temperature as shown in Fig. 3.10.4. While Figure 3.10.4 provides viscosity data suitable for most bearing calculations, oil suppliers commonly provide only the 40°C and 100°C values of kinematic viscosity in centistokes (mm^2/sec). The viscosity at other temperatures can be found by the following ASTM D341 equation relating kinematic viscosity ν in centistokes (mm^2/sec) to temperature T in degrees F:

$$\log\log(\nu + 0.7) = A - B\log(460 + T) \tag{3.10.2}$$

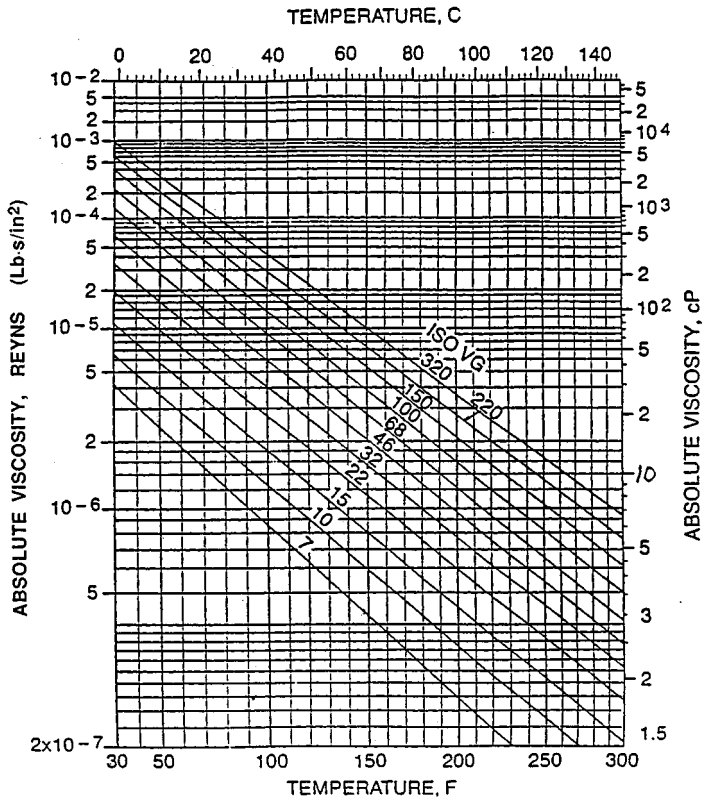


FIGURE 3.10.4 Viscosity–temperature chart for industrial petroleum oils (Ramondi and Szeri, 1984).

where A and B are constants for any particular (petroleum or synthetic) oil. For ISO VG-32 oil in Table 3.10.4, for example, A = 10.54805 and B = 3.76834 based on the 104 and 212°F (40 and 100°C) viscosities. This gives a viscosity of 10.78 cSt at a bearing operating temperature of 160°F.

Conversion of kinematic to absolute viscosity requires use of density, which also varies with temperature. A coefficient of expansion of 0.0004/°F is typical for most petroleum oils, so the variation of density with temperature is in accordance with the following relation:

$$\rho_T = \rho_{104}\left[1 - 0.0004(T - 104)\right] \tag{3.10.3}$$

For the ISO-VG-32 oil with its density of 0.857 g/cc at 100°F (40°C), this equation gives 0.838 g/cc at 160°F. Multiplying 10.78 cSt by 0.838 then gives an absolute viscosity of 9.03 cP. Multiplying by the conversion factor of 1.45×10^{-7} gives 1.31×10^{-6} lb·sec/in.^{2} (reyns).

Heat capacity C_p of petroleum oils used in bearing calculations is given by (Klaus and Tewksburg, 1984)

$$C_p = 3860 + 4.5(T) \quad \text{in.} \cdot \text{lb}/(\text{lb} \cdot °\text{F}) \tag{3.10.4}$$

The viscosity of petroleum oils is also affected significantly by pressure, and that increase can become important in concentrated contacts such as rolling element bearings where elastohydrodynamic lubrication occurs. The following relationship can be used for many oils to find the viscosity at elevated pressure:

$$\mu_p = \mu_o e^{\alpha p} \tag{3.10.5}$$

where μ_o is the viscosity at atmospheric pressure, μ_p is the viscosity at pressure p, and α is the pressure–viscosity exponent. The pressure–viscosity exponent can be obtained from lubricant suppliers.

Synthetic Oils

Synthetic oils of widely varying characteristics are finding increasing use for applications at extreme temperatures and for their unique physical and chemical characteristics. Table 3.10.6 gives a few representative examples of synthetic oils which are commercially available. Cost of the synthetics, ranging up to many times that of equivalent petroleum oils, is a common deterrent to their use where petroleum products give satisfactory performance

TABLE 3.10.6 Properties of Representative Synthetic Oils

Type	Viscosity, cSt at 100°C	Viscosity, cSt at 40°C	Viscosity, cSt at –54°C	Pour Point, °C	Flash Point, °C	Typical Uses
Synthetic hydrocarbons						
Mobil 1, 5W-30[a]	11	58	—	–54	221	Auto engines
SHC 824[a]	6.0	32	—	–54	249	Gas turbines
SHC 629[a]	19	141	—	–54	238	Gears
Organic esters						
MIL-L-7808	3.2	13	12,700	–62	232	Jet engines
MIL-L-23699	5.0	24	65,000	–56	260	Jet engines
Synesstic 68[b]	7.5	65	—	–34	266	Air compressors, hydraulics
Polyglycols						
LB-300-X[c]	11	60	—	–40	254	Rubber seals
50-HB-2000[c]	70	398	—	–32	226	Water solubility
Phosphates						
Fyrquel 150[d]	4.3	29	—	–24	236	Fire-resistant fluids for die casting,
Fyrquel 220[d]	5.0	44	—	–18	236	air compressors and hydraulic systems
Silicones						
SF-96 (50)	16	37	460	–54	316	Hydraulic and damping fluids
SF-95 (1000)	270	650	7,000	–48	316	Hydraulic and damping fluids
F-50	16	49	2,500	–74	288	Aircraft and missiles
Fluorochemicals						
Halocarbon 27[e]	3.7	30	—	–18	None	Oxygen compressors, liquid-oxygen
Krytox 103[f]	5.2	30	—	–45	None	systems

[a] Mobil Oil Corp.
[b] Exxon Corp.
[c] Union Carbide Chemials Co.
[d] Akzo Chemicals
[e] Halocarbon Products Corp.
[f] DuPont Co.

Greases

Grease is essentially a suspension of oil in a thickening agent, along with appropriate additives. The oil generally makes up between 75 and 90% of the weight of a grease, and it is held in place by the gel structure of the thickener to carry out its lubricating function. The simplicity of the lubricant supply system, ease of sealing, and corrosion protection make grease the first choice for many ball-and-roller bearings, small gear drives, and slow-speed sliding applications (Booser, 1995). Consistencies of greases vary from soap-thickened oils that are fluid at room temperature to hard brick-types that may be cut with a knife. Greases of NLGI Grade 2 stiffness (ASTM D217) are most common. Softer greases down to grade 000 provide easier feeding to multiple-row roller bearings and gear mechanisms. Stiffer Grade 3 is used in some prepacked ball bearings to avoid mechanical churning as the seals hold the grease in close proximity with the rotating balls.

Petroleum oils are used in most greases; the oils generally are in the SAE 30 range, with a viscosity of about 100 to 130 cSt at 40°C. Lower-viscosity oil grades are used for some high-speed applications and for temperatures below about –20°C. Higher-viscosity oils are used in greases for high loads and low speeds. Synthetic oils are used only when their higher cost is justified by the need for special properties, such as capability for operation below –20°C or above 125 to 150°C.

The most common gelling agents are the fatty acid metallic soaps of lithium, calcium, sodium, or aluminum in concentrations of 8 to 25%. Of these, the most popular is lithium 12-hydroxystearate; greases based on this lithium thickener are suitable for use at temperatures up to 110°C, where some lithium soaps undergo a phase change. Greases based on calcium or aluminum soaps generally have an upper temperature limit of 65 to 80°C, but this limit can be significantly raised to the 120 to 125°C range through new complex soap formulations. Calcium-complex soaps with improved high-temperature stability, for instance, are prepared by reacting both a high-molecular-weight fatty acid (e.g., stearic acid) and a low-molecular-weight fatty acid (acetic acid) with calcium hydroxide dispersed in mineral oil.

Inorganic thickeners, such as fine particles of bentonite clay, are inexpensively applied by simple mixing with oil to provide nonmelting greases for use up to about 140°C. Polyurea nonmelting organic powders are used in premium petroleum greases for applications up to about 150 to 170°C.

Additives, such as those mentioned in the subsection on petroleum oils, are added to grease to improve oxidation resistance, rust protection, or extreme pressure properties. Because of the incompatibility of oils, thickeners, and additives, greases of different types should be mixed only with caution.

Solid Lubricants

Solid lubricants provide thin solid films on sliding or rolling/sliding surfaces to reduce friction and wear. They are particularly useful for applications involving high operating temperatures, vacuum, nuclear radiation, or other environments which limit the use of oils or greases. Solid lubricant films do not prevent moving surfaces from contacting each other, so they cannot eliminate wear and their lives are limited by wear. The properties of some of the most common solid lubricants are given in Table 3.10.7.

TABLE 3.10.7 Properties of Selected Solid Lubricants

Material	Acceptable Usage Temperature, °C				Average Friction Coefficient, *f*		Remarks
	Minimum		Maximum				
	In Air	In N_2 or Vacuum	In Air	In N_2 or Vacuum	In Air	In N_2 or Vacuum	
Molybdenum disulfide, MoS_2	–240	–240	370	820	0.10–0.25	0.05–0.10	Low *f*, carries high load, good overall lubricant, can promote metal corrosion
Graphite	–240	—	540	Unstable in vacuum	0.10–0.30	0.02–0.45	Low *f* and high load capacity in air, high *f* and wear in vacuum, conducts electricity
PTFE	–70	–70	290	290	0.02–0.15	0.02–0.15	Lowest *f* of solid lubricants, load capacity moderate and decreases at elevated temperature
Calcium flouride–barium fluoride eutectic, CaF_2–BaF_2	430	430	820	820	0.10–0.25 above 540°C 0.25–0.40 below 540°C	Same as in air	Can be used at higher temperature than other solid lubricants, high *f* below 540°C

Modified from Booser, E.R., in *Encyclopedia of Chemical Technology*, 4th ed., Vol. 15, John Wiley & Sons, New York, 1995, 463–517.

The most important inorganic solid lubricants are layer–lattice solids such as molybdenum disulfide (MoS_2) and graphite. These materials are characterized by strong covalent or ionic bonding between atoms in individual layers, but relatively weak van der Waals bonds between layers, enabling the layers

to slide easily relative to one another. Graphite is a very effective lubricant film when moisture or water vapor is present, since adsorbed water vapor lubricates the sliding layers, but it has poor friction properties in vacuum or other low-humidity applications. Molybdenum disulfide does not require the presence of adsorbed water vapor, so it is widely used in vacuum or space applications.

The most commonly used organic solid lubricant is polytetrafluoroethylene (PTFE) which can be used either as a fused surface coating or as a self-lubricating material (see subsection on plastics). Its low friction is attributed to the smooth profile of the PTFE molecule. The chemical inertness of PTFE makes it attractive for chemical and food-processing applications.

New ceramic-based solid lubricants have been developed for high-temperature applications, such as heat engines or space uses. One of the most promising of these is a calcium fluoride–barium fluoride eutectic, which can be used at temperatures exceeding 800°C.

Fluid Film Bearings

Journal Bearings

A journal bearing consists of an approximately cylindrical bearing body or sleeve around a rotating cylindrical shaft. In general, journal bearings are found in motors, pumps, generators, appliances, and internal combustion engines in which a fluid lubricant is used; and in smaller mechanisms such as switches, clocks, small motors, and circuit breakers in which a solid lubricant such as graphite, grease, or certain plastics serves to reduce friction. Air (gas) bearings are designed to utilize both fluid mechanics principles when operating and solid lubricant–surfaced materials for start, stop, and emergency operations.

A hydrodynamic journal bearing maintains separation of shaft from bearing because the lubricant viscosity and the speed of the shaft create pressure in the converging portion of the fluid film which carries load. The governing equations were first developed by Reynolds (1886). Their solution has led to numerous computer solutions, including those used for this section.

Journal Bearing Design. Figure 3.10.5 shows schematics of frequently used types of journal bearing in which one or more lobes of cylindrical shape are positioned around the shaft, their axis being assumed parallel to the shaft axis. The features of each design and applications where it is often found are listed in Table 3.10.8.

Noncontact journal bearings are designed to assure a continuous supply of lubricant to the load-carrying section, and the bearing grooves in Figure 3.10.5 are designed for that purpose. Oil must be resupplied to the bearing because of the continuous loss of lubricant as it is forced from the bearing by the load-carrying pressures generated within it. The subsection on lubricant supply methods describes some of the many systems designed to assure this supply and to cool the lubricant at the same time.

Controlling Variables. Definitions of the variables involved in journal bearing analysis are contained in Table 3.10.9. Because of the large range of many variables, nondimensional quantities are often used which are independent of the dimensional unit system involved. Examples are given in the English system unless otherwise stated.

Calculating Bearing Performance. Journal bearing performance can be calculated directly from dedicated computer codes which take account of load, speed, oil type, and delivery system, as well as bearing dimensions. This subsection presents two approximate solutions: a simple thermal approach and a set of interpolation tables based on computer solutions.

Thermal Approach. It is assumed that the bearing is operating at a constant but elevated temperature. A predicted operating temperature can then be found as a function of an assumed viscosity. A solution is found when the assumed viscosity equals the lubricant viscosity at that temperature. Three equations are used for this method. For radial loads, the power dissipation is

$$H_p = j\pi^3 \mu \left(N/60\right)^2 D^3 L/C \quad \text{in.} \cdot \text{lb/sec} \tag{3.10.6}$$

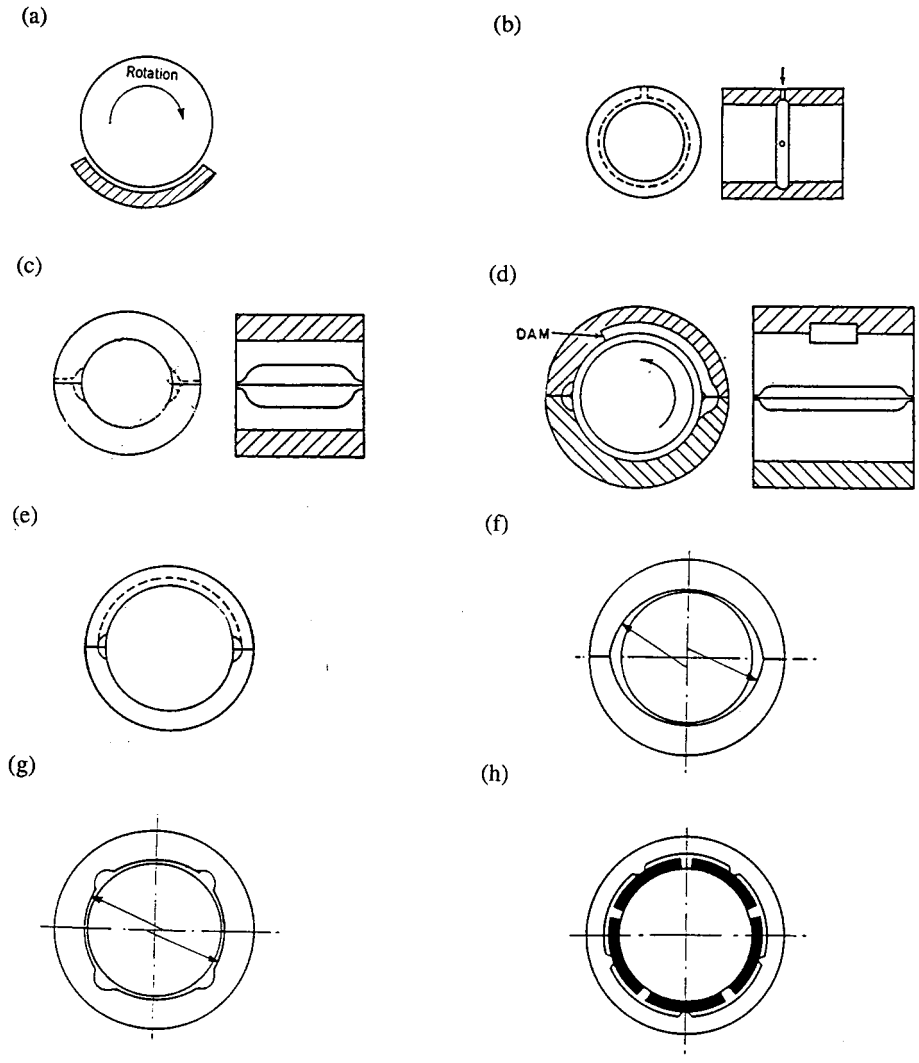


FIGURE 3.10.5 Types of pressure-fed journal bearings: (a) Partial arc. (b) Circumferential groove. (c) Cylindrical bearing–axial groove. (d) Pressure dam. (e) Cylindrical overshot. (f) Elliptical. (g) Multilobe. (h) Tilting pad.

TABLE 3.10.8 Journal Bearing Designs

Type	Typical Loading	Applications
Partial arc	Unidirectional load	Shaft guides, dampers
Circumferential groove	Variable load direction	Internal combustion engines
Axial groove types		
Cylindrical	Medium to heavy unidirectional load	General machinery
Pressure dam	Light loads, unidirectional	High-speed turbines, compressors
Overshot	Light loads, unidirectional	Steam turbines
Multilobe	Light loads, variable direction	Gearing, compressors
Preloaded	Light loads, unidirectional	Minimize vibration
Tilting pad	Moderate variable loads	Minimize vibration

where $j = 1$ for a shaft-centered design. The lubricant flow rate is

TABLE 3.10.9 Journal Bearing Parameters

B	Bearing damping coefficient	lb/in./sec.
C	Radial clearance	in.
C_α	Adiabatic constant	—
C_p	Heat capacity	in.·lb/lb°F
D	Diameter	in.
H_p	Power loss	in.·lb/sec
K	Bearing stiffness	lb/in.
L	Bearing length	in.
N	Shaft velocity	rpm
Q	Lubricant flow rate	$in.^3$/sec
R	Shaft radius	in.
R_e	Reynolds number	—
T_e	Entrance temperature	°F
T_f	Feed temperature	°F
T_q	Torque	in.·lb
ΔT_b	Temperature rise coefficient, bottom half	°F
ΔT_t	Temperature rise coefficient, top half	°F
U	Velocity	in./sec
W	Load	lb
e	Shaft eccentricity	in.
h	Film thickness	in.
j	Ratio: power loss/shaft-centered loss	—
p	Pressure	psi
q	Flow coefficient	—
w	Load coefficient	—
x	Coordinate in direction of load	in.
y	Coordinate orthogonal to load	in.
β	Exponential temperature coefficient of viscosity	—
ε	Shaft eccentricity, nondimensional	—
γ	Angular extent of film	—
ϕ	Attitude angle	—
ρ	Density	$lb/in.^3$
μ	Viscosity	$lb\cdot sec/in.^2$
ω	Shaft velocity	rad/sec
θ	Angle in direction of rotation, from bottom dead center	—
Φ	Energy dissipation	in.·lb/sec

$$Q = Q_o + qCR^2\omega/2 \quad in.^3/sec \tag{3.10.7}$$

where q is the proportion of side flow to circulating flow, and the zero speed flow, Q_o (in^3/sec), represents other flows such as from the ends of the feed grooves which are not related to the load-carrying film itself. Q_o can usually be neglected for rough estimation, and this approximation is useful for eccentricities as high as 0.7. Note that both q and j are functions of specific design as well as load and speed. The average operating temperature for a given viscosity is

$$T_2 = T_f + \frac{\left(H_p - \Phi\right)}{\left(\rho C_p Q\right)} \quad °F \tag{3.10.8}$$

where T_f is the feed temperature and Φ is the energy loss due to conduction and radiation. For diameters of 2″ or more, Φ can usually be assumed to be 0. Plotting T_2 vs. viscosity for several values of μ on a plot of the viscosity vs. T for the lubricant shows the operating temperature for the bearing as the intersection.

Flow Dynamics Solution. A more general solution for journal bearing performance is based on prediction of flow characteristics in the bearing, and of physical behavior of the bearing based on the Reynolds equation. A common two-pad journal bearing with pressurized oil feed will be used to provide specific design information. The Reynolds equation is the differential equation expressing the physical behavior of the fluid film between shaft and bearing, as written for cylindrical bearings:

$$1/R^2\left[\partial/\partial\Theta\left(h^3/\mu\right)\partial p/\partial\Theta\right]+\partial/\partial z\left(h^3/\mu\right)\partial p/\partial z=6\left(U/R\right)\partial h/\partial\Theta \tag{3.10.9}$$

where z is the axial coordinate and Θ is the angular coordinate.

Bearing Configuration. A cross section through a common type of two-pad cylindrical bearing is shown in Figure 3.10.6. Two pads having a radius of $R + C$ and an angular extent of 150°, and with load applied vertically to the lower half through a shaft, make up the bearing. Lubricant is admitted under moderate feed pressure to the two 30° grooves formed at the split as shown in Figure 3.10.6. The shaft rotates counterclockwise, and lubricant pressures are balanced when the shaft center is displaced down and to the right.

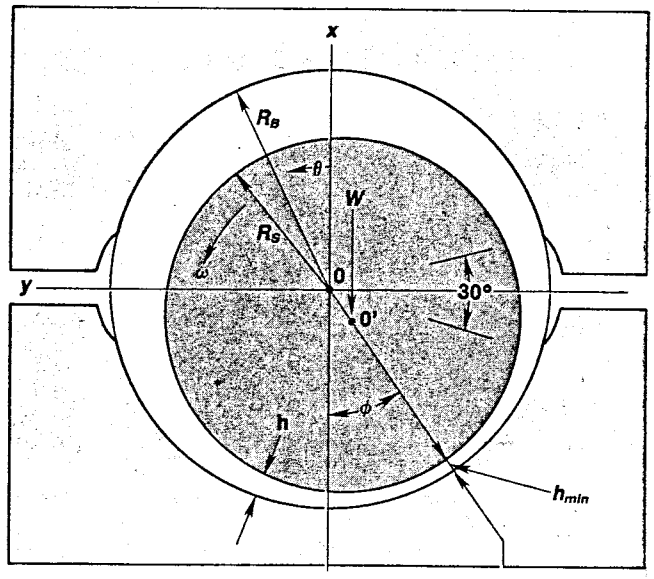


FIGURE 3.10.6 Geometry of split cylindrical journal bearing.

Lubricant Properties. Pressures in the lubricant film are generated as a function of the local shear rates and the local viscosity as described by the Reynolds equation. The local temperature rise is given by the local energy dissipation divided by the local flow rate:

$$\Delta T=\left[2\mu\omega R^2\Delta\theta\right]/\left[h^2\rho C_p\right] \tag{3.10.10}$$

As an alternative to Equation 3.10.2, an exponential relation between viscosity and temperature is used:

$$\mu=\mu_0 e^{-\beta\left(T-T_0\right)} \tag{3.10.11}$$

Assuming an ISO 32 oil, viscosity μ_0 at 104°F is 3.98 reyns, density is 0.0310, and β is 0.0170 for the range from 104 to 212°F. The value for β may be determined from Figure 3.10.4. A nondimensional

coefficient, C_α, the *adiabatic coefficient*, is used as an indicator of the severity of thermal heating in a design. It contains the several factors responsible for temperature rise.

$$C_\alpha = 2\mu_f \beta\omega\left(R/C\right)^2 / \rho C_p \tag{3.10.12}$$

Bearing Performance Tables. Using computer-generated solutions of the Reynolds equation for the pressure field over the bearing surface, the relevant performance properties of the bearing/lubricant system can be determined. A number of programs are commercially available for this purpose. In order to illustrate how the behavior of a journal bearing can be determined, a two-pad split cylindrical bearing with an *L*/*D* of 0.5 has been selected, and a proprietary computer code has been used to develop a group of performance quantities for such a bearing. The code accounts for the internal lubricant circulation, including mixing in the inlet groove of feed lubricant with warm lubricant from the upper half, resulting in the viscosity μ_f. The primary characteristics of the bearing are load, stiffness, and damping. Each of these factors is presented in nondimensional terms in Table 3.10.10, and the corresponding dimensional and other quantities are given as follows:

Film thickness, in. $$h = C\left(1 + e\cos\Theta\right) \tag{3.10.13a}$$

Shaft eccentricity, in. $$e = \varepsilon C \tag{3.10.13b}$$

Load, lb $$W = w\left[6\mu\omega\left(R/C\right)^2 DL\right] \tag{3.10.13c}$$

Flow, $\text{in.}^3/\text{sec}$ $$Q = q\left(CR^2\,\omega/2\right) \tag{3.10.13d}$$

Power loss, in. · lb/sec $$H_p = j\left(2\pi\mu\omega^2 R^3\, L/C\right) \tag{3.10.13e}$$

Stiffness, lb/in. $$S_{xx} = K_{xx} 6\mu\omega R\left(R/C\right)^2 \tag{3.10.13f}$$

Damping, lb · sec/in. $$D_{xx} = B_{xx} 12\mu R\left(R.C\right)^3 \tag{3.10.13g}$$

The axial length/diameter ratio also influences the performance of a journal bearing. To illustrate this, Table 3.10.11 presents the performance of longer bearings (*L*/*D* = 0.75 and 1.00) for comparison to the more common *L*/*D* = 0.5 results in Table 3.10.10.

Comparing Tables 3.10.10 and 3.10.11, the use of longer bearings has several important effects on operation and performance. Comparing key variables, the effects at an eccentricity of ratio of 0.7 are as follows:

Variable	L/D = 0.5	L/D = 0.75	L/D = 1.00
Load, w	0.28	0.69	1.21
Flow, q	0.69	0.82	0.88
Attitude angle, ϕ	36.4	36.1	35.8
Power ratio, j	1.00	1.15	1.17
Stiffness, K_{xx}	1.38	3.06	5.03
Damping, B_{xx}	0.99	2.52	4.31

Effect of Turbulence. Turbulence is a mixing phenomenon that occurs in larger high-speed bearings. When this behavior occurs, the simple viscous drag behavior that has been assumed in the preceding

TABLE 3.10.10 Performance of *L/D* = 0.5 Bearing

Part 1: C_α = 0.0

ε	0.2	0.5	0.7	0.8	0.9	0.95
ϕ	66.5	48.01	36.44	30.07	22.18	16.46
w	0.0246	0.0997	0.2784	0.5649	1.6674	4.4065
q	0.3037	0.6014	0.6927	0.6946	0.6487	0.588
j	0.7779	0.8534	1.1005	1.3905	2.008	3.084
ΔT_b	0	0	0	0	0	0
ΔT_t	0	0	0	0	0	0
K_{xx}	0.041	0.2805	1.379	4.063	22.67	—
K_{xy}	0.1465	0.3745	1.063	2.476	9.390	34.47
K_{yx}	–0.055	–0.072	0.0063	0.193	1.710	8.002
K_{yy}	0.046	0.170	0.4235	0.883	2.622	7.555
B_{xx}	0.142	0.352	0.989	2.311	8.707	32.30
B_{xy}, B_{yx}	0.023	0.094	0.236	0.522	1.547	4.706
B_{yy}	0.056	0.105	0.174	0.302	0.630	1.390

Part 2: C_α = 0.1

ε	0.2	0.5	0.7	0.8	0.9	0.95
ϕ	69.9	50.2	38.7	32.35	24.83	19.8
w	0.022	0.087	0.233	0.451	1.184	2.621
q	0.312	0.620	0.721	0.728	0.692	0.642
j	0.686	0.723	0.863	0.997	1.253	1.545
ΔT_b	0.274	0.403	0.642	0.907	1.519	2.346
ΔT_t	0.243	0.211	0.183	0.168	0.151	0.142
K_{xx}	0.038	0.2365	1.041	2.935	13.66	50.44
K_{xy}	0.126	0.3135	0.870	1.851	3.078	18.30
K_{yx}	–0.047	–0.061	–0.021	0.139	1.068	3.961
K_{yy}	0.037	0.140	0.3585	0.669	1.784	4.327
B_{xx}	0.121	0.286	0.776	1.592	4.97	14.00
B_{xy}, B_{yx}	0.016	0.071	0.195	0.341	0.850	2.10
B_{yy}	0.047	0.086	0.156	0.216	0.394	0.757

Part 3: C_α = 0.2

ε	0.2	0.5	0.7	0.8	0.9	0.95
ϕ	73.4	52.2	40.8	34.55	27.23	22.5
w	0.020	0.077	0.198	0.368	0.890	1.779
q	0.320	0.639	0.747	0.759	0.730	0.760
j	0.613	0.628	0.712	0.791	0.933	1.092
ΔT_b	0.520	0.7520	1.162	1.594	2.521	3.651
ΔT_t	0.472	0.415	0.363	0.333	0.301	0.284
K_{xx}	0.035	0.1925	0.830	2.156	8.86	28.6
K_{xy}	0.11	0.272	0.704	1.477	4.515	11.72
K_{yx}	–0.041	–0.062	–0.018	0.074	0.640	2.371
K_{yy}	0.029	0.125	0.2895	0.551	1.375	2.932
B_{xx}	0.104	0.242	0.596	1.21	3.90	7.830
B_{xy}, B_{yx}	0.011	0.061	0.140	0.212	0.634	1.21
B_{yy}	0.040	0.080	0.121	0.187	0.326	0.501

TABLE 3.10.10 (continued) Performance of *L/D* = 0.5 Bearing

Part 4 C_α = 0.4

ε	0.2	0.5	0.7	0.8	0.9	0.95
ϕ	80.2	56.0	44.5	38.4	31.3	26.7
w	0.016	0.061	0.148	0.260	0.562	1.000
q	0.331	0.6720	0.795	0.815	0.797	0.760
j	0.504	0.498	0.534	0.570	0.637	0.716
ΔT_b	0.946	1.33	1.97	2.61	3.87	5.26
ΔT_t	0.898	0.801	0.712	0.658	0.597	0.562
K_{xx}	0.029	0.137	0.538	1.295	4.56	12.6
K_{xy}	0.085	0.206	0.503	0.985	2.67	6.17
K_{yx}	–0.0315	–0.0548	0.0298	0.0233	0.321	1.136
K_{yy}	0.019	0.094	0.214	0.382	0.860	1.68
B_{xx}	0.079	0.175	0.397	0.734	1.75	3.44
B_{xy}, B_{yx}	0.0041	0.042	0.094	0.166	0.329	0.55
B_{yy}	0.030	0.064	0.092	0.131	0.120	0.276

TABLE 3.10.11 Performance of Long Bearings

Part 1: *L/D* = 0.75, C_α = 0.0

ε	0.2	0.5	0.7	0.8	0.9	0.95
ϕ	64.74	46.54	36.13	30.17	22.64	17.03
w	0.0705	0.2714	0.6947	1.311	3.440	8.241
q	0.392	0.738	0.825	0.811	0.737	0.6545
j	0.777	0.871	1.145	1.450	2.184	3.233
ΔT_b	0	0	0	0	0	0
ΔT_t	0	0	0	0	0	0
K_{xx}	0.121	0.706	3.065	8.506	41.5	—
K_{xy}	0.418	0.992	2.517	5.228	18.1	59.0
K_{yx}	–0.123	–0.189	0.052	0.404	3.18	16.2
K_{yy}	0.113	0.429	1.012	1.891	5.33	13.49
B_{xx}	0.423	0.982	2.52	5.16	17.7	54.4
B_{xy}, B_{yx}	0.057	0.249	0.609	1.10	3.24	7.58
B_{yy}	0.127	0.263	0.444	0.641	1.35	2.32

Part 2: *L/D* = 1.00, C_α = 0.00

ε	0.2	0.5	0.7	0.8	0.9	0.95
ϕ	63.2	45.3	35.8	30.3	22.9	17.4
w	0.138	0.506	1.214	2.18	5.34	12.15
q	0.444	0.800	0.879	0.856	0.769	0.679
j	0.782	0.886	1.174	1.768	2.250	3.323
ΔT_b	0	0	0	0	0	0
ΔT_t	0	0	0	0	0	0
K_{xx}	0.234	1.254	5.026	13.24	60.9	—
K_{xy}	0.818	1.795	4.142	8.12	26.8	83.5
K_{yx}	–0.201	–0.313	–0.075	0.671	4.96	24.9
K_{yy}	0.198	0.732	1.64	2.95	8.04	19.5
B_{xx}	0.82	1.87	4.31	8.27	26.5	75.9
B_{xy}, B_{yx}	0.10	0.45	0.97	1.68	4.78	10.36
B_{yy}	0.21	0.46	0.70	0.98	2.02	3.24

subsections is broken up by numerous eddies which increase the drag. The Reynolds number is a nondimensional quantity that expresses this factor:

$$R_e = hU\rho/\mu \tag{3.10.14}$$

where h is the local film thickness, U is the relative velocity of one surface with respect to the other, ρ is the fluid density, and μ is the local viscosity.

The influence of turbulence on an $L/D = 0.5$ bearing is shown in Table 3.10.12. Examination of Table 3.10.12 shows that the principal effects of operation in the turbulent regime with a Reynolds number above about 1000 are in the greater power required (j) and the maximum bearing temperature. Load capacity and flow are only moderately affected.

TABLE 3.10.12. Influence of Turbulence ($\varepsilon = 0.7$, $C_\alpha = 0.2$, arc = 150°)

R_e	0	1000	2000	4000
ϕ	40.8	43.8	46.4	49.2
w	0.198	0.171	0.197	0.221
q	0.747	0.809	0.862	0.914
j	0.712	0.983	1.459	2.124
ΔT_b	1.162	0.585	0.918	1.404
K_{xx}	0.830	0.627	0.634	0.647
K_{xy}	0.704	0.575	0.577	0.645
K_{yx}	–0.018	–0.034	–0.047	–0.078
K_{yy}	0.289	0.257	0.282	0.330
B_{xx}	0.596	0.483	0.513	0.534
B_{xy}, B_{yx}	0.140	0.125	0.132	0.136
B_{yy}	0.121	—	—	0.104

Example Calculation. The problem is to design a two-pad cylindrical bearing for support of a rotor having a bearing load of 8000 lb, a shaft diameter of 6 in., and a speed of 3600 rpm. Assume use of ISO VG-32 oil fed at a temperature of 120°F. Compute operating characteristics for a 3-in.-long bearing. Assume a radial clearance of 0.0044 in.

Feed viscosity, $\mu_f = 3.98 \times 10^{-6}\, e^{-0.00170(120-104)} = 3.03 \times 10^{-6}$ reyn
Angular velocity, $\omega = 3600 \times 2\pi/60 = 377$ rad/sec
Adiabatic coefficient: $C_\alpha = 2 \times 3.03 \times 10^{-6} \times 0.0170 \times 377 \times (3/0.0044)^2/0.0310/4320 = 0.1345$
Load coefficient (from Equation 3.10.13c: $w = 8000/[6 \times 3.03 \times 10^{-6} \times 377 \times 3 \times 6 \times (33/0.0044)^2] = 0.139$

The desired solution lies between Part 2 and Part 3 of Table 3.10.10. By using linear interpolation between the tabulated values for C_α of 0.1 and 0.2, and values of ε of 0.7 and 0.8, an approximate operating point of $C_\alpha = 0.1345$ yields the following coefficients: $\varepsilon = 0.729$, $w = 0.279$, $q = 0.733$, $j = 0.860$, and $\Delta T_b = 0.915$.

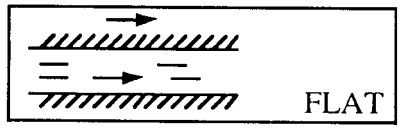
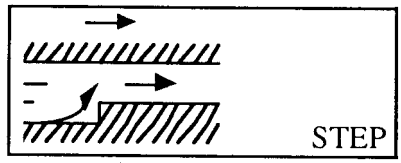
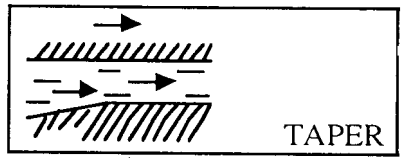
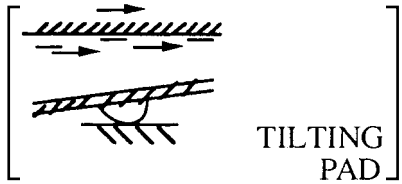
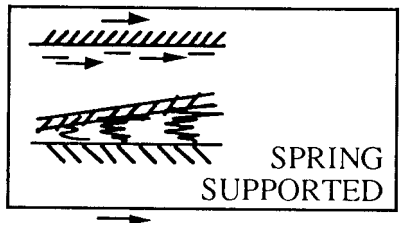
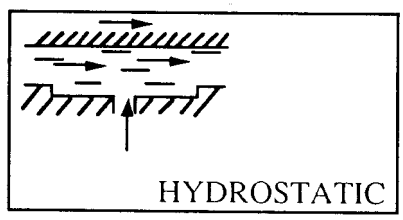
By using Equations 3.10.13, the dimensional operating results are:

Shaft velocity: $\omega = 3600 \times 2\pi/60 = 377$ rad/sec
Flow: $Q = 0.733 \times 0.0044 \times 3^2 \times 377 = 5.47$ in^3/sec
Power loss: $H_p = 0.860 \times 2\pi \times 3.03 \times 10^{-6} \times 377^2 \times 3^3 \times 3/0.0044 = 42.8$ in.·lb/sec
Oil temperature: $T_b = 120 + 0.915/0.0170 = 174$°F

Thrust Bearings

Types of Thrust Bearings. Oil film thrust bearings range from coin-size flat washers to sophisticated assemblies many feet in diameter. Of the six common types of thrust bearings shown in Table 3.10.13, the first five are hydrodynamic. As with journal bearings, each of these generates oil film pressure when

TABLE 3.10.13 Common Thrust Bearings and Their Range of Application

Type	O.D., in.	Unit Load, psi
	0.5–20	20–100
	0.5–10	100–300
	2–35	150–300
	4–120	250–700
	50–120	350–700
	3–50	500–3000

Source: Booser, E.R. and Wilcock, D.F., *Machine Design*, June 20, 1991, 69–72. With permission.

a rotating thrust face pumps oil by shear into a zone of reduced downstream clearance. When thrust load increases, film thickness drops until a new balance is reached between inflow and outflow, raising pressure until the higher bearing load is balanced. The hydrostatic bearing uses a separate oil pump to supply the pressurized flow.

Flat-land bearings, simplest to fabricate and least costly, are the first choice for simple positioning of a rotor and for light loads in electric motors, appliances, pumps, crankshafts, and other machinery. They carry less load than the other types because flat parallel planes do not directly provide the required pumping action. Instead, their action depends on thermal expansion of the oil and warping of the bearing

material induced by heating from the passing oil film. The resulting slight oil wedge then gives a load rating of about 10 to 20% of that for the other types.

Step bearings also offer relatively simple design. With a coined or etched step, they lend themselves to mass production as small-size bearings and thrust washers. Step height for optimum load capacity approximately equals the minimum film thickness, often 0.001 in. or less. Circumferential length beyond the step is ideally 45% of the total bearing segment (Wilcock and Booser, 1956). Step thrust bearings are well suited for low-viscosity fluids such as water, gasoline, fuels, and solvents. Minimum film thickness in these applications is so small that features such as pivots and tapers are usually impractical. Step height must be small enough for good load capacity, yet large enough to accommodate some wear without becoming worn away. Step erosion by contaminants is sometimes a problem.

Tapered-land bearings provide reliable, compact designs for mid- to large-size high-speed machines such as turbines, compressors, and pumps. Taper height normally should be about one to three times the minimum film thickness. For greater load capacity and to minimize wear during starting, stopping, and at low speeds, a flat land is commonly machined at the trailing edge to occupy from 10 to 30% of the circumferential length of each segment. Because operation of these bearings is sensitive to load, speed, and lubricant viscosity, they are typically designed for the rated range of operating conditions for specific machines.

Tilting-pad thrust bearings are used increasingly in turbines, compressors, pumps, and marine drives in much the same range of applications as tapered-land designs. They usually have a central supporting pivot for each of their three to ten bearing segments. Each of these thrust pad segments is free to adjust its position to form a nearly optimum oil wedge for widely varying loads, speeds, and lubricants, and with rotation in both directions. A secondary leveling linkage system is commonly introduced to support the individual pads; this provides a further advantage over tapered-land designs by accommodating some misalignment. Off-the-shelf units are available to match rotor shaft diameters from about 2 to 12 in., and custom designs range up to 120 to 170 in. in outside diameter. Recent trends to increase load capacity have led to offsetting pivots from the circumferential midpoint of a pad to about 60% beyond the leading edge, to substituting copper for steel as the backing for a tin babbitt bearing surface, and to nonflooded lubrication to minimize parasitic power loss from useless churning of oil passing through the bearing housing.

Springs or other flexible supports for thrust segments are employed for bearings ranging up to 10 ft or more in outside diameter and carrying millions of pounds of thrust. This flexible mounting avoids the high load concentration encountered by pivots in supporting large tilting-pads. Rubber backing can provide this flexible mounting for smaller thrust pads.

Hydrostatic thrust bearings are used where sufficient load support cannot be generated by oil film action within the bearing itself. This may be the case with low-viscosity fluids, or for load support on an oil film at standstill and very low speeds. The fluid is first pressurized by an external pump and then introduced into pockets in the bearing surface to float the load. A compact hydrostatic thrust bearing can sometimes be constructed with a single pocket at the end of a rotor. Larger bearings usually use three or more pressurized pockets to resist misalignment or to support off-center loads. Hydraulic flow resistance in the supply line to each pocket, or constant flow to each pocket (as with ganged gear pumps) then provides any asymmetric pressure distribution needed to support an off-center load. Bearing unit load is commonly limited to about 0.5 (0.75 with fixed flow systems) times the hydrostatic fluid supply pressure—up to 5000 psi with conventional lubricating oils.

Design Factors for Thrust Bearings. In preliminary sizing, the inside diameter d of a thrust bearing is made sufficiently larger than the shaft to allow for assembly, and to provide for any required oil flow to the thrust bearing inside diameter (see Figure 3.10.7). This clearance typically ranges from about $^1/_8$ in. for a 2-in. shaft to $^1/_2$ in. for a 10-in. shaft. Bearing outside diameter D is then set to provide bearing

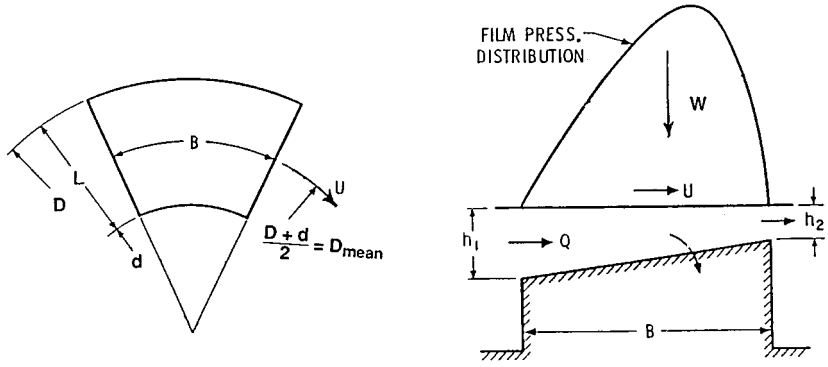


FIGURE 3.10.7 Sector for tapered land thrust bearing.

area sufficient to support total thrust load W (lb or N) with an average loading P (psi or N/m^2), using typical values from Table 3.10.13:

$$D = \left(\frac{4W}{\pi k_g P} + d^2 \right)^{0.5} \tag{3.10.15}$$

where k_g (typically 0.80 to 0.85) is the fraction of area between d and D not occupied by oil-distributing grooves. This bearing area is then divided by radial oil-feed groove passages, usually into "square" sectors with circumferential breadth B at their mean diameter equal to their radial length L.

While square pads usually produce optimum oil film performance, other proportions may be advantageous. With very large bearing diameters, for instance, square sectors may involve such a long circumferential dimension that the oil film overheats before reaching the next oil-feed groove. With a radially narrow thrust face, on the other hand, square sectors may become so numerous as to introduce excessive oil groove area, and their short circumferential length would interfere with hydrodynamic oil film action.

Performance Analysis. Performance analyses for sector thrust bearings using a fixed taper also hold approximately for most thrust bearing shapes (other than flat lands) with the same ratio of inlet to outlet oil film thickness (Wilcock and Booser, 1956; Neale, 1970; Fuller, 1984). Both for centrally pivoted pad thrust bearings and for spring-supported thrust bearings, use of an inlet-to-outlet film thickness ratio of two is usually appropriate in such an analysis.

While computer analyses in polar coordinates and with local oil film viscosity calculated over the whole oil film area gives more-exact solutions, the following constant viscosity approximations are made by relating a rectangular bearing sector (width B, length L) to the circular configuration of Figure 3.10.7. This rectangular representation allows more ready evaluation of a range of variables and gives results which are quite similar to a more accurate polar coordinate analysis.

Employing the nomenclature of Figure 3.10.7, the following relations give minimum film thickness h_2, frictional power loss H, and oil flow into a sector Q. The dimensionless coefficients involved are given in Table 3.10.14 for a range of sector proportions L/B and ratios of inlet to outlet film thicknesses h_1/h_2.

$$h_2 = K_h \left(\mu UB/P \right)^{0.5} \qquad \text{in. (m)} \tag{3.10.16a}$$

$$H = K_f \, \mu U^2 BL/h_2 \qquad \text{lb} \cdot \text{in./sec} \left(\text{N} \cdot \text{m/sec} \right) \tag{3.10.16b}$$

$$Q = K_q U L h_2 \qquad \text{in.}^3/\text{sec}\ \left(\text{m}^3/\text{sec}\right) \tag{3.10.16c}$$

$$\Delta T = H/\left(Q\rho C_p\right) = \frac{K_f}{K_q K_h^2 \rho C_p} P \quad °\text{F}\ (°\text{C}) \tag{3.10.16d}$$

where B = circumferential breadth of sector at mean radius, in. (m)
K_h, K_f, K_q = dimensionless coefficients
L = radial length of sector, in. (m)
P = unit loading on projected area of sector, W/BL, lb/in.^{2} (N/m^2)
U = surface velocity at mean diameter of sector, in./sec (m/sec)
W = load on a sector, lb (N)
C_p = oil specific heat, in.·lb/(lb·°F) (J/(kg·°C))
h_1, h_2 = leading edge and trailing edge film thicknesses, in. (m)
ρ = oil density, lb/in.^{3} (n/m^3)
μ = oil viscosity at the operating temperature, lb·sec/in.^{2} (N·sec/m^2)

Example. The following example involves a bearing for 25,000 lb thrust load in a 1200 rpm compressor whose rotor has a 5-in.-diameter journal. ISO-32 viscosity grade oil is to be fed at 120°F. Allowing $^3/_8$ in. radial clearance along the shaft sets thrust bearing bore $d = 5.75$ in. Taking unit loading $P = 300$ psi allows a margin for uncertainty in expected thrust load, and outside diameter D is given by Equation 3.10.14):

$$D = \left(\frac{4 \times 25000}{\pi(0.85)(300)} + 5.75^2\right)^{0.5} = 12.6 \text{ in.}$$

With 15% of the area used for oil feed passages, $k_g = 0.85$. Thrust bearing segment radial length $L = (12.6 - 5.75)/2 = 3.425$ in. With mean diameter $(5.75 + 12.6)/2 = 9.175$ in., total circumferential breadth of all pads at the mean diameter $= \pi D_m k_g = 24.5$ in. The number of sectors (and grooves) for $B = L$ is then $24.5/3.425 = 7.2$. Using seven lands, adjusted circumferential breadth B for each sector $= 24.5/7 = 3.5$ in. (For simpler fabrication, six or eight sectors should also be considered.) Runner velocity at the mean diameter, $U = \pi(9.175)(1200/60) = 576.5$ in./sec.

For square pads ($L/B = 1$) in a pivoted-paid bearing with $h_1/h_2 = 2$, which represents experience with centrally pivoted pads, Table 3.10.14 gives the following performance coefficients:

$$K_h = 0.261, \quad K_f = 0.727, \quad K_q = 0.849$$

Temperature rise is given by Equation (3.10.16d) which assumes that the total frictional power loss H goes into heating the total oil flow Q passing over the pad.

$$\Delta T = \frac{K_f P}{K_q K_h^2 \rho C_p} = \frac{0.727(300)}{0.849(0.261)^2(0.0313)(4535)} = 27°\text{F}$$

Adding this ΔT to the 120°F feed temperature gives 147°F as the representative oil operating temperature with a viscosity from Figure 3.10.4 of 1.6×10^{-6} lb·sec/in.^{2}. Temperature rise to the maximum oil film temperature would be expected to be about 53°F, twice the 27°F. If this bearing were in a housing fully flooded with oil, feed temperature to each pad would become the housing oil temperature, essentially the same as oil draining from the housing.

TABLE 3.10.14 Thrust Bearing Performance Characteristics

L/B	0.25	0.5	0.75	1.0	1.5	2.0	∞
			h_1/h_2 = 1.2				
K_h	0.064	0.115	0.153	0.180	0.209	0.225	0.266
K_f	0.912	0.913	0.914	0.915	0.916	0.917	0.919
K_q	0.593	0.586	0.579	0.574	0.567	0.562	0.549
			h_1/h_2 = 1.5				
K_h	0.084	0.151	0.200	0.234	0.275	0.296	0.351
K_f	0.813	0.817	0.821	0.825	0.830	0.833	0.842
K_q	0.733	0.714	0.696	0.680	0.659	0.647	0.610
			h_1/h_2 = 2				
K_h	0.096	0.170	0.223	0.261	0.305	0.328	0.387
K_f	0.698	0.708	0.718	0.727	0.739	0.747	0.768
K_q	0.964	0.924	0.884	0.849	0.801	0.772	0.690
			h_1/h_2 = 3				
K_h	0.100	0.173	0.225	0.261	0.304	0.326	0.384
K_f	0.559	0.579	0.600	0.617	0.641	0.655	0.696
K_q	1.426	1.335	1.236	1.148	1.024	0.951	0.738
			h_1/h_2 = 4				
K_h	0.098	0.165	0.212	0.244	0.282	0.302	0.352
K_f	0.476	0.503	0.529	0.551	0.581	0.598	0.647
K_q	1.888	1.745	1.586	1.444	1.242	1.122	0.779
			h_1/h_2 = 6				
K_h	0.091	0.148	0.186	0.211	0.241	0.256	0.294
K_f	0.379	0.412	0.448	0.469	0.502	0.521	0.574
K_q	2.811	2.560	2.273	2.013	1.646	1.431	0.818
			h_1/h_2 = 10				
K_h	0.079	0.121	0.148	0.165	0.185	0.195	0.221
K_f	0.283	0.321	0.353	0.377	0.408	0.426	0.474
K_q	4.657	4.182	3.624	3.118	2.412	2.001	0.834

After Khonsari, M.M., in *Tribology Data Handbook*, CRC Press, Boca Raton, FL, 1997.

Minimum film thickness h_2 becomes, from Equation (3.10.16a)

$$h_2 = 0.261\left[\left(1.6\times10^{-6}\right)(576.5)(3.5)/300\right]^{0.5} = 0.00086 \text{ in.}$$

With a fixed tapered land, rather than a centrally pivoted pad for which it could be assumed that h_1/h_2 = 2, several iterations might be required with different assumed values of the h_1/h_2 ratio in order to determine the performance coefficients in Table 3.10.14. The proper value of h_1/h_2 will be the one that gives the same final calculated value of h_2 from the above equation as was assumed in the preliminary selection of K_h, K_f, and K_q.

After finding the values for h_2 *and* K_f, the power loss H can be determined using Equation (3.10.16b). For this example the power loss would be H = 5510 lb·in./sec.

The total oil feed to the bearing should satisfy two requirements: (1) provide a full oil film over the bearing segment and (2) maintain reasonably cool operation with no more than 30 to 40°F rise in the

oil temperature from feed to drain. Equation (3.10.16c) can be used to find the oil feed Q needed at the sector inlet to form a full film. The oil feed needed for a 40°F rise is given by the following heat balance using typical density and specific heat values for petroleum oil:

$$Q = H \Big/ \left(\rho C_p \Delta T\right) \tag{3.10.17}$$

The required oil feed will be the larger of the values determined by (3.10.16c) and (3.10.17).

The above calculations are for a single sector; power loss and oil feed would be multiplied by the number of sectors (seven) to obtain values for the total bearing. Consideration would normally follow for other pad geometries, and possibly other lubricants and oil flow patterns, in a search for the most-promising design. More-detailed calculations of film thickness, film temperatures, oil flow, and power loss could then be obtained by one of a number of computer codes available from bearing suppliers or other sources.

Oil Film Bearing Materials

Selection of the material for use in a journal or thrust bearing depends on matching its properties to the load, temperature, contamination, lubricant, and required life.

Babbitts. Of the common bearing materials, listed in Table 3.10.15, first consideration for rotating machines is usually a babbitt alloy containing about 85% tin or lead together with suitable alloying elements. With their low hardness, they have excellent ability to embed dirt, conform to shaft misalignment, and rate highest for compatibility with steel journals. Tin babbitts, containing about 3 to 8% copper and 5 to 8% antimony, are usually the first choice for their excellent corrosion resistance. SAE 12 (ASTM Grade 2) tin babbitt is widely used in both automotive and industrial bearings. The much lower cost of lead babbitt, however, with 9 to 16% antimony and as much as 12% tin for improved corrosion resistance, brings SAE 13, 14, and 15 grades into wide use for both general automotive and industrial applications (Booser, 1992).

TABLE 3.10.15 Characteristics of Oil Film Bearing Materials

Material	Brinell Hardness	Load Capacity, psi	Max Operating Temp., °F	Compatibility[a]	Conformability and Embeddability[a]	Corrosion Resistance[a]	Fatigue Strength[a]
Tin babbitt	20–30	800–1500	300	1	1	1	5
Lead babbitt	15–25	800–1200	300	1	1	3	5
Copper lead	20–30	1500–2500	350	2	2	5	4
Leaded bronze	60–65	3000–4500	450	3	4	4	3
Tin bronze	65–80	5000+	500	5	5	2	2
Aluminum alloy	45–65	4000+	300	4	3	1	2
Zinc alloy	90–125	3000	250	4	5	5	3
Silver overplated	—	5000+	300	2	4	2	1
Two-component, babbitt surfaced	—	3000+	300	2	4	2	3
Three-component, babbitt surfaced	—	4000+	300	1	2	2	1

[a] Arbitrary scale: 1 = best, 5 = worst.

To achieve the high fatigue strength needed in reciprocating engines, only a very thin layer (commonly 0.001″) of babbitt is used so that much of the reciprocating load is taken on a stronger backing material (DeHart, 1984; Kingsbury, 1992). For bimetal bushings such as those used in automobile engines, oil

grooves and feed holes are formed in a continuous steel strip coated with babbitt. The strip is then cut to size and the individual segments are rolled into finished bearings.

For heavy-duty reciprocating engines, three-layer bearings are common. By using a steel strip backing, a thin overlay of SAE 19 or 190 lead babbitt is either electroplated or precision cast on an intermediate layer about 0.1 to 0.3″ thick of copper–nickel, copper–lead, leaded bronze, aluminum, or electroplated silver.

Copper Alloys. Copper–lead alloys containing 20 to 50% lead, either cast or sintered on a steel back, provide good fatigue resistance for heavy-duty main and connecting rod bearings for automotive, truck, diesel, and aircraft engines. The 10% lead–10% tin leaded bronze has been a traditional selection for bearings in steel mills, appliances, pumps, automotive piston pins, and trunions. This has been replaced in many applications by CA932 (SAE 660) containing 3% zinc for easier casting. The harder tin bronzes require reliable lubrication, good alignment, and 300 to 400 Brinell minimum shaft hardness. Cast tin bronze bushings are used at high loads and low speeds in farm machinery, earthmoving equipment, rolling mills, and in automotive engines for connecting rod bearings.

Utility of copper alloy bearings is limited to relatively low surface speeds by the tendency to form a copper transfer film on a steel shaft. Above about 1500 to 3000 ft/min, selective plucking of softer copper material from hotter load zones in the bearing may result in welded lumps forming on the cooler, stronger transfer layer on the mating steel shaft.

Zinc Alloys. Zinc alloys containing about 10 to 30% aluminum find some use for lower cost and better wear life in replacing leaded bronzes. They are used for both oscillating and rotating applications involving speeds up to 1400 ft/min and temperatures up to 250°F.

Aluminum Alloys (DeHart, 1984; Shabel et al., 1992). Although finding only minor use in general industrial applications because of their limited compatibility with steel journals, aluminum alloys containing 6.5% tin, 1% copper, and up to 4% silicon are used as solid, bimetal, and trimetal bearings in automotive engines, reciprocating compressors, and aircraft equipment. Good journal finish and shaft hardness of Rockwell B 85 or higher are required. The good fatigue and corrosion resistance of aluminum alloys have led to use of a number of unique alloys containing major additions of silicon, lead, or tin to provide better compatibility characteristics.

Dry and Semilubricated Bearings

Various plastics, porous bronze and porous iron, carbon–graphite, rubber, and wood are widely used for bearings operating dry or with sparse lubrication (Booser, 1992). Unique properties of these materials have led to their broad use in applications once employing oil film and ball and roller bearings. While these materials provide good performance under conditions of poor or nonexistent lubrication at low speeds, performance commonly improves the closer the approach to full film lubrication.

Plastics

Most commercial plastics find some use both dry and lubricated in slow-speed bearings at light loads (Jamison, 1994). The most commonly used thermoplastics for bearings are PTFE, nylon, and acetal resins. Thermosetting plastics used for bearings include phenolics, polyesters, and polyimides. Table 3.10.16 compares characteristics of typical plastic bearings with those of carbon–graphite, wood, and rubber which are used in similar applications.

In addition to the maximum temperature which can be tolerated, three operating limits shown in Table 3.10.16 are normally defined for plastic bearings: (1) maximum load at low speed, which reflects the compressive yield strength, (2) maximum speed for running under very light load, and (3) a $P\nu$ load-speed limit at intermediate speeds, which serves as a measure of the maximum tolerable surface temperature. Since wear volume in dry sliding is approximately proportional to total load and the distance of sliding, $P\nu$ also gives a measure of wear depth d in the modified form of Archard's relation (3.10.1), $d = k(P\nu)t$, where t is the operating time and wear factor k = wear coefficient K/hardness H.

TABLE 3.10.16. Representative Limiting Conditions for Nonmetallic Bearing Materials

Material	Maximum Temperature, °C	*Pv* Limit, MN/(m·sec)[a]	Maximum Pressure, *P*, MN/m²[b]	Maximum speed, v, m/sec
Thermoplastics				
Nylon	90	0.90	5	3
Filled	150	0.46	10	—
Acetal	100	0.10	5	3
Filled	—	0.28	—	—
PTFE	250	0.04	3.4	0.3
Filled	250	0.53	17	5
Fabric	—	0.88	400	0.8
Polycarbonate	105	0.03	7	5
Polyurethane	120	—	—	—
Polysulfone	160	—	—	—
Thermosetting				
Phenolics	120	0.18	41	13
Filled	160	0.53	—	—
Polyimides	260	4	—	8
Filled	260	5	—	8
Others				
Carbon–graphite	400	0.53	4.1	13
Wood	70	0.42	14	10
Rubber	65	—	0.3	20

[a] See Table 3.10.18.
[b] To convert MN/m² to psi, multiply by 145.

Typical values of this wear factor *k* are given in Table 3.10.17. Since *k* values involve substantial variability, prototype tests are highly desirable for any planned application. Added fillers can reduce the wear factor for the base polymer by a factor of 10 to 1000 and more (Blanchet and Kennedy, 1992). Common fillers include inorganic powders such as clay, glass fibers, graphite, molybdenum disulfide, and powdered metal, and also silicone fluid as an internally available lubricant.

TABLE 3.10.17 Wear Factors for Plastic Bearings[a]

Material	Wear Factor *k*, m²/N No Filler	Filled[b]
Nylon-6, 6	4.0	0.24
PTFE	400	0.14[c]
Acetal resin	1.3	4.9
Polycarbonate	50	3.6
Polyester	4.2	1.8
Poly(phenylene oxide)	60	4.6
Polysulfone	30	3.2
Polyurethane	6.8	3.6

[a] See Booser (1992).
[b] With 30 wt% glass fiber, unless otherwise noted.
[c] 15% glass fiber.

Porous Metals

Bearings of compressed and sintered bronze, iron, and aluminum alloy powder are produced at the rate of millions per week for shaft sizes ranging from about 1.6 to 150 mm. These sleeve bearings and thrust washers are used in a wide variety of small electric motors, appliances, business machines, machine tools, automotive accessories, and farm and construction equipment (Morgan, 1984; Cusano, 1994). Traditional powder metal bearings consist of 90% copper and 10% tin (Table 3.10.18). The common pore volume of 20 to 30% is usually impregnated with a petroleum oil of SAE 30 viscosity. To promote

formation of an oil film, high porosity with its high oil content is employed for higher speeds, often with an oil wick or grease providing a supplementary lubricant supply. Lower porosity with up to 3.5% added graphite is used for lower speeds and oscillation where oil film formation is difficult.

Porous iron bearings are used for lower cost, often with some copper and graphite added for high load capacity at low speed. Iron with up to 40% of added 90–10 bronze powder provides many of the characteristics of porous bronze bearings while enjoying the lower cost of the iron. Porous aluminum containing 3 to 5% copper, tin, and lead finds limited use for providing cooler operation, better conformability, and lower weight.

Table 3.10.18 gives approximate operating limits for porous metal bearings. Generally, maximum P values for sleeve bearings range up to 50,000 psi·ft/min. $P\nu$ levels for thrust bearings should generally not exceed about 20% of this value.

TABLE 3.10.18 Operating Limits for Porous Metal Bearings

Porous Metal	Nominal Composition, wt%	Pressure limit, P, MN/m²		Speed Limit ν, m/sec	$P\nu$ Limit MN/(m·sec)
		Static	Dynamic		
Bronze	Cu 90, Sn 10	59	28	6.1	1.8[a]
Iron		52	25	2.0	1.3
Iron–copper	Fe 90, Cu 10	140	28	1.1	1.4
Iron–copper–carbon	Fe 96, Cu 3, C 0.7	340	56	0.2	2.6
Bronze-iron	Fe 60, Cu 36, Sn 4	72	17	4.1	1.2
Aluminum		28	14	6.1	1.8

Note: To convert MN/m² to psi, multiply by 145.

[a] Approximately equivalent to 50,000 psi · ft/min limit often quoted by U.S. suppliers.

Rolling Element Bearings

Types of Rolling Element Bearings

Rolling element bearings may be classified according to the type of rolling element, i.e., ball or roller, and the loading direction. Ball and roller bearings can be designed to carry either radial or thrust loads, or a combination of the two. Standard rolling element bearing configurations are shown in Figure 3.10.8, and the capabilities of the different types are summarized in Figure 3.10.9.

Ball bearings usually consist of a number of hardened and precisely ground balls interposed between two grooved and hardened rings or races. A cage or separator is used to keep the balls equally spaced around the groove. The most common *radial ball bearing* is a deep groove, or Conrad, type which is designed to carry large radial loads, with a small thrust load capability. The radial capacity of a deep groove bearing can be increased by inserting more balls in the bearing, by means of either a face-located filling notch (which decreases the thrust capacity) or a split inner or outer ring (which requires a means to hold the ring halves axially together). The thrust capability of a radial ball bearing can be increased by inducing angular contact between ball and rings. A single-row angular contact bearing can carry thrust load in only one direction, with the thrust load capacity being dependent on the contact angle (angle between the line of action of the force and the plane perpendicular to the shaft axis). Duplex angular contact bearings consist of two angular contact bearings mounted together so they can carry thrust loads in either direction with little axial play, or they can be mounted in tandem to double the axial and radial load-carrying capacity. Self-aligning ball bearings are designed to accommodate more shaft misalignment than is possible with other radial ball bearings.

Thrust ball bearings are used primarily in machinery with a vertically oriented shaft which requires a stiff axial support. Many such bearings have a 90° contact angle and, as a result, can carry essentially no radial load; they also have limited high-speed capability. If thrust loads in both directions are to be carried, a second row of balls must be added.

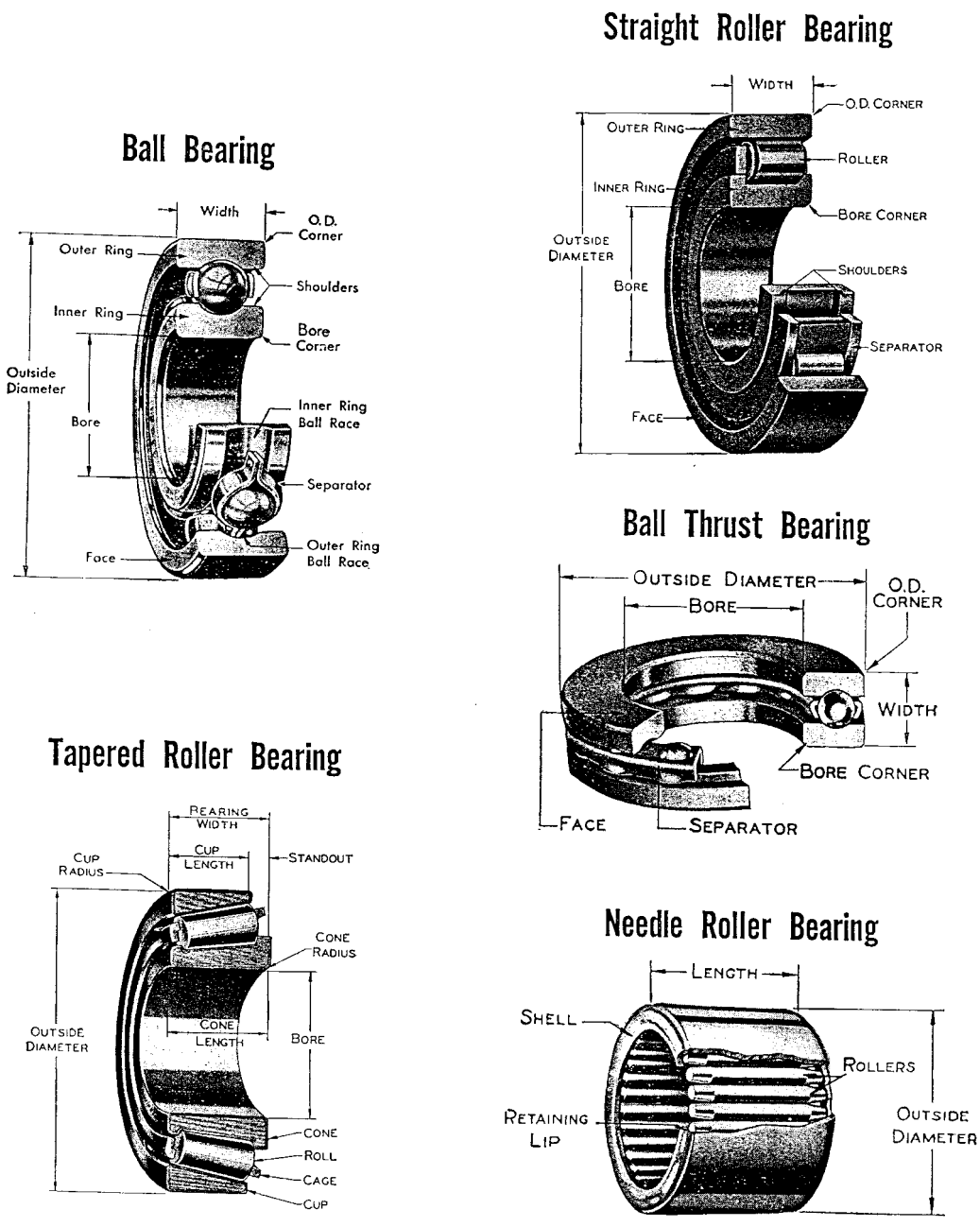


FIGURE 3.10.8 Major types of rolling element bearings.

Roller bearings can be made with cylindrical, tapered, or spherical rollers. As with ball bearings, the rollers are contained between two rings, with a cage or separator used to keep the rollers separated. The cage can be guided by either the rollers or one of the rings. Since roller bearings operate with line contacts, as opposed to the point (or elliptical) contacts that occur in ball bearings, roller bearings are stiffer (less radial displacement per unit load) and have a greater load-carrying capacity than a ball bearing of similar size. Roller bearings are more expensive than ball bearings of comparable size.

CHARACTERISTICS OF STANDARD ROLLING ELEMENT BEARING CONFIGURATIONS

	TYPE	SIZE RANGE IN INCHES		AVERAGE RELATIVE RATINGS				DIMENSIONS	
				Capacity		Limiting Speed	Permissible Misalignment		
		Bore	O.D.	Radial	Thrust			Metric	Inch
BALL BEARINGS	CONRAD TYPE	.1181 to 41.7323	.3750 to 55.1181	Good	Fair ↔	Conrad is basis for comparison 1.00	± 0° 8′	X	X
	MAXIMUM TYPE	.6693 to 4.3307	1.5748 to 8.4646	Excellent	Poor ↔	1.00	± 0° 3′	X	
	ANGULAR CONTACT 15°/40° 15° 40°	.3937 to 7.4803	1.0236 to 15.7480	Good	Good (15°) Excellent (40°) ←	1.00 / 0.70	± 0° 2′	X	
	ANGULAR CONTACT 35°	.3937 to 4.3307	1.1811 to 9.4488	Excellent	Good ←	0.70	0°	X	
	SELF-ALIGNING	.1969 to 4.7244	.7480 to 9.4488	Fair	Fair ↔	1.00	± 4°	X	
CYLINDRICAL ROLLER BEARINGS	SEPARABLE INNER RING NON-LOCATING	.4724 to 19.6850	1.2598 to 28.3465	Excellent	0	1.00	± 0° 4′	X	
	SEPARABLE INNER RING ONE DIR. LOCATING	.4724 to 12.5984	1.2598 to 22.8346	Excellent	Fair ←	1.00	± 0° 4′	X	
	SELF-CONTAINED TWO DIR. LOCATING	.4724 to 3.9370	1.4567 to 8.4646	Excellent	Fair ↔	1.00	± 0° 4′	X	
TAPERED ROLLER BEARINGS	SEPARABLE	.6205 to 6.0000	1.5700 to 10.0000	Good	Good →	0.60	± 0° 2′	X	X
SPHERICAL ROLLER BEARINGS	SELF-ALIGNING	.9843 to 12.5984	2.0472 to 22.8346	Good	Fair ↔	0.50	± 4°	X	
	SELF-ALIGNING	.9843 to 35.4331	2.0472 to 46.4567	Excellent	Good ↔	0.75	± 1°	X	
NEEDLE BEARINGS	COMPLETE BEARINGS with or without locating rings & lubricating groove	.2362 to 14.1732	.6299 to 17.3228	Good	0	0.60	= 0° 2′	X	X
	DRAWN CUP	.1575 to 2.3622	.3150 to 2.6772	Good	0	0.30	± 0° 2′	X	X
THRUST BEARINGS	SINGLE DIRECTION BALL Grooved Race	.2540 to 46.4567	.8130 to 57.0866	Poor	Excellent →	0.30	0°	X	X
	SINGLE DIRECTION CYL. ROLLER	1.1811 to 23.6220	1.8504 to 31.4960	0	Excellent →	0.20	0°	X	
	SELF-ALIGNING SPHERICAL ROLLER	3.3622 to 14.1732	4.3307 to 22.0472	Poor	Excellent →	0.50	± 3°	X	

FIGURE 3.10.9 Characteristics of standard rolling element bearing configurations.

Radial cylindrical roller bearings are designed to carry primarily radial loads. Cylindrical roller bearings have a high radial load capacity and low friction, so they are suitable for high-speed operation. Their thrust load capacity is limited to that which can be carried by contact (sliding) between the ends of the rollers and the flange or shoulder on the ring which contains them. The rollers in many cylindrical

roller bearings are actually slightly crowned to relieve stress concentrations which would otherwise occur at the ends of the rollers and to compensate for misalignment of the bearing. In order to increase the load-carrying capacity of roller bearings, a second row of rollers is added instead of using longer rollers. This is because long rollers (i.e., length/diameter > 1.7) tend to skew in the roller path, thus limiting their high-speed capability and sometimes shortening their life. Needle bearings have long rollers, however, and they are useful when there are severe radial space limitations and when neither high load capacity nor high speeds are required.

Spherical roller bearings usually have an outer ring with a spherical inside diameter, within which are barrel-shaped rollers. This makes these bearings self-aligning, and also gives them a larger contact area between roller and ring than is the case for other rolling element bearings. Because of this, spherical roller bearings have a very high radial load-carrying capacity, along with some ability to carry thrust loads. They have higher friction between roller and ring, and this limits their high-speed capability.

Tapered roller bearings have tapered rollers, ideally shaped like truncated cones, contained between two mating cones of different angles, the inner cone and the outer cup. The contact angle of the bearing determines its thrust load capability; a steeper angle is chosen for more thrust capacity. If a single row of rollers is used, the bearing is separable and can carry thrust loads in only one direction. If the thrust is double acting, a second bearing can be mounted in a back-to-back configuration or a double row bearing can be selected. In tapered roller bearings there is sliding contact between the ends of the rollers and the guide flange on the inner cone, and this sliding contact requires lubrication to prevent wear and reduce friction.

Thrust roller bearings can be either cylindrical, needle, tapered, or spherical (Figure 3.10.9). In each case there is high load-carrying capacity, but the sliding that occurs between rollers and rings requires lubrication and cooling.

Rolling Element Bearing Materials

Ball and roller bearings require materials with excellent resistance to rolling contact fatigue and wear, as well as good dimensional stability and impact resistance. The rolling elements are subjected to cyclic contact pressures which can range from 70 to 3500 MPa (100 to 500 ksi) or more, and the bearing materials must be hard enough to resist surface fatigue under those conditions. Of the through-hardening steels which meet these requirements, the most popular is AISI 52100, which contains about 1% carbon and 1.5% chromium. In general, balls and rollers made from 52100 are hardened to about Rockwell C60. Standard bearings made from 52100 may suffer from unacceptable dimensional changes resulting from metallurgical transformations at operating temperatures above 140°C (285°F). Special stabilization heat treatments enable operation at higher temperatures, with successful operation at temperatures as high as 200°C (390°F) having been achieved in cases involving low loads. The strength and fatigue resistance of the material diminish if the bearing temperature increases above about 175°C (350°F, however, so above that temperature materials with better hot-hardness, such as M50 tool steel, are required. Carburizing steels such as AISI 8620 have been developed for tapered roller bearings and other heavily loaded types that benefit from the tougher core and case compressive residual stress developed during carburizing. For applications in oxidative or corrosive environments, a hardened martensitic stainless steel such as SAE 440C may be chosen. For the highest operating temperatures, ceramic materials may be used in bearings. The most promising of the ceramics for rolling element bearing applications is silicon nitride. Its major use so far in bearings has been in hybrid bearings with ceramic balls or rollers and metallic rings, but all-ceramic bearings have also been developed.

The temperature limits of these bearing materials are given in Table 3.10.19. For all bearing materials, great care must be taken in the processing and production stages to ensure that no defects or inclusions are present that could serve as an initiation site for fatigue cracks. For most high-performance metallic bearings, this requires a very clean steel production process, such as vacuum arc remelting. Heat treatment of the material is also important to produce the required dimensional stability. The production process for ceramic bearings is even more critical, because a defect in a ceramic bearing element could result in catastrophic fracture of the brittle material.

TABLE 3.10.19 Temperature Limits for Rolling Element Bearing Materials

Material	Maximum Operating Temperature	
	°C	°F
AISI 52100	140–175	285–350
AISI 8620 (carburized)	150	300
440C stainless steel	170	340
M50 tool steel	315	600
Hybrid Si_3N_4–M50	425	800
All-ceramic (Si_3N_4)	650	1200

Bearing cages or retainers have as their primary purpose the separation of the rolling elements. In some cases, they also provide some solid lubrication to the bearing. Low-carbon steel is the most common cage material, but bronze (silicon iron bronze or aluminum bronze) and polymers (particularly nylon 6-6) are used in many applications.

Selection of Rolling Element Bearings

It has been stated that if a rolling element bearing in service is properly lubricated, properly aligned, kept free of abrasive particles, moisture, and corrosive agents, and properly loaded, then all causes of damage will be eliminated except one, contact fatigue (Harris, 1991). The fatigue process results in a spall which may originate on or just beneath the contact surface. Studies of rolling contact fatigue life by Lundberg and Palmgren (1947; 1952) and others showed that most rolling element bearings have fatigue lives which follow a Weibull statistical distribution, wherein the dependence of strength on volume is explained by the dispersion in material strength. Most bearings today are designed according to the Lundberg–Palmgren model, which has been standardized by international (ISO, 1990) and national standards (e.g., ANSI/AFBMA, 1990), although recent work (Ioannides and Harris, 1985) has found that modern bearings have longer lives than those predicted by the standard methods.

The basic rating life of rolling element bearings is the L_{10} life, which is the number of revolutions at constant speed for which there is a 10% probability of failure (or 90% reliability). The basic dynamic load-carrying capacity, or load rating, of a bearing is the constant load C which corresponds to an L_{10} life of one million revolutions. For any other bearing load F, the L_{10} life can be determined by the following relationship:

$$L_{10} = (C/F)^n \tag{3.10.18}$$

where the load-life exponent $n = 3$ for ball bearings, and $n = 10/3$ for roller bearings.

The equivalent bearing load includes contributions from both radial and thrust loads, and can be calculated by the following expression:

$$F = XF_r + YF_a \tag{3.10.19}$$

where X is a radial load factor, Y is a thrust load factor, F_r is the radial load applied to the bearing, and F_a is the applied thrust (or axial) load.

Values for the dynamic load rating C, as well as the load factors X and Y for any bearing configuration can be found in manufacturers' catalogs, or they can be calculated according to formulas given in bearing texts by Harris (1991) or Eschmann et al. (1985). Those references also give life adjustment factors which can be used to adjust the desired L_{10} life to account for special operating conditions, special material selections, special lubrication conditions, or for a reliability different from 90%.

The bearing user will generally select a commercially available bearing by the following procedure:

1. Determine the axial and thrust loads acting at the bearing location.
2. Determine the required bearing life (L_{10}).

3. Select the most appropriate bearing type from among those given in Figure 3.10.9.
4. Use the X and Y values appropriate to the type of bearing and loading conditions in Equation (3.10.19) to find the equivalent dynamic bearing load F.
5. Determine the required dynamic load capacity C from Equation (3.10.18).
6. Select a bearing with a dynamic load capacity at least as large as the required value from a manufacturer's catalog.
7. Provide an appropriate mounting arrangement for the bearing. Manufacturers' catalogs can be consulted for guidance in designing the mounting and selecting appropriate fits for the bearing. The importance of the fit cannot be overemphasized, since improper fit can result in considerable reduction in bearing life.
8. Provide adequate lubrication for the bearing (see below). Seals and/or shields may be integrated into the bearing to retain or protect the lubricant in the bearing.

Rolling Bearing Lubrication

The primary load-carrying contacts between rolling elements and rings exhibit nearly pure rolling. There are many sliding contacts in rolling element bearings, however, including those where roller ends contact the internal flanges of rings, where rolling elements contact separator/cage, and where the separator contacts the guiding (piloting) ring of the bearing. All of those contacts must be lubricated to limit friction and wear, and either a grease or an oil can be used for that purpose.

Under most normal operating conditions, rolling element bearings can be grease lubricated. *Greases* coat surfaces with a thin boundary lubricant film of oil, thickener, and additive molecules, thus providing protection against sliding wear, and provide oil to lubricate the concentrated rolling contacts (see below). The selection of a grease depends on its effective temperature range, oil viscosity, consistency, and rust-inhibiting properties. For normal applications, a bearing should be filled with grease up to 30 to 50% of its free volume. Overfilling will cause overheating, particularly at high speeds. Grease will deteriorate with time and will leak out. For that reason, there should be a relubrication schedule, with grease being added at intervals which can be estimated by the following expression (Neale, 1993):

$$\text{relubrication interval (hours)} = \left(k/d^{1/2}\right)\left[\left(14 \times 10^6/n\right) - 4d^{1/2}\right] \tag{3.10.20}$$

where k = 10 for radial ball bearings, 5 for cylindrical roller bearings, and 1 for spherical or tapered roller bearings; d = bearing bore diameter (mm); and n = speed (rpm)

Oil lubrication is required when high speed or high operating temperatures preclude the use of grease. It is necessary to choose an oil of proper viscosity and appropriate viscosity–temperature characteristics in order to insure sufficient thickness of oil film in the lubricated concentrated contacts. If viscosity is too low, the film thickness will not prevent metal/metal contact, but if the viscosity is too high, excessive friction will occur. Oil can be applied to a bearing by one of several methods (listed in order of increasing effectiveness at increasing bearing speed): *oil bath*, in which the rolling elements carry the oil through the bearing; *oil circulating system*, in which the oil is pumped from the bearing through an external filter and heat exchanger and back to the bearing; *oil mist*, in which an airstream carries oil droplets to the bearing; and *oil jets*, in which the oil is injected into the bearing through carefully positioned nozzles. The quantity and entry velocity of the oil must be carefully selected and controlled in order to dissipate the heat generated in the bearing.

The lubrication mechanism in the concentrated contacts of rolling element bearings is *elastohydrodynamic lubrication* (EHL). EHL typically occurs in lubricated, nonconforming elastic contacts, such as the elliptical contact that occurs between ball and raceway or the rectangular contact between roller and ring. These lubricated contacts have a very small area and the contact pressures are very high. Because of those high pressures, the contacting surfaces deform and the lubricant viscosity increases, thereby aiding its ability to sustain heavy loading without oil-film breakdown. A diagram of these phenomena is shown in Figure 3.10.10. The most important parameter of the concentrated contact, from

the point of view of rolling element bearing performance, is minimum EHL film thickness, h_o. The following expression can be used to find minimum film thickness in most rolling element bearing contacts (Hamrock and Dowson, 1977):

$$h_o = 3.63R_x U^{0.68} G^{-0.49} W_p^{-0.73}\left(1 - e^{-0.68\kappa}\right) \tag{3.10.21}$$

where $R_x = (R_{x1}\, R_{x2})/(R_{x1} + R_{x2})$, $R_y = (R_{y1}\, R_{y2})/(R_{y1} + R_{y2})$, ellipticity parameter $\kappa = R_x/R_y$, $U = \mu\,(u_1 + u_2)/2E'R_x$, μ = absolute viscosity, u_1 and u_2 are velocities of rolling element and ring, $E' = E/(1-\nu^2)$, E = modulus of elasticity, ν = Poisson's ratio, $G = \alpha E'$, α = pressure–viscosity exponent, $W_p = W/E'\,R_x^2$, and W = radial load.

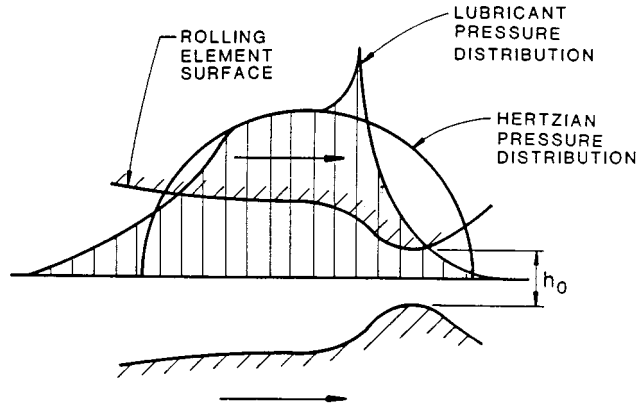


FIGURE 3.10.10 Typical pressure and film thickness distributions in elastohydrodynamic contact.

The minimum film thickness must be large enough to prevent metal/metal contact within the lubricated conjunctions. The criterion for this is stated as

$$h_o \geq 1.5\left(r_{q1}^2 + r_{q2}^2\right)^{0.5} \tag{3.10.22}$$

where r_{q1} and r_{q2} are the rms surface roughness of the rolling element and ring, respectively. If the minimum film thickness is less than this value, complete EHL will not occur, and this could result in wear, surface fatigue, and eventual early bearing failure (i.e., well before the predicted L_{10} life).

An alternative to oil or grease lubrication for rolling element bearings operating under severe conditions is *solid lubrication*. Solid lubricants can be used effectively in high temperatures or vacuum conditions where liquid lubricants are impractical or would provide marginal performance. Solid lubricants do not prevent solid/solid contact, so wear of the lubricant coating can be expected; bearing life is governed by the depletion of the solid lubricant film.

Lubricant Supply Methods

Lubrication systems for oil film bearings can generally be grouped into three classifications: self contained devices for small machines; centralized systems, common in manufacturing plants; and circulating systems dedicated to a single piece of equipment such as a motor, turbine, or compressor. Upper speed limits for common journal bearing lubrication methods are indicated in Figure 3.10.11 (Wilcock and Booser, 1987). Submerging the bearing directly in an oil bath is a common alternative for vertical machines.

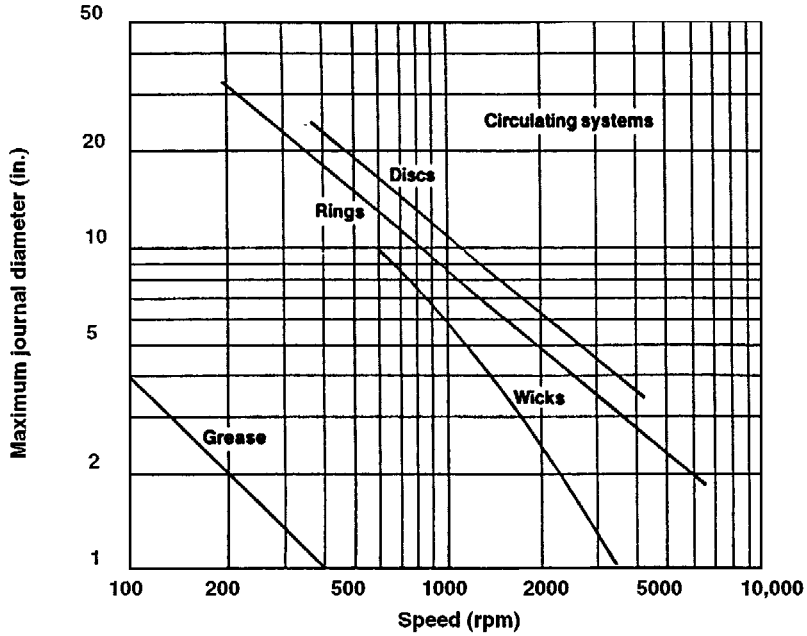


FIGURE 3.10.11 Upper limits for journal bearing lubrication methods. (From Wilcock, D.F. and Booser, E.R., *Mach. Des.*; April 23, 101–107, 1987. With permission.)

Self-Contained Units

Lifting of oil by *capillary action* from a small reservoir is used to feed small bearings in business machines, household appliances, electrical instruments, controls, and timer motors. In capillary tubes, the height h to which oil will rise is (Wilcock and Booser, 1987)

$$h = 2\sigma\cos\theta/(r\rho) \tag{3.10.23}$$

where σ = surface tension, lb/in., r = capillary radius (or spacing between two parallel plates), in.; ρ = oil density, lb/in.^{3}. Because oils wet most surfaces readily, the cosine of the contact angle can be taken as unity. As an example, with $\sigma = 1.7 \times 10^{-4}$ lb/in. for a petroleum oil, the rise in an 0.005-in.-radius capillary will be $h = 2(1.7 \times 10^{-4})(1)/(0.005)(0.0307) = 2.2$ in.

Wick lubrication is applied in millions of fractional horsepower motors annually. Although wicks generally are not efficient at raising oil more than about 2 in., lift may be up to 5 in. in railway journal bearings. By referring to Figure 3.10.12, petroleum oil delivery by a typical wick can be estimated by the following equation (Elwell, 1994):

$$Q = kAF_o(h_u - h)/(\mu L) \quad \text{in.}^3/\text{min} \tag{3.10.24}$$

where the constant k reflects both the capillary spacing in the wick and the surface tension of the oil; A is the wick cross-section area, in.^{2}; F_o is volume fraction of oil in the saturated wick (often about 0.75); h_u is the ultimate wicking height, about 7.5 in. for SAE Grade F-1 felt; h is oil delivery height above the reservoir surface, in.; L is wick length, in.; and μ is viscosity at the wick temperature, lb·sec/in.$^2 \times 10^6$. k is approximately 0.26 for SAE Grade F-1 felt.

Oil rings hanging over a journal, as illustrated in Figure 3.10.13 and Table 3.10.20, are used to lift oil to journal bearings in electric motors, pumps, and medium-size industrial machines (Elwell, 1994).

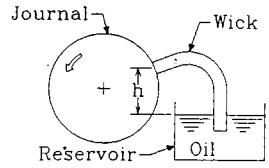


FIGURE 3.10.12 Wick-lubricated journal bearing. (From Elwell, R.C., in *Handbook of Lubrication and Tribology*, Vol. 3, CRC Press, Boca Raton, FL, 1994, 515–533. With permission.)

At very low journal surface speeds below about 2 to 3 ft/sec, the ring will run synchronously with its journal. At higher speeds, increasing viscous drag on the ring in its reservoir will slow the ring surface velocity; oil ring rpm at higher speeds is often in the range of $^1/_{10}$ the journal rpm. Above about 45 ft/sec journal surface velocity, oil delivery drops to an unusably low level as centrifugal throw-off and windage interfere.

These self-contained systems usually supply much less oil to a bearing than needed to form a full hydrodynamic oil film (Elwell, 1994). With the starved oil supply generating an oil wedge of reduced circumferential extent, power loss will be lowered at the expense of reduced load capacity (smaller minimum film thickness).

Centralized Distribution Systems

Limitations with individual localized lubricating devices have led to widespread use of centralized systems for factory production-line equipment, construction and mining machines, and small applications. Oil or soft grease is pumped from a central reservoir in pulses or as metered continuous feed. Oil mist is piped for distances up to 300 ft for machines in petrochemical plants and steel mills. Polymer additives in the 50,000 to 150,000 molecular-weight range greatly reduce the escape of stray oil mist into the atmosphere.

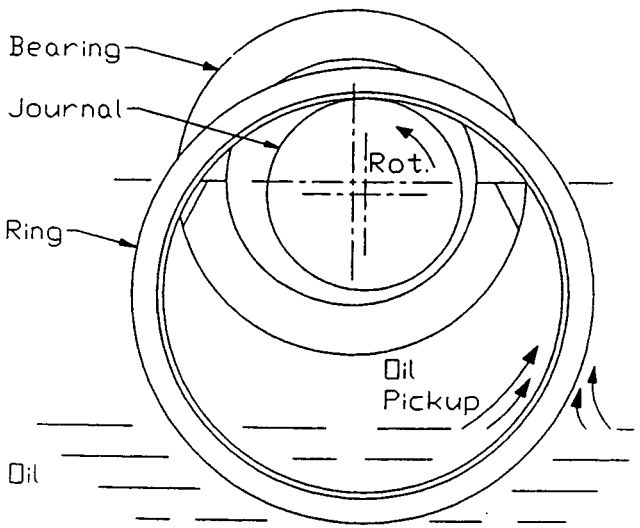
Circulating Systems

Where bearing design, reliability requirements, or equipment considerations preclude use of a simpler oil feed, a circulating system is employed involving an oil reservoir, pump, cooler, and filter (Twidale and Williams, 1984). These systems become quite compact with the space limitations in aircraft, marine, or automobile engines where the reservoir may simply be the machine sump with capacity to hold only a 20- to 60-sec oil supply. Characteristics of typical oil-circulating systems for industrial use are given in Table 3.10.21 (Wilcock and Booser, 1987).

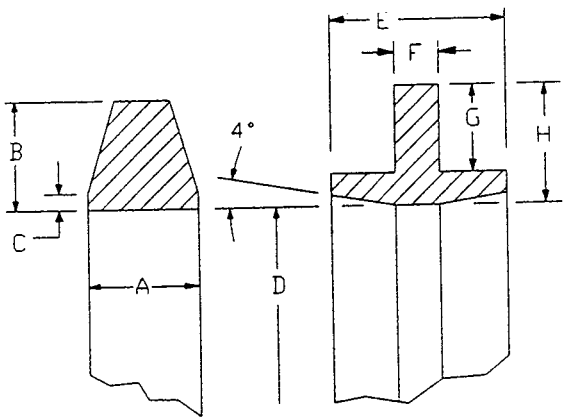
Dynamic Seals

Fluid seals commonly accompany bearings in a wide variety of machinery, both to control leakage of lubricating oil and process fluids and to minimize contamination. Static seals, such as O-rings and gaskets, provide sealing between surfaces which do not move relative to each other. Dynamic seals, which will be the focus of this discussion, restrain flow of fluid between surfaces in relative motion. Most dynamic seals could be classified as either contact or clearance seals. Contact seals are used when the surfaces are in sliding contact, while clearance seals imply that the surfaces are in close proximity to each other but do not contact. The major types of dynamic seals are listed in Table 3.10.22. Further details about the design and analysis of dynamic sealing elements can be found in the handbook article by Stair (1984) and in the book by Lebeck (1991).

As an example of a contact seal, ball bearings are often sealed to retain their lubricant and to keep out contaminants over a long lifetime. For small grease-lubricated bearings, the sealing function is often accomplished by lightly loaded contact between a rubber lip seal component and the bearing ring. For shafts ranging up to about 2 in. (5 cm) in diameter, the lip seal may be replaced by a closely fitted, but noncontacting, shield which helps contain the grease and restricts intrusion of dirt and other contaminants.



Oil-ring bearing elements.



Ring cross-sections.

FIGURE 3.10.13 Oil-ring bearing elements and ring cross-sections. (From Elwell, R.C., in *Handbook of Lubrication and Tribology*, Vol. 3, CRC Press, Boca Raton, FL, 1994, 515–533. With permission.)

TABLE 3.10.20 Typical Oil-Ring Dimensions, mm (in.)

A	B	C	D	E	F	G	H
6 (0.24)	6 (0.24)	1 (0.04)	100 (3.94)				
7 (0.28)	7 (0.28)	1 (0.04)	135 (5.31)				
8 (0.31)	8 (0.31)	2 (0.08)	165 (6.50)				
16 (0.63)	13 (0.51)	2 (0.08)	200 (7.87)	16 (0.63)	5 (0.20)	11 (0.43)	10 (0.39)
			335 (13.2)	21 (0.83)	6 (0.24)	13 (0.51)	14 (0.55)
			585 (23.0)	25 (1.0)	7 (0.28)	14 (0.55)	20 (0.79)
			830 (32.7)	32 (1.3)	8 (0.31)	16 (0.63)	27 (1.06)

TABLE 3.10.21 Typical Oil Circulating Systems

Application	Duty	Oil Viscosity at 40°C (cSt)	Oil Feed (gpm)	Pump Type	Reservoir Dwell Time (min)	Type	Rating (μm)
Electrical machinery	Bearings	32–68	2	Gear	5	Strainer	50
General	Bearings	68	10	Gear	8	Dual cartridge	100
Paper mill dryer section	Bearings, gears	150–220	20	Gear	40	Dual cartridge	120
Steel mill	Bearings	150–460	170	Gear	30	Dual cartridge	150
	Gears	68–680	240	Gear	20	Dual cartridge	
Heavy duty gas turbines	Bearings, controls	32	600	Gear	5	Pleated paper	5
Steam turbine-generators	Bearings	32	1000	Centrifugal	5	Bypass 15%/hr	10

TABLE 3.10.22 Characteristics of Dynamic Seals

Type of Seal	Type of Motion		Extent of Use	Friction	Leakage	Life
	Rotating	Reciprocating				
Contact						
Face seals	x		H	L	L	M–H
Lip seals	x		H	L	L	L–M
Piston rings		x	H	H	L	L–M
O-Rings	x	x	M	H	L	L
Packings	x	x	H	M	M	L
Diaphragms		x	L	L	L	H
Controlled clearance						
Hydrodynamic	x		L	L	M	H
Hydrostatic	x		L	L	M	H
Floating bushing	x	x	M	M	M–H	H
Fixed geometry clearance						
Labyrinth	x		H	H	H	H
Bushing	x	x	M	H	H	M–H
Special						
Magnetic fluid	x	x	L	L	L	M
Centrifugal	x		L	M	L	H

H = High, M = Moderate, L = Low.
Modified from Stair, W.K., in *Handbook of Lubrication*, Vol. 2, CRC Press, Boca Raton, FL 1984, 581–622.

For more severe sealing requirements in rolling element bearings, labyrinth clearance seals may be used (Harris, 1991).

The most common seals for rotating shafts passing through fixed housings, such as pumps or gearboxes, are radial lip seals and mechanical face seals. These contact seals can be designed to handle a wide range of sealed fluids, temperatures, velocities, and pressures (Stair, 1984; Lebeck, 1991). Material selection and surface finish considerations are similar to those discussed in the subsections on sliding friction and its consequences and on dry and semilubricated bearings.

When high surface speeds are encountered, wear and frictional heating may prohibit use of rubbing contact seals. For such applications, clearance seals such as close-fitting labyrinth seals or controlled-clearance fluid film seals can be used. Fluid film seals can be either hydrostatic or hydrodynamic; both types have a pressurized film of fluid which prevents contact between the sealed surfaces and use pressure balancing to restrain leakage. The principles governing their operation are similar to those discussed in the subsection on fluid film bearings. Details of fluid film seals can be found in Shapiro (1995).

References

ANSI/AFBMA, 1990. Load Ratings and Fatigue Life for Ball Bearings, ANSI/AFBMA 9–1990, AFBMA, Washington, D.C.

Archard, J.F. 1980. Wear theory and mechanisms, in *Wear Control Handbook,* M.B. Peterson and W.O. Winer, Eds., ASME, New York.

Bhushan, B. and Gupta, B.K. 1991. *Handbook of Tribology*, McGraw-Hill, New York.

Blanchet, T.A. and Kennedy, F.E. 1992. Sliding wear mechanism of polytetrafluoroethylene (PTFE) and PTFE composites, *Wear*, 153:229–243.

Blau, P.J., Ed. 1992. *Friction, Lubrication and Wear Technology, Metals Handbook*, Vol. 18, 10th ed., ASM International, Metals Park, OH.

Booser, E.R. 1992. Bearing materials, in *Encyclopedia of Chemical Technology*, Vol. 4, pp. 1–21, John Wiley & Sons, New York.

Booser, E.R. 1995. Lubricants and lubrication, in *Encyclopedia of Chemical Technology*, 4th ed., Vol. 15, pp. 463–517, John Wiley & Sons, New York.

Booser, E.R. and Wilcock, D.F. 1987. New technique simplifies journal bearing design, *Mach. Des.*, April 23, pp. 101–107.

Booser, E.R. and Wilcock, D.F. 1991. Selecting thrust bearings, *Mach. Des.*, June 20, pp. 69–72.

Crook, P. and Farmer, H.N. 1992. Friction and wear of hardfacing alloys, in *Friction, Lubrication and Wear Technology, Metals Handbook*, Vol. 18, pp. 758–765, ASM International, Metals Park, OH.

Cusano, C. 1994. Porous metal bearings, in *Handbook of Lubrication and Tribology*, Vol. 3, pp. 491–513, CRC Press, Boca Raton, FL.

DeHart, A.O. 1984. Sliding bearing materials, in *Handbook of Lubrication*, Vol. 2, pp. 463–476, CRC Press, Boca Raton, FL.

Derner, W.J. and Pfaffenberger, E.E. 1984. Rolling element bearings, in *Handbook of Lubrication*, Vol. 2, pp. 495, CRC Press, Boca Raton, FL.

Elwell, R.C. 1994. Self-contained bearing lubrication: rings, disks, and wicks, in *Handbook of Lubrication and Tribology*, Vol. 3, pp. 515–533, CRC Press, Boca Raton, FL.

Engineering Sciences Data Unit (ESDU). 1965. *General Guide to the Choice of Journal Bearing Type*, Item 65007, Institution of Mechanical Engineers, London.

Engineering Sciences Data Unit (ESDU). 1967. *General Guide to the Choice of Thrust Bearing Type*, Item 67073, Institution of Mechanical Engineers, London.

Eschmann, P., Hasbargen, L., and Weigand, K. 1985. *Ball and Roller Bearings*, John Wiley & Sons, New York.

Fenske, G.R. 1992. Ion implantation, in *Friction, Lubrication and Wear Technology, Metals Handbook,* Vol. 18, pp. 850–860, ASM International, Metals Park, OH.

Fuller, D.D. 1984. Theory and practice of lubrication for engineers, 2nd ed., John Wiley & Sons, New York.

Hamrock, B. and Dowson, D. 1977. Isothermal elastohydrodynamic lubrication of point contacts, *ASME J. Lubr. Technol.*, 99(2): 264–276.

Harris, T.A. 1991. *Rolling Bearing Analysis*, 3rd ed., John Wiley & Sons, New York.

Ioannides, S. and Harris, T.A. 1985. A new fatigue life model for rolling bearings, *ASME J. Tribology*, 107:367–378.

ISO, 1990. Rolling Bearings Dynamic Load Ratings and Rating Life, International Standard ISO 281.

Jamison, W.E. 1994. Plastics and plastic matrix composites, in *Handbook of Lubrication and Tribology*, Vol. 3, pp. 121–147, CRC Press, Boca Raton, FL.

Khonsari, M.M. 1997. In *Tribology Data Handbook*, CRC Press, Boca Raton, FL.

Kingsbury, G.R. 1992. Friction and wear of sliding bearing materials, in *ASM Handbook*, Vol. 18 pp. 741–757, ASM International, Metals Park, OH.

Klaus, E.E. and Tewksbury, E.J. 1984. Liquid lubricants, in *Handbook of Lubrication*, Vol. 2, pp. 229–254, CRC Press, Boca Raton, FL.

Kushner, B.A. and Novinski, E.R. 1992. Thermal spray coatings, in *Friction, Lubrication and Wear Technology, Metals Handbook*, Vol. 18, pp. 829–833, ASM International, Metals Park, OH.

Lebeck, A.O. 1991. *Principles and Design of Mechanical Face Seals,* John Wiley & Sons, New York.

Lundberg, G. and Palmgren, A. 1947. Dynamic capacity of rolling bearings, *Acta Polytech. Mech. Eng. Ser.*, 1(3):196.

Lundberg, G. and Palmgren, A. 1952. Dynamic capacity of roller bearings, *Acta Polytech. Mech. Eng. Ser.*, 2(4):210.

Morgan, V.T. 1984. *Porous Metal Bearings and Their Application*, MEP-213, Mechanical Engineering Publications, Workington, U.K.

Neale, M.J. 1993. *Bearings*, Butterworth-Heinemann, Oxford.

Neale, P.B. 1970. *J. Mech. Eng. Sci.*, 12:73–84.

Peterson, M.B. and Winer, W.O., Eds., 1980. *Wear Control Handbook*, ASME, New York.

Rabinowicz, E. 1980. Wear coefficients metals, in *Wear Control Handbook*, M.B. Peterson and W.O. Winer, Eds., pp. 475–506, ASME, New York.

Rabinowicz, E. 1995. *Friction and Wear of Materials*, 2nd ed., John Wiley & Sons, New York.

Ramondi, A.A. and Szeri, A.Z. 1984. Journal and thrust bearings, in *Handbook of Lubrication,* Vol. 2, pp. 413–462, CRC Press, Boca Raton, FL.

Reynolds, O. 1886. On the theory of lubrication and its application to Mr. Beauchamp Tower's experiments, *Philos. Trans R. Soc.*, 177:157–234.

Schmitt, G.F. 1980. Liquid and solid particles impact erosion, in *Wear Control Handbook,* M.B. Peterson and W.O. Winer, Eds., pp. 231–282, ASME, New York.

Shabel, B.S. Granger, D.A., and Tuckner, W.G. 1992. Friction and wear of aluminum–silicon alloys, in *ASM Handbook*, Vol. 18, pp. 785–794, ASM International, Metals Park, OH.

Shapiro, W. 1995. Hydrodynamic and hydrostatic seals, in *Handbook of Lubrication and Tribology*, Vol. 3, pp. 445–468, CRC Press, Boca Raton, FL.

Stair, W.K. 1984. Dynamic seals, in *Handbook of Lubrication*, Vol. 2, pp. 581–622, CRC Press, Boca Raton, FL.

Twidale, A.J. and Williams, D.C.J. 1984. Circulating oil systems, in *Handbook of Lubrication*, Vol. 2, pp. 395–409, CRC Press, Boca Raton, FL.

Weil, R. and Sheppard, K. 1992. Electroplated coatings, in *Friction, Lubrication and Wear Technology, Metals Handbook*, Vol. 18, pp. 834–839, ASM International, Ohio.

Wilcock, D.F. and Booser, E.R. 1956. Bearing design and application, McGraw-Hill, New York.

Wilcock, D.F. and Booser, E.R. 1987. Lubrication techniques for journal bearings, *Machine Des.*, April 23, 101–107.

3.11 Pumps and Fans

Robert F. Boehm

Introduction

Pumps are devices that impart a pressure increase to a liquid. Fans are used to increase the velocity of a gas, but this is also accomplished through an increase in pressure. The pressure rise found in pumps can vary tremendously, and this is a very important design parameter along with the liquid flow rate. This pressure rise can range from simply increasing the elevation of the liquid to increasing the pressure hundreds of atmospheres. Fan applications, on the other hand, generally deal with small pressure increases. In spite of this seemingly significant distinction between pumps and fans, there are many similarities in the fundamentals of certain types of these machines as well as with their application and theory of operation.

The appropriate use of pumps and fans depends upon the proper choice of device and the proper design and installation for the application. A check of sources of commercial equipment shows that many varieties of pumps and fans exist. Each of these had special characteristics that must be appreciated for achieving proper function. Preliminary design criteria for choosing between different types is given by Boehm (1987).

As is to be expected, the wise applications of pumps and fans requires knowledge of fluid flow fundamentals. Unless the fluid mechanics of a particular application are understood, the design could be less than desirable.

In this section, pump and fan types are briefly defined. In addition, typical application information is given. Also, some ideas from fluid mechanics that are especially relevant to pump and fan operation are reviewed.

Pumps

Raising of water from wells and cisterns is the earliest form of pumping (a very detailed history of early applications is given by Ewbank, 1842). Modern applications are much broader, and these find a wide variety of machines in use. Modern pumps function on one of two principles. By far the majority of pump installations are of the *velocity head* type. In these devices, the pressure rise is achieved by giving the fluid a movement. At the exit of the machine, this movement is translated into a pressure increase. The other major type of pump is called *positive displacement*. These devices are designed to increase the pressure of the liquid while essentially trying to compress the volume. A categorization of pump types has been given by Krutzsch (1986), and an adaptation of this is shown below.

I. Velocity head
   A. Centrifugal
      1. Axial flow (single or multistage)
      2. Radial flow (single or double suction)
      3. Mixed flow (single or double suction)
      4. Peripheral (single or multistage)
   B. Special Effect
      1. Gas lift
      2. Jet
      3. Hydraulic ram
      4. Electromagnetic
II. Positive displacement
   A. Reciprocating
      1. Piston, plunger
         a. Direct acting (simplex or duplex)

b. Power (single or double acting, simplex, duplex, triplex, multiplex)
2. Diaphragm (mechanically or fluid driven, simplex or multiplex)

B. Rotary
1. Single rotor (vane, piston, screw, flexible member, peristaltic)
2. Multiple rotor (gear, lobe, screw, circumferential piston)

In the next subsection, some of the more common pumps are described.

Centrifugal and Other Velocity Head Pumps

Centrifugal pumps are used in more industrial applications than any other kind of pump. This is primarily because these pumps offer low initial and upkeep costs. Traditionally, pumps of this type have been limited to low-pressure-head applications, but modern pump designs have overcome this problem unless very high pressures are required. Some of the other good characteristics of these types of devices include smooth (nonpulsating) flow and the ability to tolerate nonflow conditions.

The most important parts of the centrifugal pump are the *impeller* and *volute*. An impeller can take on many forms, ranging from essentially a spinning disk to designs with elaborate vanes. The latter is usual. Impeller design tends to be somewhat unique to each manufacturer, as well as finding a variety of designs for a variety of applications. An example of an impeller is shown in Figure 3.11.1. This device imparts a radial velocity to the fluid that has entered the pump perpendicular to the impeller. The volute (there may be one or more) performs the function of slowing the fluid and increasing the pressure. A good discussion of centrifugal pumps is given by Lobanoff and Ross (1992).

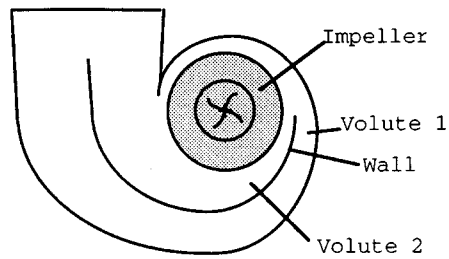


FIGURE 3.11.1. A schematic of a centrifugal pump is shown. The liquid enters perpendicular to the figure, and a radial velocity is imparted by clockwise spin of the impeller.

A very important factor in the specification of a centrifugal pump is the *casing orientation* and *type*. For example, the pump can be oriented vertically or horizontally. Horizontal mounting is most common. Vertical pumps usually offer benefits related to ease of priming and reduction in required net positive suction head (see discussion below). This type also requires less floor space. Submersible and immersible pumps are always of the vertical type. Another factor in the design is the way the casing is split, and this has implications about ease of manufacture and repair. Casings that are split perpendicular to the shaft are called *radially split*, while those split parallel to the shaft axis are denoted as *axially split*. The latter can be *horizontally split* or *vertically split*. The number of *stages* in the pump greatly affects the pump-output characteristics. Several stages can be incorporated into the same casing, with an associated increase in pump output. Multistage pumps are often used for applications with total developed head over 50 atm.

Whether or not a pump is self-priming can be important. If a centrifugal pump is filled with air when it is turned on, the initiation of pumping action may not be sufficient to bring the fluid into the pump. Pumps can be specified with features that can minimize priming problems.

There are other types of velocity head pumps. *Jet pumps* increase pressure by imparting momentum from a high-velocity liquid stream to a low-velocity or stagnant body of liquid. The resulting flow then

goes through a diffuser to achieve an overall pressure increase. *Gas lifts* accomplish a pumping action by a drag on gas bubbles that rise through a liquid.

Positive-Displacement Pumps

Positive-displacement pumps demonstrate high discharge pressures and low flow rates. Usually, this is accomplished by some type of pulsating device. A piston pump is a classic example of positive-displacement machines. Rotary pumps are one type of positive-displacement device that do not impart pulsations to the existing flow (a full description of these types of pumps is given by Turton, 1994). Several techniques are available for dealing with pulsating flows, including use of double-acting pumps (usually of the reciprocating type) and installation of pulsation dampeners.

Positive-displacement pumps usually require special seals to contain the fluid. Costs are higher both initially and for maintenance compared with most pumps that operate on the velocity head basis. Positive-displacement pumps demonstrate an efficiency that is nearly independent of flow rate, in contrast to the velocity head type (see Figure 3.11.2 and the discussion related to it below).

Reciprocating pumps offer very high efficiencies, reaching 90% in larger sizes. These types of pumps are more appropriate for pumping abrasive liquids (e.g., slurries) than are centrifugal pumps.

A characteristic of positive displacement pumps which may be valuable is that the output flow is proportional to pump speed. This allows this type of pump to be used for metering applications. Also a positive aspect of these pumps is that they are self-priming, except at initial start-up.

Very high head pressures (often damaging to the pump) can be developed in positive-displacement pumps if the downstream flow is blocked. For this reason, a pressure-relief-valve bypass must always be used with positive-displacement pumps.

Pump/Flow Considerations

Performance characteristics of the pump must be considered in system design. Simple diagrams of pump applications are shown in Figure 3.11.2. First, consider the left-hand figure. This represents a flow circuit, and the pressure drops related to the piping, fittings, valves, and any other flow devices found in the circuit must be estimated using the laws of fluid mechanics. Usually, these resistances (pressure drops) are found to vary approximately with the square of the liquid flow rate. Typical characteristics are shown in Figure 3.11.3. Most pumps demonstrate a flow vs. pressure rise variation that is a positive value at zero flow and decreases to zero at some larger flow. Positive-displacement pumps, as shown on the right-hand side of Figure 3.11.3, are an exception to this in that these devices usually cannot tolerate a zero flow. An important aspect to note is that a closed system can presumably be pressurized. A contrasting situation and its implications are discussed below.

The piping diagram show on the right-hand side of Figure 3.11.2 is a once-through system, another frequently encountered installation. However, the leg of piping through "pressure drop 1" shown there can have some very important implications related to *net positive suction head*, often denoted as **NPSH**. In simple terms, NPSH indicates the difference between the local pressure and the thermodynamic saturation pressure at the fluid temperature. If NPSH = 0, the liquid can vaporize, and this can result in a variety of outcomes from noisy pump operation to outright failure of components. This condition is called **cavitation**. Cavitation, if it occurs, will first take place at the lowest pressure point within the piping arrangement. Often this point is located at, or inside, the inlet to the pump. Most manufacturers specify how much NPSH is required for satisfactory operation of their pumps. Hence, the actual NPSH (denoted as **NPSHA**) experienced by the pump must be larger than the manufacturer's required NPSH (called **NPSHR**). If a design indicates insufficient NPSH, changes should be made in the system, possibly including alternative piping layout, including elevation and/or size, or use of a pump with smaller NPSH requirements.

The manufacturer should be consulted for a map of operational information for a given pump. A typical form is shown in Figure 3.11.4. This information will allow the designer to select a pump that satisfied the circuit operational requirements while meeting the necessary NPSH and most-efficient-operation criteria.

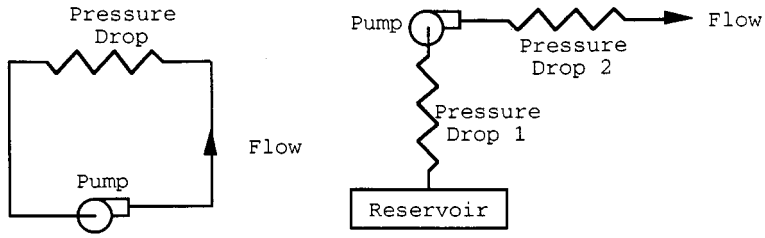


FIGURE 3.11.2. Typical pump applications, either in circuits or once-through arrangements, can be represented as combined fluid resistances as shown. The resistances are determined from fluid mechanics analyses.

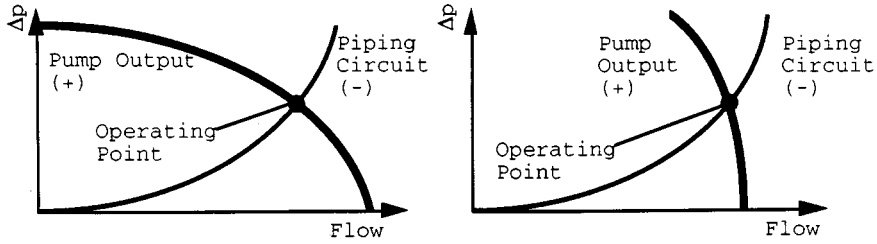


FIGURE 3.11.3. An overlay of the pump flow vs. head curve with the circuit piping characteristics gives the operating state of the circuit. A typical velocity head pump characteristic is shown on the left, while a positive-displacement pump curve is shown on the right.

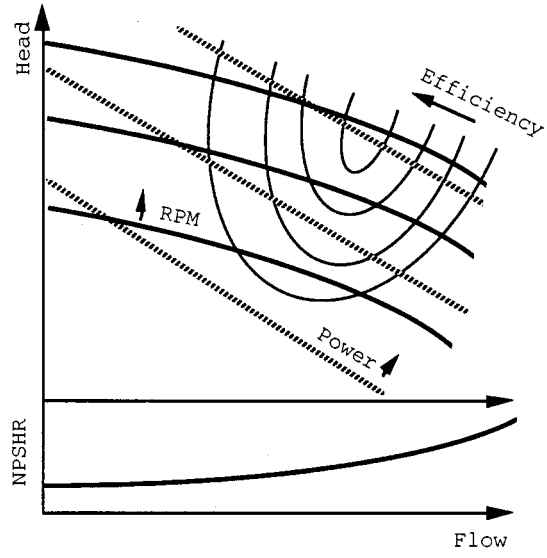


FIGURE 3.11.4. A full range of performance information should be available from the pump manufacturer, and this may include the parameters shown.

Several options are available to the designer for combining pumps in systems. Consider a comparison of the net effect between operating pumps in series or operating the same two pumps in parallel. Examples of this for pumps with characteristics such as centrifugal units are shown in Figure 3.11.5. It is clear that one way to achieve high pumping pressures with centrifugal pumps is to place a number of units in series. This is a related effect to what is found in *multistage* designs.

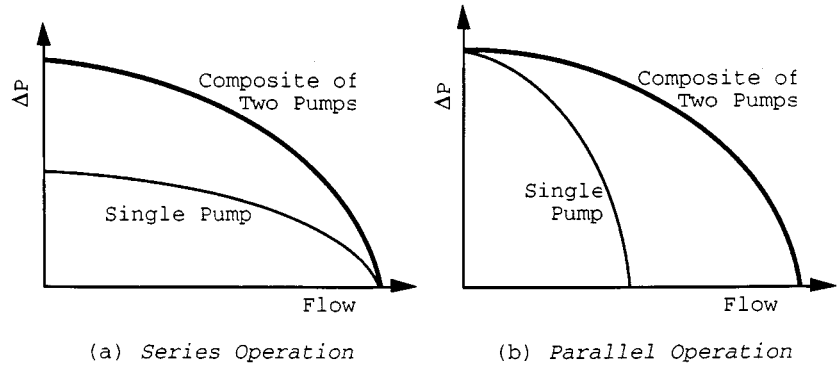


FIGURE 3.11.5. Series (a) and parallel (b) operation of centrifugal pumps are possible. The resultant characteristics for two identical pumps are shown.

Fans

As noted earlier, fans are devices that cause air to move. This definition is broad and can include a flapping palm branch, but the discussion here deals only with devices that impart air movement due to *rotation of an impeller inside a fixed casing*. In spite of this limiting definition, a large variety of commercial designs are included.

Fans find application in many engineering systems. Along with the chillers and boilers, they are the heart of heating, ventilating, and air conditioning (HVAC) systems. When large physical dimensions of a unit are not a design concern (usually the case), centrifugal fans are favored over axial flow units for HVAC applications. Many types of fans are found in *power plants*. Very large fans are used to furnish air to the boiler, as well as to draw or force air through cooling towers and pollution-control equipment. *Electronic cooling* finds applications for small units. Even automobiles have several fans in them. Because of the great engineering importance of fans, several organizations publish rating and testing criteria (see, for example, ASME, 1990).

Generally fans are classified according to how the air flows through the impeller. These flows may be *axial* (essentially a propeller in a duct), *radial* (conceptually much like the centrifugal pumps discussed earlier), *mixed*, and *cross*. While there are many other fan designations, all industrial units fit one of these classifications. Mixed-flow fans are so named because both axial and radial flow occur on the vanes. Casings for these devices are essentially like those for axial-flow machines, but the inlet has a radial-flow component. On cross-flow impellers, the gas traverses the blading twice.

Characteristics of fans are shown in Figure 3.11.6. Since velocities can be high in fans, often both the total and the static pressure increases are considered. While both are not shown on this figure, the curves have similar variations. Of course the total ΔP will be greater than will the static value, the difference being the velocity head. This difference increases as the volume flow increases. At zero flow (the shutoff point), the static and total pressure difference values are the same. Efficiency variation shows a sharp optimum value at the design point. For this reason, it is critical that fan designs be carefully tuned to the required conditions.

A variety of vane type are found on fans, and the type of these is also used for fan classification. Axial fans usually have vanes of airfoil shape or vanes of uniform thickness. Some vane types that might be found on a centrifugal (radial-flow) fan are shown in Figure 3.11.7.

One aspect that is an issue in choosing fans for a particular application is fan efficiency. Typical efficiency comparisons of the effect of blade type on a centrifugal fan are shown in Figure 3.11.8. Since velocities can be high, the value of aerodynamic design is clear. Weighing against this are cost and other factors.

An additional aspect that may be important in the choice of fans is noise generation. This may be most critical in HVAC applications. It is difficult to describe noise characteristics in brief terms because of the frequency-dependent nature of these phenomena. However, a comparison of specific sound power

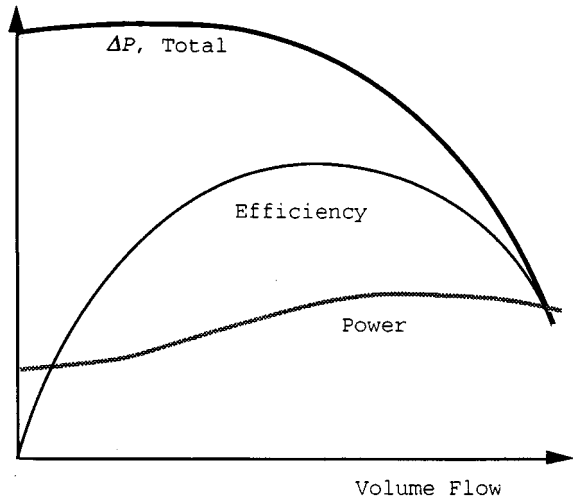


FIGURE 3.11.6. Shown are characteristics of a centrifugal fan. The drawbacks to operating away from optimal conditions are obvious from the efficiency variation.

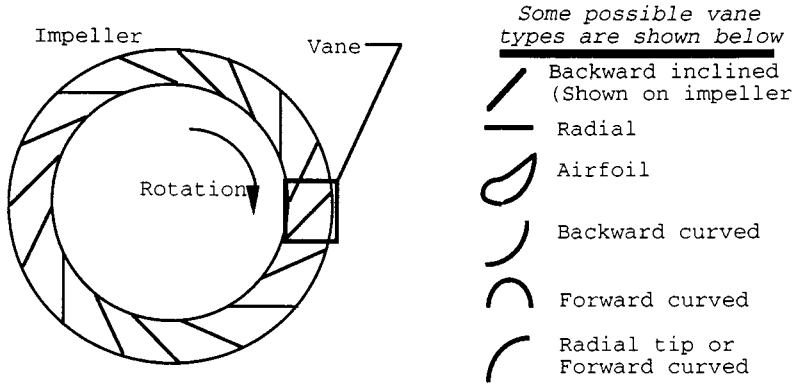


FIGURE 3.11.7. A variety of vane types that might be used on a centrifugal fan are shown.

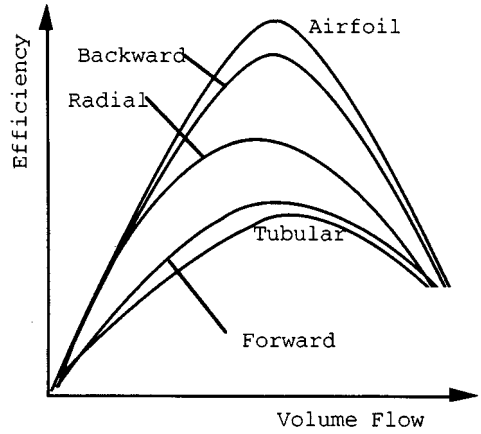


FIGURE 3.11.8. Efficiency variation with volume flow of centrifugal fans for a variety of vane types is shown.

level (usually denoted by K_w) shows backward-curved centrifugal fans with aerodynamic blades perform best among the designs. Details of noise characteristics are given elsewhere (ASHRAE, 1991).

While each type of fan has some specific qualities for certain applications, most installations use centrifugal (radial-flow) fans. A primary exception is for very-high-flow, low-pressure-rise situations where axial (propeller) fans are used.

Similarities exist between fans and pumps because the fluid density essentially does not vary through either type of machine. Of course, in pumps this is because a liquid can be assumed to be incompressible. In fans, a gas (typically air) is moved with little pressure change. As a result, the gas density can be taken to be constant. Since most fans operate near atmospheric pressure, the ideal gas assumptions can be used in determining gas properties.

Flow control in fan applications, where needed, is a very important design concern. Methods for accomplishing this involve use of dampers (either on the inlet or on the outlet of the fan), variable pitch vanes, or variable speed control. Dampers are the least expensive to install, but also the most inefficient in terms of energy use. Modern solid state controls for providing a variable frequency power to the drive motor is becoming the preferred control method, when a combination of initial and operating costs is considered.

Defining Terms

Cavitation: Local liquid conditions allow vapor voids to form (boiling).
NPSH: Net positive suction head is the difference between the local absolute pressure of a liquid and the thermodynamic saturation pressure of the liquid based upon the temperature of the liquid. Applies to the inlet of a pump.
NPSHA: Actual net positive suction head is the NPSH at the given state of operation of a pump.
NPSHR: Required net positive suction head is the amount of NPSH required by a specific pump for a given application.

References

ASHRAE, 1991. *ASHRAE Handbook 1991, HVAC Applications,* American Society of Heating, Refrigerating, and Air Conditioning Engineers, Atlanta, Chapter 42.

ASME, 1990. *ASME Performance Test Codes, Code on Fans*, ASME PTC 11-1984 (reaffirmed 1990), American Society of Mechanical Engineers, New York.

Boehm, R.F. 1987. *Design Analysis of Thermal Systems*, John Wiley and Sons, New York, 17–26.

Ewbank, T. 1842. *A Description and Historical Account of Hydraulic and Other Machines for Raising Water*, 2nd ed., Greeley and McElrath, New York.

Krutzsch, W.C. 1986. Introduction: classification and selection of pumps, in *Pump Handbook*, 2nd ed., I. Karassik et al., Eds., McGraw-Hill, New York, Chapter 1.

Lobanoff, V. and Ross, R. 1992. *Centrifugal Pumps: Design & Application*, 2nd ed., Gulf Publishing Company, Houston.

Turton, R.K. 1994. *Rotodynamic Pump Design*, Cambridge University Press, Cambridge, England.

Further Information

Dickson, C. 1988. *Pumping Manual*, 8th ed., Trade & Technical Press, Morden, England.

Dufour, J. and Nelson, W. 1993. *Centrifugal Pump Sourcebook*, McGraw-Hill, New York.

Fans. 1992. In *1992 ASHRAE Handbook, HVAC Systems and Equipment*, American Society of Heating, Refrigerating, and Air Conditioning Engineers, Atlanta, GA, Chapter 18.

Garay, P.N. 1990. *Pump Application Book*, Fairmont Press, Liburn, GA.

Krivchencko, G.I. 1994. *Hydraulic Machines, Turbines and Pumps*, 2nd ed., Lewis Publishers, Boca Raton, FL.

Stepanoff, A.J. 1993. *Centrifugal and Axial Flow Pumps: Theory, Design, and Application* (Reprint Edition), Krieger Publishing Company, Malabar, FL.

3.12 Liquid Atomization and Spraying

Rolf D. Reitz

Sprays are involved in many practical applications, including in the process industries (e.g., spray drying, spray cooling, powdered metals); in treatment applications (e.g., humidification, gas scrubbing); in coating applications (e.g., surface treatment, spray painting, and crop spraying); in spray combustion (e.g., burners, furnaces, rockets, gas turbines, diesel and port fuel injected engines); and in medicinal and printing applications. To be able to describe sprays it is necessary to obtain a detailed understanding of spray processes.

In the simplest case, the liquid to be sprayed is injected at a high velocity through a small orifice. Atomization is the process whereby the injected liquid is broken up into droplets. Atomization has a strong influence on spray vaporization rates because it increases the total surface area of the injected liquid greatly. Fast vaporization may be desirable in certain applications, but undesirable in others, where the liquid is required to impinge on a target. The trajectories of the spray drops are governed by the injected momentum of the drop, drag forces, and interactions between the drops and the surrounding gas. Control of these and other spray processes can lead to significant improvements in performance and in quality of product, and to reduction of emission of pollutants.

Spray Characterization

Practical atomizers generate sprays with a distribution of drop sizes, with average sizes in the diameter range from a few microns (1 μm = 10^{-6} m) to as large as 0.5 mm. It is important to quantify the details of the distribution depending on the application. For example, the smaller drops in a spray vaporize fast, and this is helpful to control ignition processes in some combustion systems. On the other hand, the large drops carry most of the mass and momentum of the injected liquid and these drops are able to penetrate into the high-pressure gases in engine combustion chambers. Typical average drop sizes for broad classes of sprays are shown schematically in Figure 3.12.1. It should be noted that the terminology used to describe sprays in Figure 3.12.1 is qualitative and is not universally agreed upon.

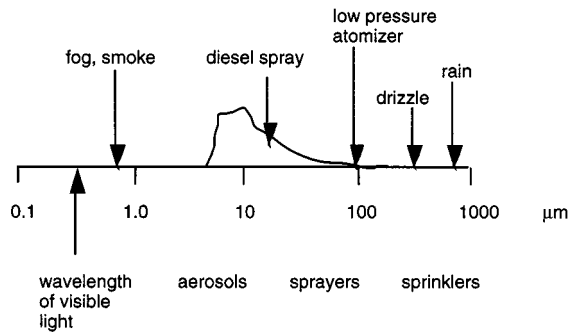


FIGURE 3.12.1 Typical average spray drop sizes for various classes of sprays. A representative size distribution is depicted for the diesel spray.

Methods for characterizing the size distribution of spray drops are discussed in References 1 and 2. A probability distribution function, $F(D)$, is introduced that represents the fraction of drops per unit diameter range about the diameter, D, as shown in Figure 3.12.2. The spray drop sizes span a range from a minimum diameter, D_{min}, to a maximum diameter, D_{max}. It is also convenient to introduce a mean or average drop diameter instead of having to specify the complete drop size distribution. The number median drop diameter (NMD) represents that drop whose diameter is such that 50% of the drops in the spray have sizes less than this size. Spray drop size distribution data can also be represented as a volume

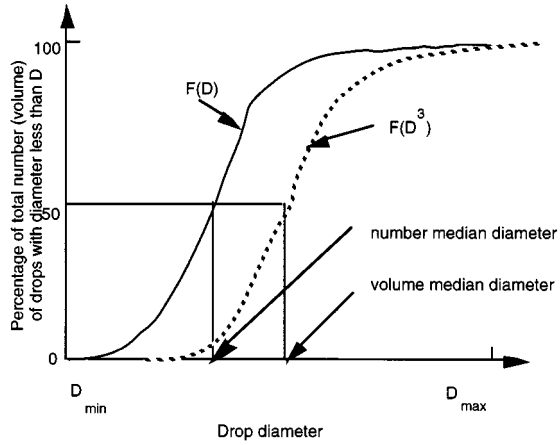


FIGURE 3.12.2 Cumulative spray drop number and volume distributions.

(or mass) distribution function, $F(D^3)$; this gives more weight to the large drops in the distribution. In this case, a volume median diameter (VMD) or a mass median diameter (MMD) can also be defined, as indicated in Figure 3.12.2.

Various other mean diameters are also in common use. These are summarized using the standard notation of Mugele and Evans[2] as

$$\left(D_{jk}\right)^{j-k} = \frac{\int_{D_{\min}}^{D_{\max}} D^j f(D)\, dD}{\int_{D_{\min}}^{D_{\max}} D^k f(D)\, dD} \tag{3.12.1}$$

where $f(D) = dF(D)/dD$ is the drop size probability density function (usually normalized such that $\int_{D_{\min}}^{D_{\max}} f(D)dD = 1$). Commonly used mean diameters are D_{10} (i.e., $j = 1$, $k = 0$, sometimes called the length mean diameter[3] and D_{32} (i.e., $j = 3$, $k = 2$, called the Sauter mean diameter or SMD). The Sauter mean diameter has a useful physical interpretation in combustion applications since drop vaporization rates are proportional to the surface area of the drop. It represents the size of that drop that has the same volume-to-surface area ratio as that of the entire spray.

Several distribution functions have been found to fit experimental data reasonably well. Among these are the Nukiyama–Tanasawa and the Rosin–Rammler distributions which have the general form[3] $f(D) = aD^p \exp(\{-bD\}^q)$, where the constants a, p, b, and q characterize the size distribution. The higher the parameter, q, the more uniform the distribution, and typically $1.5 < q < 4$. Other distributions have been proposed which consist of logarithmic transformations of the normal distribution, such as $f(D) = a \exp(-y^2/2)$, where $y = \delta \ln(\eta D/(D_{\max} - D))$, and a, δ, and η are constants. In this case, the smaller δ, the more uniform the size distribution. It should be noted that there is no theoretical justification for any of these size distributions. Spray drop size distributions can be measured nonintrusively by using optical laser diffraction and phase/Doppler instruments. A discussion of these techniques and their accuracy is reviewed by Chigier.[4]

Atomizer Design Considerations

Atomization is generally achieved by forcing a liquid or a liquid–gas mixture through a small hole or slit under pressure to create thin liquid sheets or jets moving at a high relative velocity with respect to the surrounding ambient gas. Desirable characteristics of atomizers include the ability to atomize the

liquid over a wide range of flow rates, low power requirements, and low susceptibility to blockage or fouling. In addition, atomizers should produce consistent sprays with uniform flow patterns in operation.

Atomizers can be broadly characterized as those producing hollow cone or solid cone sprays, as depicted in Figure 3.12.3. In solid cone (or full cone) sprays the spray liquid is concentrated along the spray axis, Figure 3.12.3(a). These sprays are useful in applications requiring high spray penetration, such as in diesel engines. In hollow cone sprays the axis region is relatively free of drops, giving wide spray dispersal, Figure 3.12.3(b). These sprays are often used in furnaces, gas turbines, and spray-coating applications.

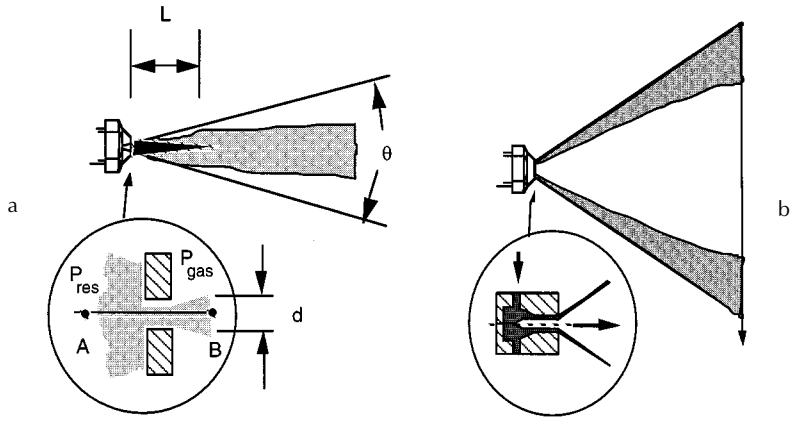


FIGURE 3.12.3 Schematic diagram of (a) solid cone and (b) hollow cone pressure atomizer sprays.

Many different atomizer designs are found in applications. Common atomizer types include pressure, rotary, twin-fluid (air-assist, air-blast, effervescent), flashing, electrostatic, vibratory, and ultrasonic atomizers, as discussed next.

Atomizer Types

In *pressure atomizers* atomization is achieved by means of a pressure difference, $\Delta P = P_{res} - P_{gas}$, between the liquid in the supply reservoir pressure, P_{res}, and the ambient medium pressure, P_{gas}, across a nozzle. The simplest design is the plain orifice nozzle with exit hole diameter, d, depicted in Figure 3.12.3(a). The liquid emerges at the theoretical velocity $U = \sqrt{2\Delta P/\rho_{liquid}}$, the (Bernoulli) velocity along the streamline A–B in Figure 3.12.3(a), where ρ_{liquid} is the density of the liquid. The actual injection velocity is less than the ideal velocity by a factor called the discharge coefficient, C_D, which is between 0.6 and 0.9 for plain hole nozzles. C_D accounts for flow losses in the nozzle.

Four main jet breakup regimes have been identified, corresponding to different combinations of liquid inertia, surface tension, and aerodynamic forces acting on the jet, as shown in Figure 3.12.4. At low injection pressures the low-velocity liquid jet breaks up due to the unstable growth of long-wavelength waves driven by surface tension forces (Rayleigh regime). As the jet velocity is increased, the growth of disturbances on the liquid surface is enhanced because of the interaction between the liquid and the ambient gas (the first and second wind-induced breakup regimes). At high injection pressures the high-velocity jet disintegrates into drops immediately after leaving the nozzle exit (atomization regime). Criteria for the boundaries between the regimes are available.[5] Aerodynamic effects are found to become very important relative to inertial effects when the jet Weber number, $We_j > 40$, where $We_j = \rho_{gas} U^2 d/\sigma$, ρ_{gas} is the gas density, and σ is the liquid surface tension.

Experiments show that the unstable growth of surface waves is aided by high relative velocities between the liquid and the gas, and also by high turbulence and other disturbances in the liquid and gas flows, and by the use of spray liquids with low viscosity and low surface tension.

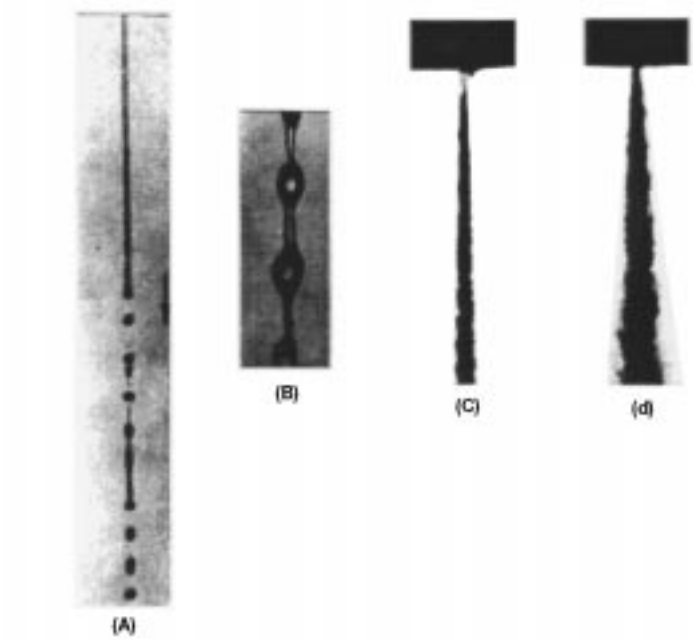


FIGURE 3.12.4 (a) Rayleigh breakup. Drop diameters are larger than the jet diameter. Breakup occurs many nozzle diameters downstream of nozzle. (b) First wind-induced regime. Drops with diameters of the order of jet diameter. Breakup occurs many nozzle diameters downstream of nozzle. (c) Second wind-induced regime. Drop sizes smaller than the jet diameter. Breakup starts some distance downstream of nozzle. (d) Atomization regime. Drop sizes much smaller than the jet diameter. Breakup starts at nozzle exit.

Liquid breakup characteristics such as the spray drop size, the jet breakup length, and the spray angle have been related to the unstable wave growth mechanism. The wavelengths and growth rates of the waves can be predicted using results from a linear stability analysis with[6]

$$\frac{\Lambda}{2} = 9.02\frac{\left(1+0.45Z^{0.5}\right)\left(1+0.4T^{0.7}\right)}{\left(1+0.87\,\mathrm{We}_2^{1.67}\right)^{0.6}} \tag{3.12.2a}$$

$$\Omega\left(\frac{\rho_1 a^3}{\sigma}\right)^{0.5} = \frac{0.34+0.38\,\mathrm{We}_2^{1.5}}{(1+Z)\left(1+1.4T^{0.6}\right)} \tag{3.12.2b}$$

where Λ is the wavelength, Ω is the growth rate of the most unstable surface wave, and a is the liquid jet radius. The maximum wave growth rate increases, and the corresponding wavelength decreases with increasing Weber number, $\mathrm{We}_2 = \rho_{gas}U^2a/\sigma$, where U is the relative velocity between the liquid and the gas. The liquid viscosity appears in the Ohnesorge number, $Z = \mathrm{We}_1^{1/2}/\mathrm{Re}_1$. Here, the Weber number We_1 is based on the liquid density, the Reynolds number is $\mathrm{Re}_1 = Ua/\nu_1$, ν_1 is the liquid viscosity, and the parameter $T = Z\mathrm{We}_2^{1/2}$. The wave growth rate is reduced and the wavelength is increased as the liquid viscosity increases.

The size of the drops formed from the breakup process is often assumed to be proportional to the wavelength of the unstable surface waves in modeling studies.[6] However, the drop sizes in the primary breakup region near the nozzle exist have also been found to be influenced by the length scale of the energy-containing eddies in the turbulent liquid flow.[7] There is uncertainty about atomization mechanisms since spray measurements are complicated by the high optical density of the spray in the breakup region

(e.g., see Figure 3.12.4(d)). As the drops penetrate into the ambient gas, they interact with each other through collisions and coalescence, and the spray drop size changes dynamically within the spray as a result of secondary breakup and vaporization effects. The drop trajectories are determined by complex drop drag, breakup, and vaporization phenomena, and by interactions with the turbulent gas flow.[6]

High-pressure diesel sprays are intermittent and are required to start and stop quickly without dribble between injections. This is accomplished by means of a plunger arrangement that is actuated by a cam and spring system in mechanical "jerk" pump systems (see Figure 3.12.5). Modern electronic injectors include electromagnetic solenoids that permit the duration and injection pressure to be varied independently of each other and of engine speed. Experiments on diesel-type injector nozzles show that the penetration distance, S, of the tip of the spray at time, t, after the beginning of the injection is given by[8]

$$\begin{aligned} S &= 0.39Ut\left(\rho_{\text{liquid}}/\rho_{\text{gas}}\right)^{1/2} && \text{for } t < t_b \\ S &= 2.46\sqrt{U\,dt}\left(\rho_{\text{liquid}}/\rho_{\text{gas}}\right)^{1/4} && \text{for } t > t_b \end{aligned} \tag{3.12.3}$$

where the "breakup time" is $t_b = 40.5d(\rho_{\text{liquid}}/\rho_{\text{gas}})^{1/2}/U$. The jet breakup length (see Figure 3.12.3(a)), $L = Ut_b$ is independent of the injection velocity. On the other hand, for low-speed jets, or for jets injected into a low-gas-density environment, $t_b = 1.04C\,(\rho_{\text{liquid}}\,d^3/\sigma)^{1/2}$, where C is a constant typically between 12 and 16 and σ is the surface tension. In this case L increases with the injection velocity.[9] The functional form of the above jet breakup time and length correlations can be derived for an inviscid liquid in the limits of large and small Weber number, We_2 from the unstable wave growth rate in Equation (3.12.2) with $t_b \sim \Omega^{-1}$.

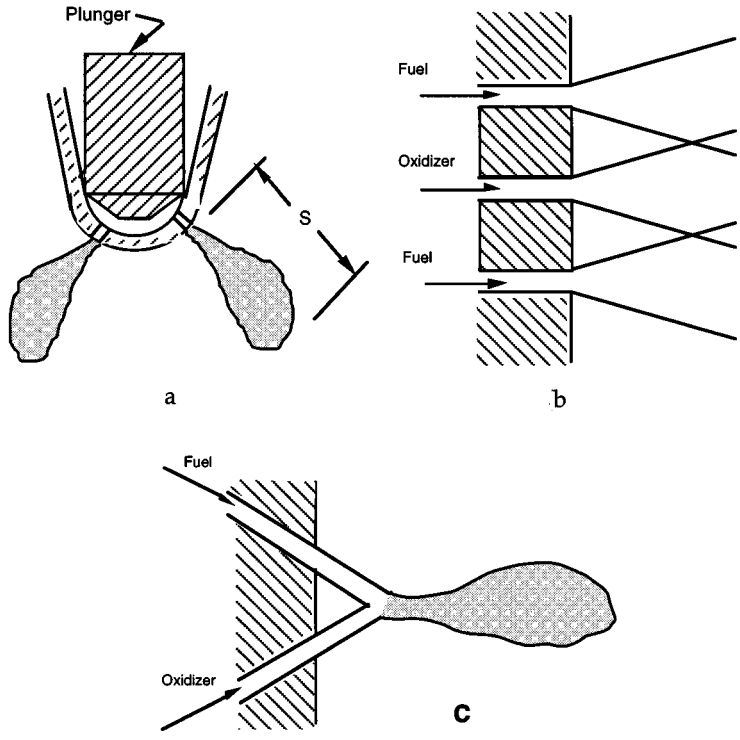


FIGURE 3.12.5 (a) Diesel injector multihole spray nozzle, (b) showerhead, and (c) doublet impingement nozzles.

For high-speed diesel-type jets in the atomization regime the resulting spray diverges in the form of a cone with cone angle, θ, that is usually in the range from 5 to 20°. θ increases with gas density following $\tan\theta = A(\rho_{gas}/\rho_{liquid})^{1/2}$, where A is a constant that depends on the nozzle passage length and (weakly) on the injection velocity.[9] Very high injection pressures are required to produce small drops. In diesel engines ΔP is typically as high as 200 Mpa, and drops are produced with mean diameters of the order of 10 μm (see Figure 3.12.1). Drop size correlations have been proposed for plain-orifice sprays, such as that presented in Table 3.12.1.[3] Note, however, that these correlations do not account for the fact that the spray drop size varies with time, and from place to place in the spray. Moreover, experimental correlations often do not include some parameters that are known to influence spray drop sizes, such as the nozzle passage length and its entrance geometry. Therefore, overall drop size correlations should only be used with caution.

TABLE 3.12.1 Representative Drop Size Correlations for Various Spray Devices (Dimensional quantities are in SI units, kg, m, s)

Device	Correlation	Notes
Plain orifice	$\text{SMD} = 3.08\nu_1^{0.385}\left(\rho_{liquid}\sigma\right)^{0.737}\rho_{gas}^{0.06}\,\Delta P^{-0.54}$	Use SI units
Fan spray	$\text{SMD} = 2.83 d_h\left(\sigma\mu_{liquid}^2/\rho_{gas}d_h^3\,\Delta P^2\right)^{0.25}$ $+ 0.26 d_h\left(\sigma\rho_{liquid}/\rho_{gas}d_h\,\Delta P\right)^{0.25}$	d_h = nozzle hydraulic diameter
Rotary atomizer	$\text{SMD} = 0.119Q^{0.1}\sigma^{0.5}/Nd^{0.5}\rho_{liquid}^{0.4}\mu_{liquid}^{0.1}$	N = rotational speed (rev/sec), Q = volumetric flow rate, A_{inj} U
Pressure swirl	$\text{SMD} = 4.52\left(\sigma\mu_{liquid}^2/\rho_{gas}\,\Delta P^2\right)^{0.25}\left(t\cos\theta\right)^{0.25}$ $+ 0.39\left(\sigma\rho_{liquid}/\rho_{gas}\,\Delta P\right)^{0.25}\left(t\cos\theta\right)^{0.75}$ $t = 0.0114A_{inj}\rho_{liquid}^{1/2}d\cos\theta$	t = film thickness; θ = cone angle, d = discharge orifice diameter
Twin fluid/air blast	$\text{SMD} = 0.48d\left(\sigma/\rho_{gas}U^2d\right)^{0.4}\left(1+1/\text{ALR}\right)^{0.4}$ $+ 0.15d\left(\mu_{liquid}^2/\sigma\rho_{liquid}d\right)^{0.5}\left(1+1/\text{ALR}\right)$	ALR = air-to-liquid mass ratio
Prefilming air blast	$\text{SMD} = \left(1+1/\text{ALR}\right)\left[0.33d_h\left(\sigma/\rho_{gas}U^2d_p\right)^{0.6}\right.$ $\left.+ 0.068d_h\left(\mu_{liquid}^2/\sigma\rho_{liquid}d_p\right)^{0.5}\right]$	d_h = hydraulic diameter, d_p = prefilmer diameter, Figure 3.12.9
Ultrasonic	$\text{SMD} = \left(4\pi^3\sigma/\rho_{liquid}\omega^2\right)^{1/3}$	ω = vibration frequency

Source: Lefebvre, A.H., *Atomization and Sprays*, Hemisphere Publishing, New York, 1989. With permission.

The plain orifice design is also used in twin-fluid-type liquid rocket engines in showerhead and doublet designs (Figures 3.12.5b and 3.12.5c). In the case of doublet nozzles, shown in Figure 3.12.6c, the impinging jets create unstable liquid sheets which break up to produce the sprays. Drop size correlations are available for liquid sheets such as those formed by discharging the liquid through a rectangular slit (see *fan spray*, Table 3.12.1). Thin liquid sheets or slits lead to the production of small drops. The breakup mechanism of liquid sheets is also thought to involve the unstable growth of surface waves due to surface tension and aerodynamic forces.[5]

In *rotary atomizers* centrifugal forces are used to further enhance the breakup process. In this case the liquid is supplied to the center of a spinning disk and liquid sheets or ligaments are thrown off the edges of the disk. The drop size depends on the rotational speed of the disk, as indicated in Table 3.12.1.

A spinning wheel or cup (turbobell) is used in some spray-painting applications. The spray shape is controlled by supplying a coflowing stream of "shaping-air."

Centrifugal forces also play a role in the breakup mechanism of *pressure swirl* atomizers (*simplex* nozzles). These atomizers give wider spray cone angle than plain orifice nozzles, and are available in hollow cone and solid cone designs. As depicted in Figure 3.12.3(b) the spray liquid enters a swirl chamber tangentially to create a swirling liquid sheet. The air core vortex within the swirl chamber plays an important role in determining the thickness of the liquid sheet or film at the nozzle exit. This type of nozzle produces relatively coarse sprays. A representative SMD correction is listed in Table 3.12.1. The spray cone angle depends on the ratio of the axial and tangential liquid velocity components at the exit of the nozzle. This type of atomizer is not well suited for use in transient applications because it tends to dribble at start-up and to shut down when the air core is not fully formed.

The basic drawback of all pressure atomizers is that the flow rate depends on the square root of ΔP. The volumetric flow rate is $Q = A_{inj}\, U$, where A_{inj} is the liquid flow area at the nozzle exit, so that a factor of 20 increase in flow rate (a typical turndown ratio from idle to full load operation of a gas turbine engine) requires a factor of 400 increase in injection pressure.

This difficulty has led to so-called wide-range atomizer designs such as those shown in Figure 3.12.6. The *duplex* nozzle features two sets of tangential swirl ports; the primary (or pilot) supplies fuel at low flow rates, while the secondary ports become operational at high flow rates. Another variation is the *dual-orifice* nozzle which is conceptually two simplex nozzles arranged concentrically, one supplying the primary flow and the other supplying the secondary flow. The *spill-return* nozzle is a simplex nozzle with a rear passage that returns fuel to the injection pump. In this design the flow rate is controlled by the relative spill amount, and there are no small passages to become plugged. However, the fuel is always supplied at the maximum pressure which increases the demands on the injection pump. But high swirl is always maintained in the swirl chamber and good atomization is achieved even at low flow rates.

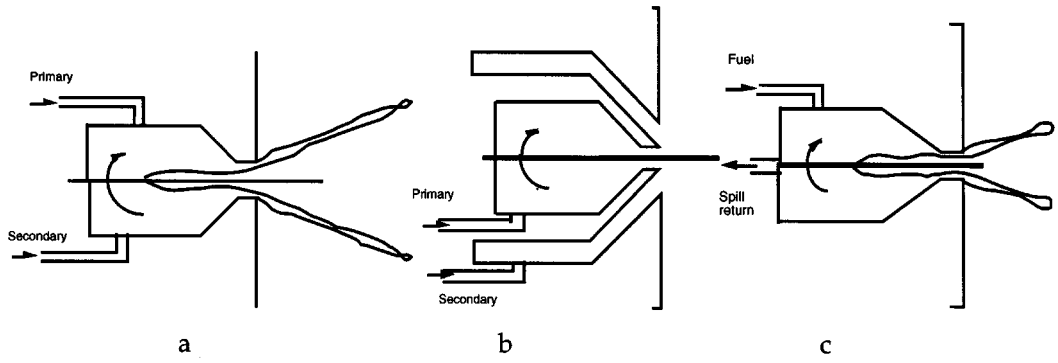


FIGURE 3.12.6 (a) Duplex, (b) dual orifice, and (c) spill-return-type nozzle designs.

In *twin-fluid injectors* atomization is aided by a flow of high-velocity gas through the injector passages. The high-velocity gas stream impinges on a relatively low-velocity liquid either internally (in *internal-mixing* nozzles, Figure 3.12.7) or externally (in *external-mixing* designs, Figure 3.12.8). The liquid and gas flows are typically swirled in opposite directions by means of swirl vanes to improve atomization. *Air-assist* refers to designs that use a relatively small amount of air at high (possibly sonic) velocities. *Air-blast* refers to designs that use large quantities of relatively low-velocity air which often supplies some of the air to help decrease soot formation in combustion systems[3]. (see Figure 3.12.9.)

In *flashing* and *effervescent* atomizers a two-phase flow is passed through the injector nozzle exit. In the former the bubbles are generated by means of a phase change which occurs as the liquid, containing a dissolved propellant gas or vapor, undergoes the pressure drop through the nozzle. This process is exploited in many household spray cans, but has the disadvantage of releasing the propellant gas required for atomization into the atmosphere. In the so-called effervescent atomizer, air bubbles are introduced

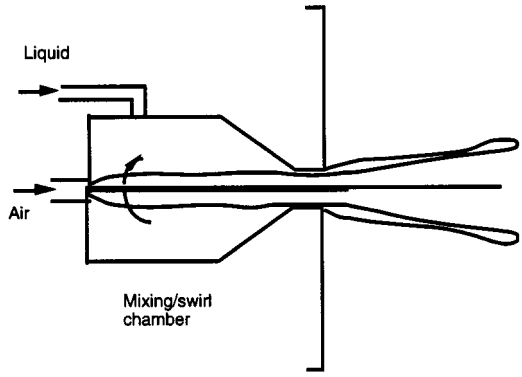


FIGURE 3.12.7 Internal-mixing twin-fluid injector design.

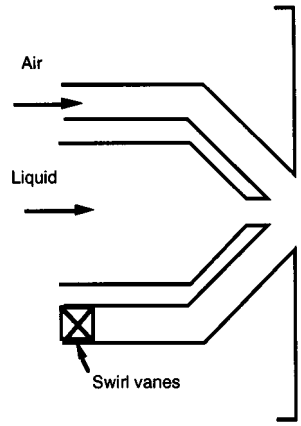


FIGURE 3.12.8 External-mixing twin-fluid injector design.

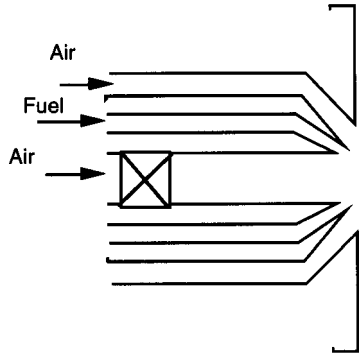


FIGURE 3.12.9 Prefilming air blast atomizer.

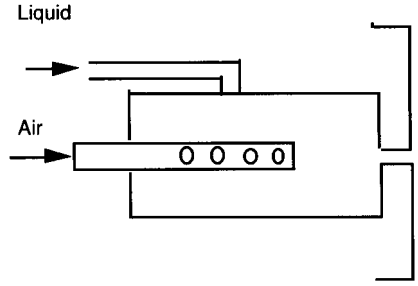


FIGURE 3.12.10 Internal-mixing, effervescent atomizer.

into the liquid upstream of the exit orifice, as depicted in Figure 3.12.10. The spray quality is found to depend weakly on the air bubble size and is independent of the nozzle exit diameter. This makes internal-mixing, air-assist atomizers very attractive for use with high-viscosity fluids and slurries where nozzle plugging would otherwise be a problem.[3]

In *electrostatic* atomizers the spray liquid is charged by applying a high-voltage drop across the nozzle. The dispersion of the spray drops is increased by exploiting electrical repulsive forces between the droplets. An electrostatic charge on the drops is also helpful in spray-coating applications, such as in automotive spray painting using electrostatic turbobell sprayers, since the charged drops are attracted to an oppositely charged target surface.

Other atomizer types include *vibratory* and *ultrasonic* atomizers (or *nebulizers*), where the drops are formed by vibrating the injector nozzle at high frequencies and at large amplitudes to produce short-wavelength disturbances to the liquid flow. Ultrasonic atomizers are used in inhalation therapy where very fine sprays (submicron sizes) are required, and an available representative drop size correlation is also listed in Table 3.12.1.

References

1. American Society for Testing and Materials (ASTM) Standard E799. 1988. Data Criteria and Processing for Liquid Drop Size Analysis.
2. Mugele, R. and Evans, H.D. 1951. Droplet size distributions in sprays, *Ind. Eng. Chem.*, 43, 1317–1324.
3. Lefebvre, A.H. 1989. *Atomization and Sprays*, Hemisphere Publishing, New York.
4. Chigier, N.A. 1983. Drop size and velocity instrumentation, *Prog. Energ. Combust. Sci.*, 9, 155–177.
5. Chigier, N. and Reitz, R.D. 1996. Regimes of jet breakup, in *Progress in Astronautics and Aeronautics Series*, K. Kuo, Ed., AIAA, New York, Chapter 4, pp. 109–135.
6. Reitz, R.D. 1988. Modeling atomization processes in high-pressure vaporizing sprays, *Atomisation Spray Technol.,* 3, 309–337.
7. Wu, P-K., Miranda, R.F., and Faeth, G.M. 1995. Effects of initial flow conditions on primary breakup of nonturbulent and turbulent round liquid jets, *Atomization Sprays*, 5, 175–196.
8. Hiroyasu, H. and Arai, M. 1978. Fuel spray penetration and spray angle in diesel engines, *Trans. JSAE*, 34, 3208.
9. Reitz, R.D. and Bracco, F.V. 1986. Mechanisms of breakup of round liquid jets, in *The Encyclopedia of Fluid Mechanics,* Vol. 3, N. Cheremisnoff, Ed., Gulf Publishing, Houston, TX, Chapter 10, 233–249.

Further Information

Information about recent work in the field of atomization and sprays can be obtained through participation in the Institutes for Liquid Atomization and Spraying Systems (ILASS-Americas, -Europe, -Japan, -Korea). These regional ILASS sections hold annual meetings. An international congress (ICLASS) is also held biennially. More information is available on the ILASS-Americas homepage at http://ucicl.eng.uci.edu/ilass. Affiliated with the ILASS organizations is the Institute's Journal publication *Atomization and Sprays* published by Begell House, Inc., New York.

3.13 Flow Measurement*

Alan T. McDonald and Sherif A. Sherif

This section deals with the measurement of mass flow rate or volume flow rate of a fluid in an enclosed pipe or duct system. Flow measurement in open channels is treated in Section 3.10 of this book.

The choice of a flow meter type and size depends on the required accuracy, range, cost, ease of reading or data reduction, and service life. Always select the simplest and cheapest device that gives the desired accuracy.

Direct Methods

Tanks can be used to determine the flow rate for steady liquid flows by measuring the volume or mass of liquid collected during a known time interval. If the time interval is long enough, flow rates may be determined precisely using tanks. Compressibility must be considered in gas volume measurements. It is not practical to measure the mass of gas, but a volume sample can be collected by placing an inverted "bell" over water and holding the pressure constant by counterweights. No calibration is required when volume measurements are set up carefully; this is a great advantage of direct methods.

Positive-displacement **flow meters** may be used in specialized applications, particularly for remote or recording uses. For example, household water and natural gas meters are calibrated to read directly in units of product. Gasoline metering pumps measure total flow and automatically compute the cost. Many positive-displacement meters are available commercially. Consult manufacturers' literature or Reference 7 for design and installation details.

Restriction Flow Meters for Flow in Ducts

Most restriction flow meters for internal flow (except the laminar flow element) are based on acceleration of a fluid stream through some form of nozzle, shown schematically in Figure 3.13.1. Flow separating from the sharp edge of the nozzle throat forms a recirculation zone shown by the dashed lines downstream from the nozzle. The main flow stream continues to accelerate from the nozzle throat to form a *vena contracta* at section (2) and then decelerates again to fill the duct. At the vena contracta, the flow area is a minimum, the flow streamlines are essentially straight, and the pressure is uniform across the channel section. The theoretical flow rate is

$$\dot{m}_{\text{theoretical}} = \frac{A_2}{\sqrt{1-\left(A_2/A_1\right)^2}}\sqrt{2\rho\left(p_1 - p_2\right)} \tag{3.13.1}$$

Equation (3.13.1) shows the general relationship for a **restriction flow meter**: Mass flow rate is proportional to the square root of the pressure differential across the meter taps. This relationship limits the flow rates that can be measured accurately to approximately a 4:1 range.

Several factors limit the utility of Equation (3.13.1) for calculating the actual mass flow rate through a meter. The actual flow area at section (2) is unknown when the vena contracta is pronounced (e.g., for orifice plates when D_t is a small fraction of D_1). The velocity profiles approach uniform flow only at large Reynolds number. Frictional effects can become important (especially downstream from the meter) when the meter contours are abrupt. Finally, the location of the pressure taps influences the differential pressure reading, $p_1 - p_2$.

The actual mass flow rate is given by

* The contributors of this chapter were asked to conform as much as possible to the nomenclature presented below, but define new terms pertaining to their particular area of specialty in the context of the chapter.

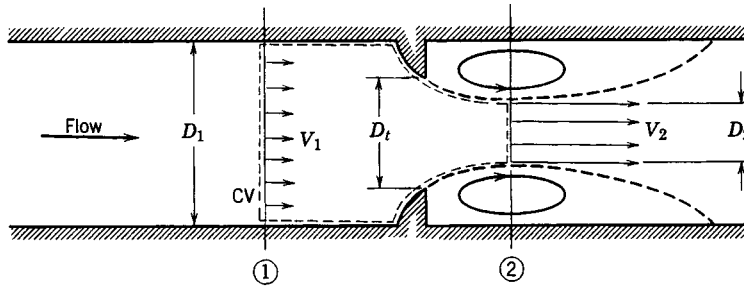


FIGURE 3.13.1 Internal flow through a generalized nozzle, showing control volume used for analysis.

$$\dot{m}_{\text{actual}} = \frac{CA_t}{\sqrt{1-\left(A_t/A_1\right)^2}}\sqrt{2\rho\left(p_1 - p_2\right)} \tag{3.13.2}$$

where C is an empirical *discharge coefficient*.

If $\beta = D_t/D_1$, then $(A_t/A_1)^2 = (D_t/D_1)^4 = \beta^4$, and

$$\dot{m}_{\text{actual}} = \frac{CA_t}{\sqrt{1-\beta^4}}\sqrt{2\rho\left(p_1 - p_2\right)} \tag{3.13.3}$$

where $1/(1 - \beta^4)^{1/2}$ is the *velocity correction factor*. Combining the discharge coefficient and velocity correction factor into a single *flow coefficient*,

$$K \equiv \frac{C}{\sqrt{1-\beta^4}} \tag{3.13.4}$$

yields the mass flow rate in the form:

$$\dot{m}_{\text{actual}} = KA_t\sqrt{2\rho\left(p_1 - p_2\right)} \tag{3.13.5}$$

Test data can be used to develop empirical equations to predict flow coefficients vs. pipe diameter and Reynolds number for standard metering systems. The accuracy of the equations (within specified ranges) is often adequate to use the meter without calibration. Otherwise, the coefficients must be measured experimentally.

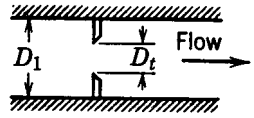
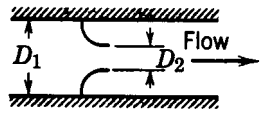
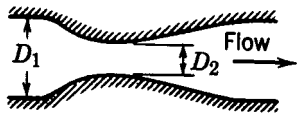
For the turbulent flow regime ($\text{Re}_D > 4000$), the flow coefficient may be expressed by an equation of the form:[7]

$$K = K_\infty + \frac{1}{\sqrt{1-\beta^4}}\frac{b}{\text{Re}_{D_1}^n} \tag{3.13.6}$$

where subscript ∞ denotes the flow coefficient at infinite Reynolds number and constants b and n allow for scaling to finite Reynolds numbers. Correlating equations and curves of flow coefficients vs. Reynolds number are given for specific metering elements in the next three subsections following the general comparison of the characteristics of orifice plate, flow nozzle, and venturi meters in Table 3.13.1 (see Reference 4).

Flow meter coefficients reported in the literature have been measured with fully developed turbulent velocity distributions at the meter inlet (Section 3.1). When a flow meter is installed downstream from

TABLE 3.13.1.Characteristics of Orifice, Flow Nozzle, and Venturi Flow Meters

Flow Meter Type	Diagram	Head Loss	Cost
Orifice		High	Low
Flow nozzle		Intermediate	Intermediate
Venturi		Low	High

a valve, elbow, or other disturbance, a straight section of pipe must be placed in front of the meter. Approximately 10 diameters of straight pipe upstream are required for venturi meters, and up to 40 diameters for orifice plate or flow nozzle meters. Some design data for incompressible flow are given below. The same basic methods can be extended to compressible flows.[7]

Orifice Plates

The orifice plate (Figure 3.13.2) may be clamped between pipe flanges. Since its geometry is simple, it is low in cost and easy to install or replace. The sharp edge of the orifice will not foul with scale or suspended matter. However, suspended matter can build up at the inlet side of a concentric orifice in a horizontal pipe; an eccentric orifice may be placed flush with the bottom of the pipe to avoid this difficulty. The primary disadvantages of the orifice are its limited capacity and the high permanent head loss caused by uncontrolled expansion downstream from the metering element.

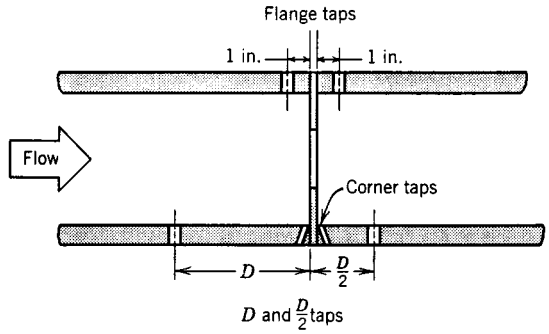


FIGURE 3.13.2 Orifice geometry and pressure tap locations.

Pressure taps for orifices may be placed in several locations as shown in Figure 3.13.2 (see Reference 7 for additional details). Since the location of the pressure taps influences the empirically determined flow coefficient, one must select handbook values of K consistent with the pressure tap locations.

The correlating equation recommended for a concentric orifice with corner taps is

$$C = 0.5959 + 0.0312\beta^{2.1} - 0.184\beta^{8} + \frac{91.71\beta^{2.5}}{\mathrm{Re}_{D_1}^{0.75}} \tag{3.13.7}$$

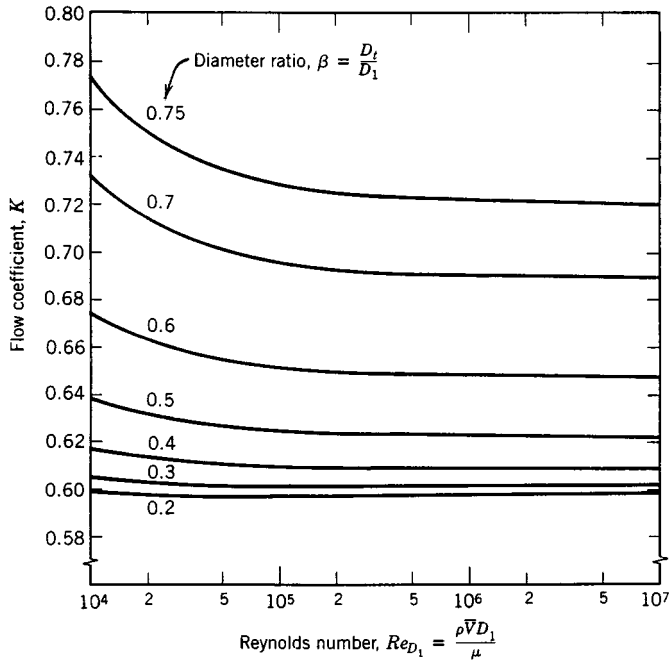


FIGURE 3.13.3 Flow coefficients for concentric orifices with corner taps.

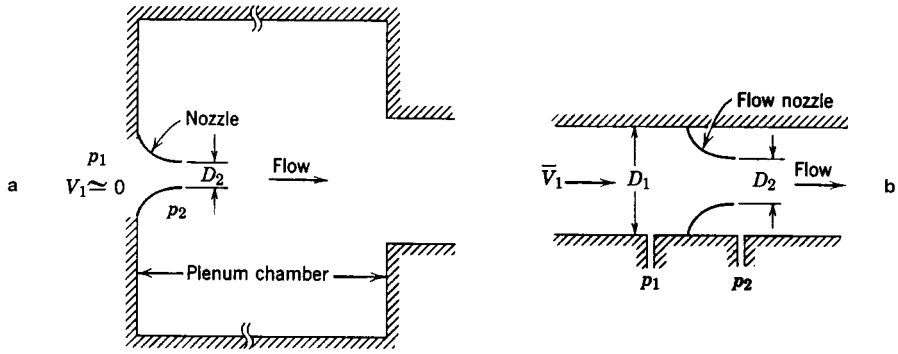


FIGURE 3.13.4 Typical installations of nozzle flow meters. (a) In plenum, (b) In duct.

Equation (3.13.7) predicts orifice discharge coefficients within ± 0.6% for $0.2 < \beta < 0.75$ and for $10^4 < \mathrm{Re}_{D_1} < 10^7$. Some flow coefficients calculated from Equation (3.13.7) are presented in Figure 3.13.3. Flow coefficients are relatively insensitive to Reynolds number for $\mathrm{Re}_{D_1} > 10^5$ when $\beta > 0.5$.

A similar correlating equation is available for orifice plates with D and $D/2$ taps. Flange taps require a different correlation for every line size. Pipe taps, located at $2^1/_2\,D$ and $8D$, no longer are recommended.

Flow Nozzles

Flow nozzles may be used as metering elements in either plenums or ducts, as shown in Figure 3.13.4; the nozzle section is approximately a quarter ellipse. Design details and recommended locations for pressure taps are given in Reference 7.

The correlating equation recommended for an ASME long-radius flow nozzles[7] is

$$C = 0.9975 - \frac{6.53\beta^{0.5}}{\mathrm{Re}_{D_1}^{0.5}} \tag{3.13.8}$$

Equation (3.13.8) predicts discharge coefficients for flow nozzles within ± 2.0% for $0.25 < \beta < 0.75$ for $10^4 < \mathrm{Re}_{D_1} < 10^7$. Some flow coefficients calculated from Equation 3.13.8 are presented in Figure 3.13.5. (K can be greater than 1 when the velocity correction factor exceeds 1.) For plenum installation, nozzles may be fabricated from spun aluminum, molded fiberglass, or other inexpensive materials. Typical flow coefficients are in the range $0.95 < K < 0.99$; the larger values apply at high Reynolds numbers. Thus, the mass flow rate can be computed within approximately ± 2% using $K = 0.97$.

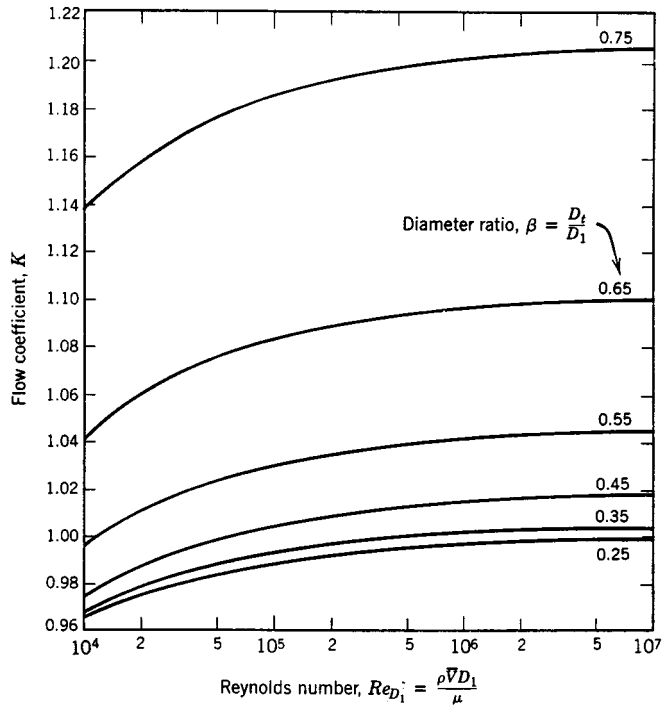


FIGURE 3.13.5 Flow coefficients for ASME long-radius flow nozzles.

Venturis

Venturi meters are generally made from castings machined to close tolerances to duplicate the performance of the standard design, so they are heavy, bulky, and expensive. The conical diffuser section downstream from the throat gives excellent pressure recovery; overall head loss is low. Venturi meters are self-cleaning because of their smooth internal contours.

Experimentally measured discharge coefficients for venturi meters range from 0.980 to 0.995 at high Reynolds numbers ($\mathrm{Re}_{D_1} > 2 \times 10^5$). Thus, $C = 0.99$ can be used to calculate mass flow rate within about ±1% at high Reynolds number.[7] Consult manufacturers' literature for specific information at Reynolds numbers below 10^5.

Orifice plates, flow nozzles, and venturis all produce pressure drops proportional to flow rate squared, according to Equation 3.13.4. In practice, a meter must be sized to accommodate the largest flow rate expected. Because the pressure drop vs. flow rate relationship is nonlinear, a limited range of flow rate

can be measured accurately. Flow meters with single throats usually are considered for flow rates over a 4:1 range.[7]

Unrecoverable head loss across a metering element may be expressed as a fraction of the differential pressure across the element. Unrecoverable head losses are shown in Figure 3.13.6.[7]

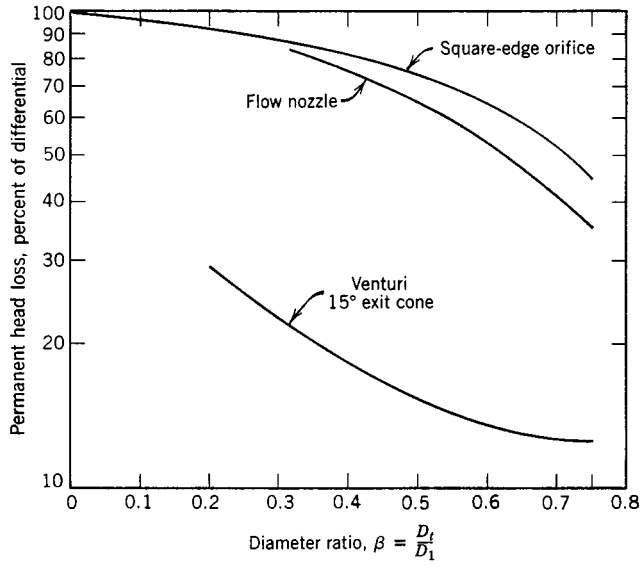


FIGURE 3.13.6 Permanent head loss produced by various flow metering elements.

Laminar Flow Elements

The laminar flow element (LFE)* produces a pressure differential proportional to flow rate. The LFE contains a metering section subdivided into many passages, each of small enough diameter to assure fully developed laminar flow. Because the pressure drop in laminar duct flow is directly proportional to flow rate, the pressure drop vs. flow rate relationship is linear. The LFE may be used with reasonable accuracy over a 10:1 flow rate range. Because the relationship between pressure drop and flow rate for laminar flow depends on viscosity, which is a strong function of temperature, fluid temperature must be known in order to obtain accurate metering.

An LFE costs approximately as much as a venturi, but is much lighter and smaller. Thus, the LFE is widely used in applications where compactness and extended range are important.

Linear Flow Meters

Several flow meter types produce outputs proportional to flow rate. Some meters produce signals without the need to measure differential pressure. The most common linear flow meters are discussed briefly below.

Float meters indicate flow rate directly for liquids or gases. An example is shown in Figure 3.13.7. The float is carried upward in the tapered clear tube by the flowing fluid until drag force and float weight are in equilibrium. Float meters (often called rotameters) are available with factory calibration for a number of common fluids and flow rate ranges.

A *turbine flow meter* is a free-running, vaned impeller mounted in a cylindrical section of tube (Figure 3.13.8). With proper design, the rate of rotation of the impeller may be made proportional to volume

* Patented and manufactured by Meriam Instrument Co., 10920 Madison Ave., Cleveland, OH.

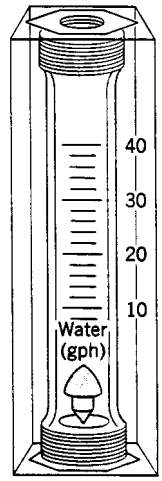


FIGURE 3.13.7 Float-type variable-area flow meter. (Courtesy of Dwyer Instrument Co., Michigan City, IN.)

flow rate over as much as a 100:1 flow rate range. Rotational speed of the turbine element can be sensed using a magnetic or modulated carrier pickup external to the meter. This sensing method therefore requires no penetrations or seals in the duct. Thus, turbine flow meters can be used safely to measure flow rates in corrosive or toxic fluids. The electrical signal can be displayed, recorded, or integrated to provide total flow information.

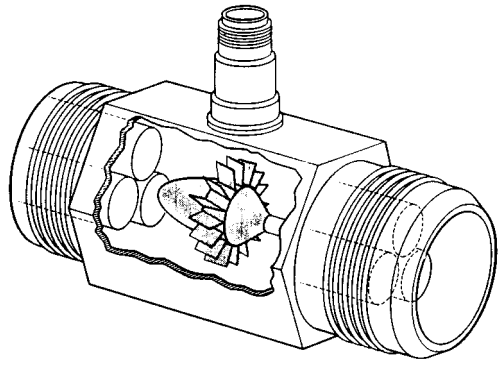


FIGURE 3.13.8 Turbine flow meter. (Courtesy of Potter Aeronautical Corp., Union, NJ.)

Vortex shedding from a bluff obstruction may be used to meter flow. Since Strouhal number, $St = fL/V$, is approximately constant ($St \cong 0.21$), vortex shedding frequency f is proportional to flow velocity. Vortex shedding causes velocity and pressure changes. Pressure, thermal, or ultrasonic sensors may be used to detect the vortex shedding frequency, and thus to infer the fluid velocity. (The velocity profile does affect the constancy of the shedding frequency.) Vortex flow meters can be used over a 20:1 flow rate range.[7]

Electromagnetic flow meters create a magnetic field across a pipe. When a conductive fluid passes through the field, a voltage is generated at right angles to the field and velocity vectors. Electrodes placed on a pipe diameter detect the resulting signal voltage, which is proportional to the average axial velocity when the profile is axisymmetric. The minimum flow speed should be above about 0.3 m/sec, but there are no restrictions on Reynolds number. The flow rate range normally quoted is 10:1.[7]

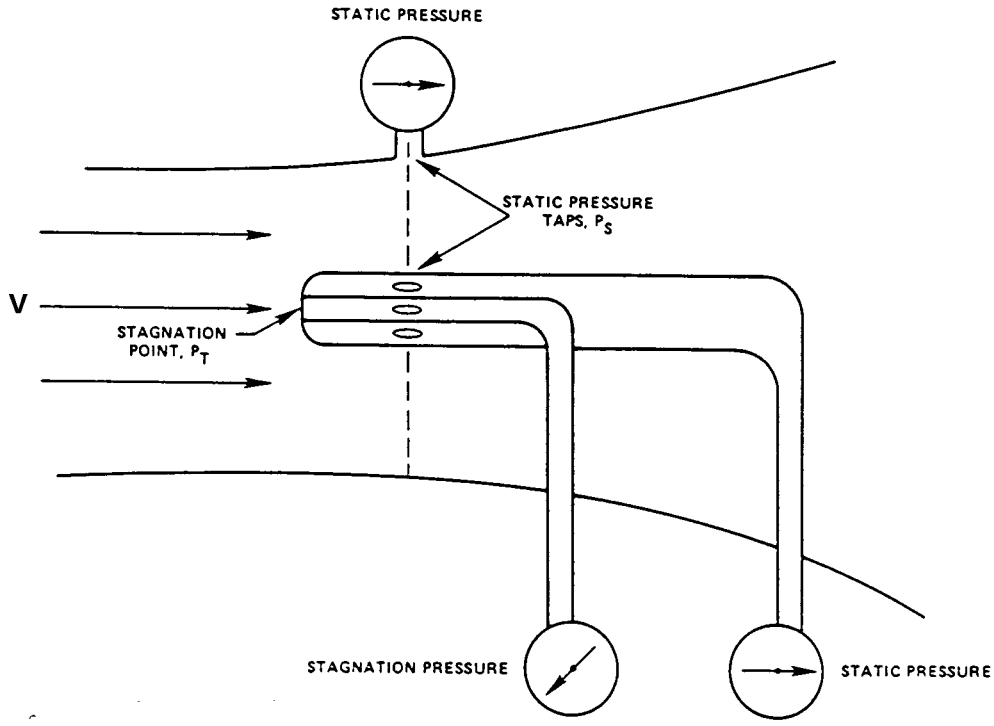
Ultrasonic flow meters also respond to average velocity at a pipe cross section. Two principal types of ultrasonic meters measure propagation time for clean liquids, and reflection frequency shift (Doppler effect) for flows carrying particulates. The speed of an acoustic wave increases in the flow direction and decreases when transmitted against the flow. For clean liquids, an acoustic path inclined to the pipe axis is used to infer flow velocity. Multiple paths are used to estimate volume flow rate accurately.

Doppler effect ultrasonic flow meters depend on reflection of sonic waves (in the megahertz range) from scattering particles in the fluid. When the particles move at flow speed, the frequency shift is proportional to flow speed; for a suitably chosen path, output is proportional to volume flow rate; ultrasonic meters may require calibration in place. One or two transducers may be used; the meter may be clamped to the outside of the pipe. Flow rate range is 10:1.[7]

A relatively new type of true mass flow meter is based on the effect of *Coriolis acceleration* on natural frequency of a bent tube carrying fluid. The bent tube is excited and its vibration amplitude measured. The instrument measures mass flow rate directly, and thus is ideal for two-phase or liquid–solid flow measurements. Pressure drop of the Coriolis meter may be high, but its useful flow rate range is 100:1.

Traversing Methods

In situations such as in air handling or refrigeration equipment, it may be impractical or impossible to install a fixed flow meter, but it may be possible to measure flow rate using a traversing technique.[1] To measure flow rate by **traverse**, the duct cross section is subdivided into segments of equal area. The fluid velocity is measured at the center of each area segment using a pitot tube (see Figure 3.13.9), a total head tube, or a suitable anemometer. The volume flow rate for each segment is approximated by the product of the measured velocity and segment area. Flow rate through the entire duct is the sum of these segmental flow rates. For details of recommended procedures see Reference 1 or 6.



$$V = \sqrt{2(P_T - P_S)/\rho}$$

FIGURE 3.13.9 Pitot tube.

Use of probes for traverse measurements requires direct access to the flow field. Pitot tubes give uncertain results when pressure gradients or streamline curvature are present, and they respond slowly. Two types of anemometers — **thermal anemometers** and **laser Doppler anemometers** — partially overcome these difficulties, although they introduce new complications.

Thermal anemometers use electrically heated tiny elements (either hot-wire or hot-film elements). Sophisticated feedback circuits are used to maintain the temperature of the element constant and to sense the input heating rate. The heating rate is related to the local flow velocity by calibration. The primary advantage of thermal anemometers is the small size of the sensing element. Sensors as small as 0.002 mm in diameter and 0.1 mm long are available commercially. Because the thermal mass of such tiny elements is extremely small, their response to fluctuations in flow velocity is rapid. Frequency responses to the 50-kHz range have been quoted.[3] Thus, thermal anemometers are ideal for measuring turbulence quantities. Insulating coatings may be applied to permit their use in conductive or corrosive gases or liquids.

Because of their fast response and small size, thermal anemometers are used extensively for research. Numerous schemes for treating the resulting data have been published. Digital processing techniques, including fast Fourier transforms, can be used to obtain mean values and moments, and to analyze signal frequency content and correlations.

Laser Doppler anemometers (LDAs) can be used for specialized applications where direct physical access to the flow field is difficult or impossible.[5] Laser beam(s) are focused to a small volume in the flow at the location of interest; laser light is scattered from particles present in the flow or introduced for this purpose. A frequency shift is caused by the local flow speed (Doppler effect). Scattered light and a reference beam are collected by receiving optics. The frequency shift is proportional to the flow speed; this relationship may be calculated, so there is no need for calibration. Since velocity is measured directly, the signal is unaffected by changes in temperature, density, or composition in the flow field. The primary disadvantages of LDAs are the expensive and fragile optical equipment and the need for careful alignment.

Viscosity Measurements

Viscometry is the technique of measuring the viscosity of a fluid. Viscometers are classified as rotational, capillary, or miscellaneous, depending on the technique employed. Rotational viscometers use the principle that a rotating body immersed in a liquid experiences a viscous drag which is a function of the viscosity of the liquid, the shape and size of the body, and the speed of its rotation. Rotational viscometers are widely used because measurements can be carried out for extended periods of time. Several types of viscometers are classified as rotational and Figure 3.13.10 is a schematic diagram illustrating a typical instrument of this type.

Capillary viscometry uses the principle that when a liquid passes in laminar flow through a tube, the viscosity of the liquid can be determined from measurements of the volume flow rate, the applied pressure, and the tube dimensions. Viscometers that cannot be classified either as rotational or capillary include the falling ball viscometer. Its method of operation is based on Stokes' law which relates the viscosity of a Newtonian fluid to the velocity of a sphere falling in it. Falling ball viscometers are often employed for reasonably viscous fluids. Rising bubble viscometers utilize the principle that the rise of an air bubble through a liquid medium gives a visual measurement of liquid viscosity. Because of their simplicity, rising bubble viscometers are commonly used to estimate the viscosity of varnish, lacquer, and other similar media.

Defining Terms

Flow meter: Device used to measure mass flow rate or volume flow rate of fluid flowing in a duct.

Restriction flow meter: Flow meter that causes flowing fluid to accelerate in a nozzle, creating a pressure change that can be measured and related to flow rate.

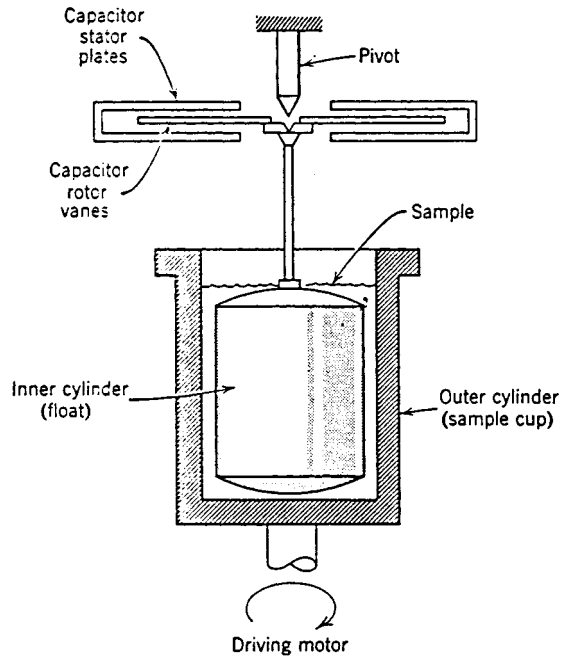


FIGURE 3.13.10 Rotational viscometer.

Thermal anemometer: Heated sensor used to infer local fluid velocity by sensing changes in heat transfer from a small electrically heated surface exposed to the fluid flow.

Traverse: Systematic procedure used to traverse a probe across a duct cross-section to measure flow rate through the duct.

References

1. *ASHRAE Handbook Fundamentals*. 1981. American Society of Heating, Refrigerating, and Air Conditioning Engineers, Atlanta, GA.
2. Baker, R.C. *An Introductory Guide to Flow Measurement*. 1989. Institution of Mechanical Engineers, London.
3. Bruun, H.H. 1995. *Hot-Wire Anemometry: Principles and Signal Analysis*. Oxford University Press, New York.
4. Fox, R.W. and McDonald, A.T. 1992. *Introduction to Fluid Mechanics*, 4th ed., John Wiley & Sons, New York.
5. Goldstein, R.J., Ed. 1996. *Fluid Mechanics Measurements*, 2nd ed., Taylor and Francis, Bristol, PA.
6. ISO 7145, *Determination of Flowrate of Fluids in Closed Conduits or Circular Cross Sections Method of Velocity Determination at One Point in the Cross Section*, ISO UDC 532.57.082.25:532.542. International Standards Organization, Geneva, 1982.
7. Miller, R.W. 1996. *Flow Measurement Engineering Handbook*, 3rd ed., McGraw-Hill, New York.
8. Spitzer, R.W., Ed. 1991. *Flow Measurement: A Practical Guide for Measurement and Control*. Instrument Society of America, Research Triangle Park, NC.
9. White, F.M. 1994. *Fluid Mechanics*, 3rd ed., McGraw-Hill, New York.

Further Information

This section presents only a survey of flow measurement methods. The references contain a wealth of further information. Baker[2] surveys the field and discusses precision, calibration, probe and tracer methods, and likely developments. Miller[7] is the most comprehensive and current reference available for information on restriction flow meters. Goldstein[5] treats a variety of measurement methods in his recently revised and updated book. Spitzer[8] presents an excellent practical discussion of flow measurement. Measurement of viscosity is extensively treated in *Viscosity and Flow Measurement: A Laboratory Handbook of Rheology*, by Van Wazer, J.R., Lyons, J.W., Kim, K.Y., and Colwell, R.E., Interscience Publishers, John Wiley & Sons, New York, 1963.

3.14 Micro/Nanotribology

Bharat Bhushan

Introduction

The emerging field of micro/nanotribology is concerned with processes ranging from atomic and molecular scales to microscale, occurring during adhesion, friction, wear, and thin-film lubrication at sliding surfaces (Bhushan, 1995, 1997; Bhushan et al., 1995). The differences between conventional tribology or macrotribology and micro/nanotribology are contrasted in Figure 3.14.1. In macrotribology, tests are conducted on components with relatively large mass under heavily loaded conditions. In these tests, wear is inevitable and the bulk properties of mating components dominate the tribological performance. In micro/nanotribology, measurements are made on at least one of the mating components, with relatively small mass under lightly loaded conditions. In this situation, negligible wear occurs and the surface properties dominate the tribological performance.

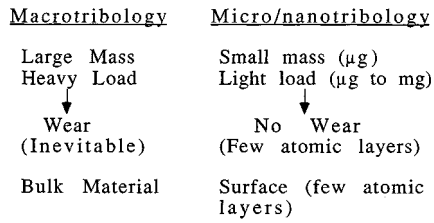


FIGURE 3.14.1 Comparison between macrotribology and microtribology.

Micro/nanotribological investigations are needed to develop fundamental understanding of interfacial phenomena on a small scale and to study interfacial phenomena in the micro- and nanostructures used in magnetic storage systems, microelectromechanical systems (MEMS), and other industrial applications (Bhushan, 1995, 1996, 1997). Friction and wear of lightly loaded micro/nanocomponents are highly dependent on the surface interactions (few atomic layers). These structures are generally lubricated with molecularly thin films. Micro- and nanotribological studies are also valuable in fundamental understanding of interfacial phenomena in macrostructures to provide a bridge between science and engineering (Bowden and Tabor, 1950, 1964; Bhushan and Gupta, 1997; Bhushan, 1996).

In 1985, Binnig et al. (1986) developed an "atomic force microscope" (AFM) to measure ultrasmall forces (less than 1 μN) present between the AFM tip surface and the sample surface. AFMs can be used for measurement of *all engineering surfaces* of any surface roughness, which may be either electrically conducting or insulating. AFM has become a popular surface profiler for topographic measurements on micro- to nanoscale. These are also used for scratching, wear, and nanofabrication purposes. AFMs have been modified in order to measure both normal and friction forces and this instrument is generally called a friction force microscope (FFM) or a lateral force microscope (LFM). New transducers in conjunction with an AFM can be used for measurements of elastic/plastic mechanical properties (such as load-displacement curves, indentation hardness, and modulus of elasticity) (Bhushan et al., 1996). A surface force apparatus (SFA) was first developed in 1969 (Tabor and Winterton, 1969) to study both static and dynamic properties of molecularly thin liquid films sandwiched between two molecularly smooth surfaces. SFAs are being used to study rheology of molecularly thin liquid films; however, the liquid under study has to be confined between molecularly smooth surfaces with radii of curvature on the order of 1 mm (leading to poorer lateral resolution as compared with AFMs) (Bhushan, 1995). Only AFMs/FFMs can be used to study *engineering surfaces* in the *dry and wet conditions* with *atomic resolution*. The scope of this section is limited to the applications of AFMs/FFMs.

At most solid–solid interfaces of technological relevance, contact occurs at numerous asperities with a range of radii; a sharp AFM/FFM tip sliding on a surface simulates just one such contact. Surface roughness, adhesion, friction, wear, and lubrication at the interface between two solids with and without liquid films have been studied using the AFM and FFM. The status of current understanding of micro/nanotribology of engineering interfaces follows.

Experimental Techniques

An AFM relies on a scanning technique to produce very high resolution, three-dimensional images of sample surfaces. The AFM measures ultrasmall forces (less than 1 nN) present between the AFM tip surface and a sample surface. These small forces are measured by measuring the motion of a very flexible cantilever beam having an ultrasmall mass. The deflection can be measured to with ±0.02 nm, so for a typical cantilever force constant of 10 N/m, a force as low as 0.2 nN can be detected. An AFM is capable of investigating surfaces of both conductors and insulators on an atomic scale. In the operation of a high-resolution AFM, the sample is generally scanned; however, AFMs are also available where the tip is scanned and the sample is stationary. To obtain atomic resolution with an AFM, the spring constant of the cantilever should be weaker than the equivalent spring between atoms. A cantilever beam with a spring constant of about 1 N/m or lower is desirable. Tips have to be sharp as possible. Tips with a radius ranging from 10 to 100 nm are commonly available.

In the AFM/FFM shown in Figure 3.14.2, the sample is mounted on a PZT tube scanner which consists of separate electrodes to precisely scan the sample in the *X*–*Y* plane in a raster pattern and to move the sample in the vertical (*Z*) direction. A sharp tip at the end of a flexible cantilever is brought in contact with the sample. Normal and frictional forces being applied at the tip–sample interface are measured simultaneously, using a laser beam deflection technique.

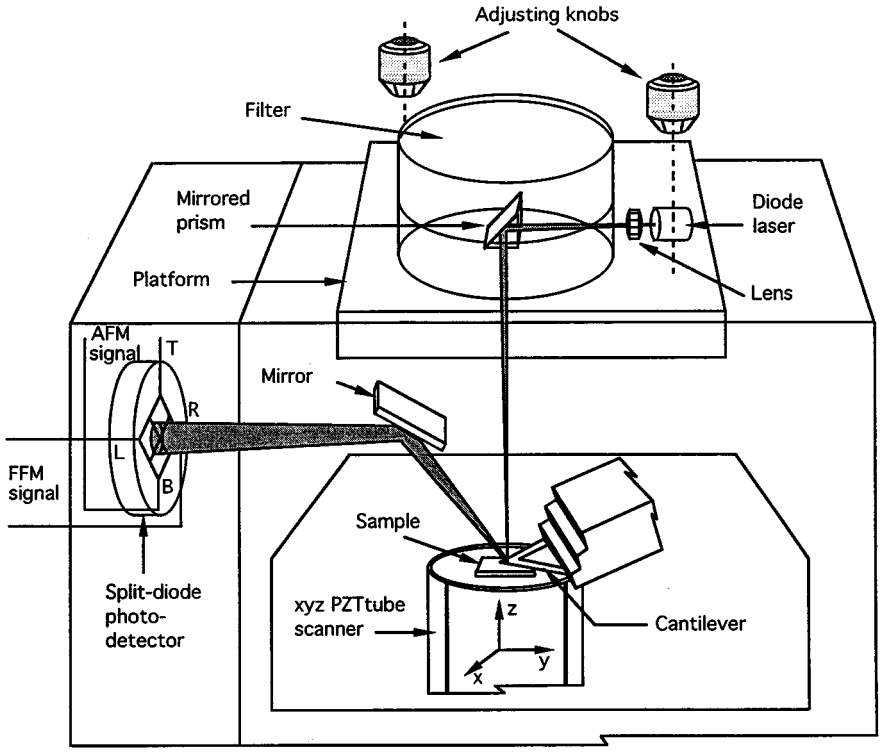


FIGURE 3.14.2 Schematic of a commercial AFM/FFM using laser beam deflection method.

Topographic measurements are typically made using a sharp tip on a cantilever beam with normal stiffness on the order of 0.5 N/m at a normal load of about 10 nN, and friction measurements are carried out in the load range of 10 to 150 nN. The tip is scanned in such a way that its trajectory on the sample forms a triangular pattern. Scanning speeds in the fast and slow scan directions depend on the scan area and scan frequency. A maximum scan size of 125 × 125 μm and scan rate of 122 Hz typically can be used. Higher scan rates are used for small scan lengths.

For nanoscale boundary lubrication studies, the samples are typically scanned over an area of 1 × 1 μm at a normal force of about 300 nN, in a direction orthogonal to the long axis of the cantilever beam (Bhushan, 1997). The samples are generally scanned with a scan rate of 1 Hz and the scanning speed of 2 μm/sec. The coefficient of friction is monitored during scanning for a desired number of cycles. After the scanning test, a larger area of 2 × 2 μm is scanned at a normal force of 40 nN to observe for any wear scar.

For microscale scratching, microscale wear, and nano-scale indentation hardness measurements, a sharp single-crystal natural diamond tip mounted on a stainless steel cantilever beam with a normal stiffness on the order of 25 N/m is used at relatively higher loads (1 to 150 μN). For wear studies, typically an area of 2 × 2 μm is scanned at various normal loads (ranging from 1 to 100 μN) for a selected number of cycles. For nanoindentation hardness measurements the scan size is set to zero and then normal load is applied to make the indents. During this procedure, the diamond tip is continuously pressed against the sample surface for about 2 sec at various indentation loads. The sample surface is scanned before and after the scratching, wear, or indentation to obtain the initial and the final surface topography, at a low normal load of about 0.3 μN using the same diamond tip. An area larger than the indentation region is scanned to observe the indentation marks. Nanohardness is calculated by dividing the indentation load by the projected residual area of the indents.

In measurements using conventional AFMs, the hardness value is based on the projected residual area after imaging the incident. Identification of the boundary of the indentation mark is difficult to accomplish with great accuracy, which makes the direct measurement of contact area somewhat inaccurate. A capacitive transducer with the dual capability of depth sensing as well as *in situ* imaging is used in conjunction with an AFM (Bhushan et al., 1996). This indentation system, called nano/picoindention, is used to make load-displacement measurements and subsequently carry out *in situ* imaging of the indent, if necessary. Indenter displacement at a given load is used to calculate the projected indent area for calculation of the hardness value. Young's modulus of elasticity is obtained from the slope of the unloading portion of the load-displacement curve.

Surface Roughness, Adhesion, and Friction

Solid surfaces, irrespective of the method of formation, contain surface irregularities or deviations from the prescribed geometrical form. When two nominally flat surfaces are placed in contact, surface roughness causes contact to occur at discrete contact points. Deformation occurs in these points and may be either elastic or plastic, depending on the nominal stress, surface roughness, and material properties. The sum of the areas of all the contact points constitutes the real area that would be in contact, and for most materials at normal loads this will be only a small fraction of the area of contact if the surfaces were perfectly smooth. In general, real area of contact must be minimized to minimize adhesion, friction, and wear (Bhushan and Gupta, 1997; Bhushan, 1996). Characterizing surface roughness is therefore important for predicting and understanding the tribological properties of solids in contact.

Surface roughness most commonly refers to the variations in the height of the surface relative to a reference plane (Bowden and Tabor, 1950; Bhushan, 1996). Commonly measured roughness parameters, such as standard deviation of surface heights (rms), are found to be scale dependent and a function of the measuring instrument, for any given surface, Figure 3.14.3 (Poon and Bhushan, 1995). The topography of most engineering surfaces is fractal, possessing a self-similar structure over a range of scales. By using fractal analysis one can characterize the roughness of surfaces with two scale-independent fractal parameters D and C which provide information about roughness at all length scales (Ganti and

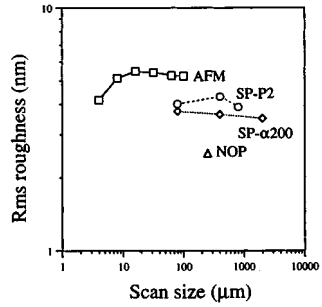


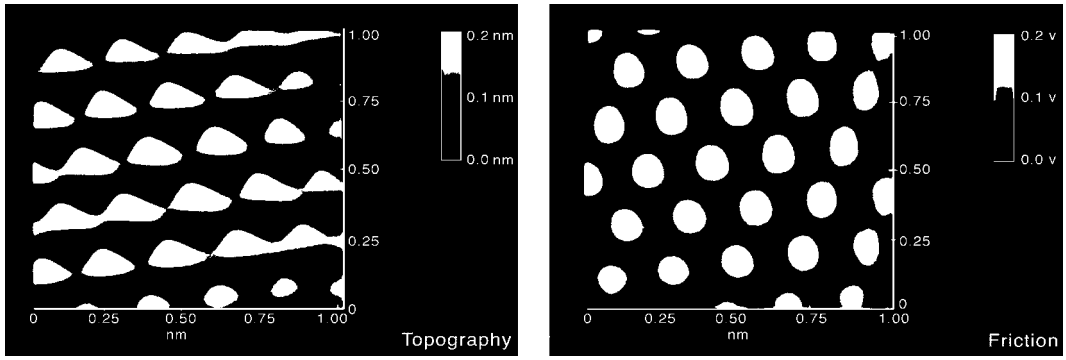
FIGURE 3.14.3 Scale dependence of standard deviation of surface heights (rms) for a glass–ceramic substrate, measured using an AFM, a stylus profiler (SP-P2 and SP-α 200), and a noncontact optical profiler (NOP).

Bhushan, 1995; Bhushan, 1995). These two parameters are instrument independent and are unique for each surface. *D* (generally ranging from 1 to 2) primarily relates to distribution of different frequencies in the surface profile, and *C* to the amplitude of the surface height variations at all frequencies. A fractal model of elastic plastic contact has been used to predict whether contacts experience elastic or plastic deformation and to predict the statistical distribution of contact points.

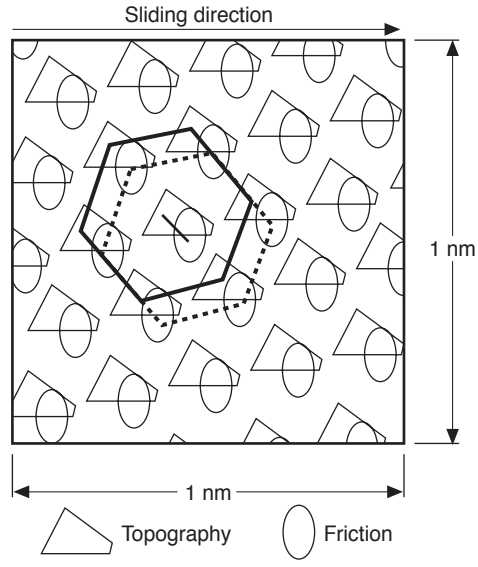
Based on atomic-scale friction measurements of a well-characterized freshly cleaved surface of highly oriented pyrolytic graphite (HOPG), the atomic-scale friction force of HOPG exhibits the same periodicity as that of corresponding topography (Figure 3.14.4(a)), but the peaks in friction and those in topography were displaced relative to each other (Figure 3.14.4(b)). A Fourier expansion of the interatomic potential has been used to calculate the conservative interatomic forces between atoms of the FFM tip and those of the graphite surface. Maxima in the interatomic forces in the normal and lateral directions do not occur at the same location, which explains the observed shift between the peaks in the lateral force and those in the corresponding topography. Furthermore, the observed local variations in friction force were explained by variation in the intrinsic lateral force between the sample and the FFM tip, and these variations may not necessarily occur as a result of an atomic-scale stick–slip process.

Friction forces of HOPG have also been studied. Local variations in the microscale friction of cleaved graphite are observed, which arise from structural changes that occur during the cleaving process. The cleaved HOPG surface is largely atomically smooth, but exhibits line-shaped regions in which the coefficient of friction is more than an order of magnitude larger. Transmission electron microscopy indicates that the line-shaped regions consist of graphite planes of different orientation, as well as of amorphous carbon. Differences in friction can also be seen for organic mono- and multilayer films, which again seem to be the result of structural variations in the films. These measurements suggest that the FFM can be used for structural mapping of the surfaces. FFM measurements can be used to map chemical variations, as indicated by the use of the FFM with a modified probe tip to map the spatial arrangement of chemical functional groups in mixed organic monolayer films. Here, sample regions that had stronger interactions with the functionalized probe tip exhibited larger friction. For further details, see Bhushan (1995).

Local variations in the microscale friction of scratched surfaces can be significant and are seen to depend on the local surface slope rather than on the surface height distribution (Bhushan, 1995). Directionality in friction is sometimes observed on the macroscale; on the microscale this is the norm (Bhushan, 1995). This is because most “engineering” surfaces have asymmetric surface asperities so that the interaction of the FFM tip with the surface is dependent on the direction of the tip motion. Moreover, during surface-finishing processes material can be transferred preferentially onto one side of the asperities, which also causes asymmetry and directional dependence. Reduction in local variations and in the directionality of frictional properties therefore requires careful optimization of surface roughness distributions and of surface-finishing processes.



a



b

FIGURE 3.14.4 (a) Gray-scale plots of surface topography (left) and friction profiles (right) of a 1 × 1 nm area of freshly cleaved HOPG, showing the atomic-scale variation of topography and friction, (b) diagram of superimposed topography and friction profiles from (a); the symbols correspond to maxima. Note the spatial shift between the two profiles.

Table 3.14.1 shows the coefficient of friction measured for two surfaces on micro- and macroscales. The coefficient of friction is defined as the ratio of friction force to the normal load. The values on the microscale are much lower than those on the macroscale. When measured for the small contact areas and very low loads used in microscale studies, indentation hardness and modulus of elasticity are higher than at the macroscale. This reduces the degree of wear. In addition, the small apparent areas of contact reduce the number of particles trapped at the interface, and thus minimize the "ploughing" contribution to the friction force.

At higher loads (with contact stresses exceeding the hardness of the softer material), however, the coefficient of friction for microscale measurements increases toward values comparable with those obtained from macroscale measurements, and surface damage also increases (Bhushan et al., 1995; Bhushan and Kulkarni, 1996). Thus, Amontons' law of friction, which states that the coefficient of

TABLE 3.14.1 Surface Roughness and Micro- and Macroscale Coefficients of Friction of Various Samples

Material	rms Roughness, nm	Microscale Coefficient of Friction vs. Si_3N_4 Tip[a]	Macroscale Coefficient of Friction vs. Alumina Ball[b]
Si(111)	0.11	0.03	0.18
C^+-implanted Si	0.33	0.02	0.18

[a] Tip radius of about 50 nm in the load range of 10 to 150 nN (2.5 to 6.1 GPa), a scanning speed of 5 m/sec and scan area of 1 × 1 μm.

[b] Ball radius of 3 mm at a normal load of 0.1 N (0.3 GPa) and average sliding speed of 0.8 mm/sec.

friction is independent of apparent contact area and normal load, does not hold for microscale measurements. These findings suggest that microcomponents sliding under lightly loaded conditions should experience very low friction and near zero wear.

Scratching, Wear, and Indentation

The AFM can be used to investigate how surface materials can be moved or removed on micro- to nanoscales, for example, in scratching and wear (Bhushan, 1995) (where these things are undesirable) and, in nanomachining/nanofabrication (where they are desirable). The AFM can also be used for measurements of mechanical properties on micro- to nanoscales. Figure 3.14.5 shows microscratches made on Si(111) at various loads after 10 cycles. As expected, the depth of scratch increases with load. Such microscratching measurements can be used to study failure mechanisms on the microscale and to evaluate the mechanical integrity (scratch resistance) of ultrathin films at low loads.

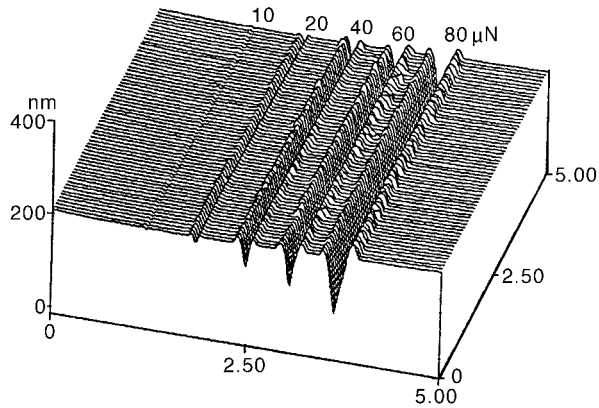


FIGURE 3.14.5 Surface profiles of Si(111) scratched at various loads. Note that the *x* and *y* axes are in micrometers and the *z* axis is in nanometers.

By scanning the sample in two dimensions with the AFM, wear scars are generated on the surface. The evolution of wear of a diamond-like carbon coating on a polished aluminum substrate is showing in Figure 3.14.6 which illustrates how the microwear profile for a load of 20 μN develops as a function of the number of scanning cycles. Wear is not uniform, but is initiated at the nanoscratches indicating that surface defects (with high surface energy) act as initiation sites. Thus, scratch-free surfaces will be relatively resistant to wear.

Mechanical properties, such as load-displacement curves, hardness, and modulus of elasticity can be determined on micro- to picoscales using an AFM and its modifications (Bhushan, 1995; Bhushan et al., 1995, 1996). Indentability on the scale of picometers can be studied by monitoring the slope of cantilever deflection as a function of sample traveling distance after the tip is engaged and the sample

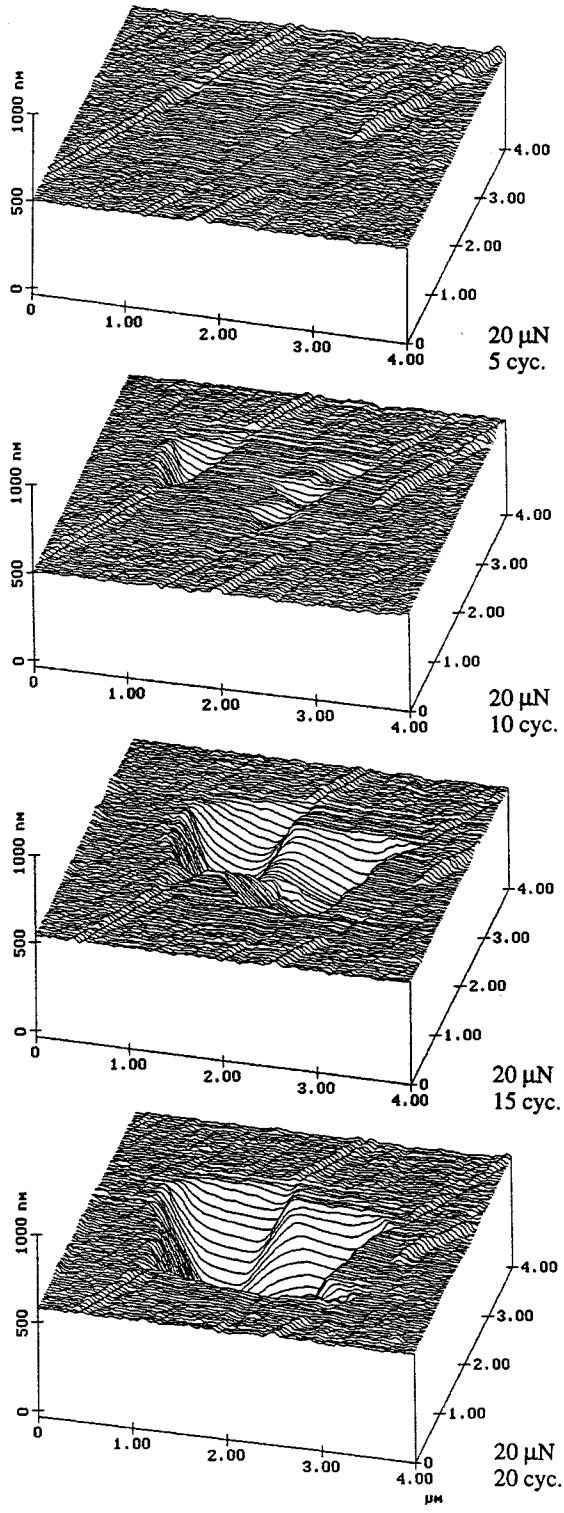


FIGURE 3.14.6 Surface profiles of diamond-like carbon-coated thin-film disk showing the worn region; the normal load and number of test cycles are indicated. (From Bhushan, B., *Handbook of Micro/Nanotribology*, CRC Press, Boca Raton, FL, 1995. With permission.)

is pushed against the tip. For a rigid sample, cantilever deflection equals the sample traveling distance; but the former quantity is smaller if the tip indents the sample. The indentation hardness on nanoscale of bulk materials and surface films with an indentation depth as small as 1 nm can be measured. An example of hardness data as a function of indentation depth is shown in Figure 3.14.7. A decrease in hardness with an increase in indentation depth can be rationalized on the basis that, as the volume of deformed materials increases, there is a higher probability of encountering material defects. AFM measurements on ion-implanted silicon surfaces show that ion implantation increases their hardness and, thus, their wear resistance (Bhushan, 1995). Formation of surface alloy films with improved mechanical properties by ion implantation is growing in technological importance as a means of improving the mechanical properties of materials.

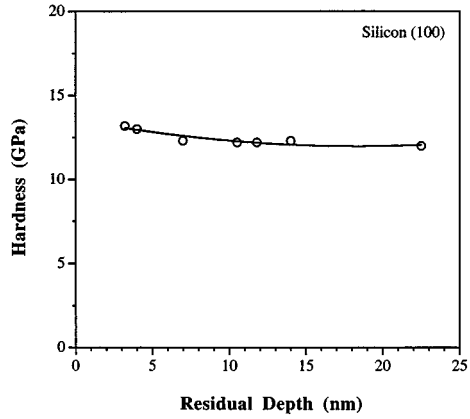


FIGURE 3.14.7 Indentation hardness as a function of residual indentation depth for Si(100).

Young's modulus of elasticity is calculated from the slope of the indentation curve during unloading (Bhushan, 1995; Bhushan et al., 1996). AFM can be used in a *force modulation mode* to measure surface elasticities: an AFM tip is scanned over the modulated sample surface with the feedback loop keeping the average force constant. For the same applied force, a soft area deforms more, and thus causes less cantilever deflection, than a hard area. The ratio of modulation amplitude to the local tip deflection is then used to create a *force modulation image*. The force modulation mode makes it easier to identify soft areas on hard substrates.

Detection of the transfer of material on a nanoscale is possible with the AFM. Indentation of C_{60}-rich fullerene films with an AFM tip has been shown to result in the transfer of fullerene molecules to the AFM tip, as indicated by discontinuities in the cantilever deflection as a function of sample traveling distance in subsequent indentation studies (Bhushan, 1995).

Boundary Lubrication

The "classical" approach to lubrication uses freely supported multimolecular layers of liquid lubricants (Bowden and Tabor, 1950, 1964; Bhushan, 1996). The liquid lubricants are chemically bonded to improve their wear resistance (Bhushan, 1995, 1996). To study depletion of boundary layers, the microscale friction measurements are made as a function of the number of cycles. For an example of the data of virgin Si(100) surface and silicon surface lubricated with about 2-nm-thick Z-15 and Z-Dol perfluoropolyether (PEPE) lubricants, see Figure 3.14.8. Z-Dol is PFPE lubricant with hydroxyl end groups. Its lubricant film was thermally bonded. In Figure 3.14.8, the unlubricated silicon sample shows a slight increase in friction force followed by a drop to a lower steady state value after some cycles. Depletion of native oxide and possible roughening of the silicon sample are responsible for the decrease in this

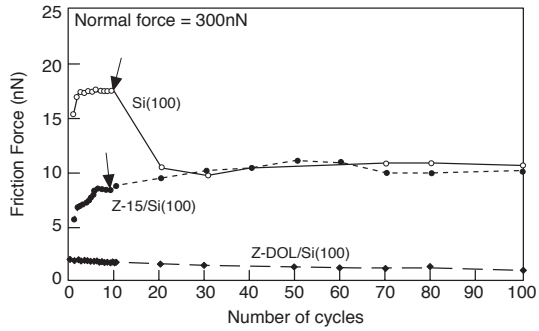


FIGURE 3.14.8 Friction force as a function of number of cycles using a silicon nitride tip at a normal force of 300 nN for the unlubricated and lubricated silicon samples.

friction force. The initial friction force for the Z-15-lubricated sample is lower than that of the unlubricated silicon and increases gradually to a friction force value comparable with that of the silicon after some cycles. This suggests the depletion of the Z-15 lubricant in the wear track. In the case of the Z-Dol-coated silicon sample, the friction force starts out to be low and remains low during the entire test. It suggests that Z-Dol does not get displaced/depleted as readily as Z-15. Additional studies of freely supported liquid lubricants showed that either increasing the film thickness or chemically bonding the molecules to the substrate with a mobile fraction improves the lubrication performance (Bhushan, 1997).

For lubrication of microdevices, a more effect approach involves the deposition of organized, dense molecular layers of long-chain molecules on the surface contact. Such monolayers and thin films are commonly produced by Langmuir–Blodgett (LB) deposition and by chemical grafting of molecules into self-assembled monolayers (SAMs). Based on the measurements, SAMs of octodecyl (C_{18}) compounds based on aminosilanes on a oxidized silicon exhibited a lower coefficient of friction of (0.018) and greater durability than LB films of zinc arachidate adsorbed on a gold surface coated with octadecylthiol (ODT) (coefficient of friction of 0.03) (Bhushan et al., 1995). LB films are bonded to the substrate by weak van der Waals attraction, whereas SAMs are chemically bound via covalent bonds. Because of the choice of chain length and terminal linking group that SAMs offer, they hold great promise for boundary lubrication of microdevices.

Measurement of ultrathin lubricant films with nanometer lateral resolution can be made with the AFM (Bhushan, 1995). The lubricant thickness is obtained by measuring the force on the tip as it approaches, contacts, and pushes through the liquid film and ultimately contacts the substrate. The distance between the sharp "snap-in" (owing to the formation of a liquid of meniscus between the film and the tip) at the liquid surface and the hard repulsion at the substrate surface is a measure of the liquid film thickness. This technique is now used routinely in the information-storage industry for thickness measurements (with nanoscale spatial resolution) of lubricant films, a few nanometers thick, in rigid magnetic disks.

References

Bhushan, B. 1995. *Handbook of Micro/Nanotribology*, CRC Press, Boca Raton, FL.

Bhushan, B. 1996. *Tribology and Mechanics of Magnetic Storage Devices*, 2nd ed., Springer, New York.

Bhushan, B. 1997. *Micro/Nanotribiology and Its Applications*, NATO ASI Series E: Applied Sciences, Kluwer, Dordrecht, Netherlands.

Bhushan, B. and Gupta, B.K. 1997. *Handbook of Tribology: Materials, Coatings and Surface Treatments*, McGraw-Hill, New York (1991); Reprint with corrections, Kreiger, Malabar, FL.

Bhushan, B. and Kulkarni, A.V. 1996. Effect of normal load on microscale friction measurements, *Thin Solid Films*, 278, 49–56.

Bhushan, B., Israelachvili, J.N., and Landman, U. 1995. Nanotribology: friction, wear and lubrication at the atomic scale, *Nature*, 374, 607–616.

Bhushan, B., Kulkarni, A.V., Bonin, W., and Wyrobek, J.T. 1996. Nano-indentation and pico-indentation measurements using capacitive transducer system in atomic force microscopy, *Philos. Mag.*, A74, 1117–1128.

Binning, G., Quate, C.F., and Gerber, Ch. 1986. Atomic force microscopy, *Phys. Rev. Lett.*, 56, 930–933.

Bowden, F.P. and Tabor, D. 1950; 1964. *The Friction and Lubrication of Solids*, Parts I and II, Clarendon, Oxford.

Ganti, S. and Bhushan, B. 1995. Generalized fractal analysis and its applications to engineering surfaces, *Wear*, 180, 17–34.

Poon, C.Y. and Bhushan, B. 1995. Comparison of surface roughness measurements by stylus profiler, AFM and non-contact optical profiler, *Wear*, 190, 76–88.

Tabor, D. and Winterton, R.H.S. 1969. The direct measurement of normal and retarded van der Walls forces, *Proc. R. Soc. London*, A312, 435–450.

Nomenclature for Fluid Mechanics

Symbol	Quantity	Unit		Dimensions (*MLtT*)
		SI	English	
a	Velocity of sound	m/sec	ft/sec	Lt^{-1}
a	Acceleration	m/sec^2	ft/sec^2	Lt^{-2}
A	Area	m^2	ft^2	L^2
b	Distance, width	m	ft	L
c_p	Specific heat, constant pressure	J/kg·K	ft·lb/lb$_m$·°R	$L^2t^{-2}T^{-1}$
c_v	Specific heat, constant volume	J/kg·K	ft·lb/lb$_m$·°R	$L^2t^{-2}T^{-1}$
C	Concentration	No./m^3	No./ft^3	L^{-3}
C	Coefficient	—	—	—
C	Empirical constant	—	—	—
D	Diameter	m	ft	L
D_H	Hydraulic diameter	m	ft	L
e	Total energy per unit mass	J/kg	ft·lb/lb$_m$	L^2t^{-2}
E	Total energy	J	ft·lb or Btu	ML^2t^{-2}
E	Modulus of leasticity	Pa	lb/ft^2	ML^1t^{-2}
Eu	Euler number	—	—	—
f	Friction factor	—	—	—
F	Force	N	lb	MLt^{-2}
Fr	Froude number	—	—	—
F_B	Buoyant force	N	lb	MLt^{-2}
g	Acceleration of gravity	m/sec^2	ft/sec^2	Lt^{-2}
g_0	Gravitation constant	kg·m/N·sec^2	lb$_m$·ft/lb·sec^2	—
G	Mass flow rate per unit area	kg/sec·m^2	lb$_m$/sec·ft^2	$ML^{-2}t^{-1}$
h	Head, vertical distance	m	ft	L
h	Enthalpy per unit mass	J/kg	ft·lb/lb$_m$	L^2t^{-2}
H	Head, elevation of hydraulic grade line	m	ft	L
I	Moment of inertia	m^4	ft^4	L^4
k	Specific heat ratio	—	—	—
K	Bulk modulus of elasticity	Pa	lb/ft^2	ML^1t^{-2}
K	Minor loss coefficient	—	—	—
L	Length	m	ft	L
L	Lift	N	lb	MLt^{-2}
l	Length, mixing length	m	ft	L
ln	Natural logarithm	—	—	—
m	Mass	kg	lb$_m$	M
m	Strength of source	m^3/sec	ft^3/sec	L^3t^{-1}
m	Mass flow rate	kg/sec	lb$_m$/sec	Mt^{-1}

Symbol	Quantity	Unit		Dimensions (*MLtT*)
		SI	**English**	
M	Molecular weight	—	—	—
M	Momentum per unit time	N	lb	MLt^{-2}
M	Mach number	—	—	—
n	Exponent, constant	—	—	—
n	Normal direction	m	ft	L
n	Manning roughness factor	—	—	—
n	Number of moles	—	—	—
N	Rotation speed	1/sec	1/sec	t^{-1}
NPSH	Net positive suction head	m	ft	L
p	Pressure	Pa	lb/ft^2	$ML^{-1}t^{-2}$
P	Height of weir	m	ft	L
P	Wetted perimeter	m	ft	L
q	Discharge per unit width	m^2/sec	ft^2/sec	L^2t^{-1}
q	Heat transfer per unit time	J/sec	Btu	ML^2t^{-3}
r	Radial distance	m	ft	L
R	Gas constant	J/kg·K	ft·lb/lb$_m$·°R	$L^2t^{-2}T^{-1}$
Re	Reynolds number	—	—	—
s	Distance	m	ft	L
s	Entropy per unit mass	J/kg·K	ft·lb/lb$_m$·°R	$L^2t^{-2}T^{-1}$
S	Entropy	J/K	ft·lb/°R	$ML^2t^{-2}T^{-1}$
S	Specific gravity, slope	—	—	—
t	Time	sec	sec	t
t	Distance, thickness	m	ft	L
T	Temperature	K	°R	T
T	Torque	N·m	lb·ft	ML^2t^{-2}
u	Velocity, Velocity component	m/sec	ft/sec	Lt^{-1}
u	Peripheral speed	m/sec	ft/sec	Lt^{-1}
u	Internal energy per unit mass	J/kg	ft·lb/lb$_m$	L^2t^{-2}
u_*	Shear stress velocity	m/sec	ft/sec	Lt^{-1}
U	Internal energy	J	Btu	ML^2t^{-2}
v	Velocity, velocity component	m/sec	ft/sec	Lt^{-1}
v_s	Specific volume	m^3/kg	ft^3/lb$_m$	$M^{-1}L^3$
V	Volume	m^3	ft^3	L^3
V	Volumetric flow rate	m^3/sec	ft^3/sec	L^3t^{-1}
V	Velocity	m/sec	ft/sec	Lt^{-1}
w	Velocity component	m/sec	ft/sec	Lt^{-1}
w	Work per unit mass	J/kg	ft·lb/lb$_m$	L^2t^{-2}
W	Work per unit time	J/sec	ft·lb/sec	ML^2t^{-3}
W_s	Shaft work	m·N	ft·lb	ML^2t^{-2}
W	Weight	N	lb	MLt^{-2}
We	Weber number	—	—	—
x	Distance	m	ft	L
y	Distance, depth	m	ft	L
Y	Weir height	m	ft	L
z	Vertical distance	m	ft	L

Greek Symbols

α	Angle, coefficient	—	—	—
β	Blade angle	—	—	—
Γ	Circulation	m^2	ft^2	L^2t^{-1}
ύ	Vector operator	1/m	1/ft	L^{-1}
γ	Specific weight	N/m^3	lb/ft^3	$ML^{-2}t^{-2}$
δ	Boundary layer thickness	m	ft	L
ε	Kinematic eddy viscosity	m^2/sec	ft^2/sec	L^2t^{-1}
ε	Roughness height	m	ft	L
η	Eddy viscosity	$N{\cdot}sec/m^2$	$lb{\cdot}sec/ft^2$	$ML^{-1}t^{-1}$
η	Head ratio	—	—	—
η	Efficiency	—	—	—
Θ	Angle	—	—	—
κ	Universal constant	—	—	—
λ	Scale ratio, undetermined multiplier	—	—	—
μ	Viscosity	$N{\cdot}sec/m^2$	$lb{\cdot}sec/ft^2$	$ML^{-1}t^{-1}$
ν	Kinematic viscosity (= μ/ρ)	m^2/sec	ft^2/sec	L^2t^{-1}
Φ	Velocity potential	m^2/sec	ft^2/sec	L^2t^{-1}
Φ	Function	—	—	—
π	Constant	—	—	—
Π	Dimensionless constant	—	—	—
ρ	Density	kg/m^3	lb_m/ft^3	ML^{-3}
σ	Surface tension	N/m	lb/ft	Mt^{-2}
σ	Cavitation index	—	—	—
τ	Shear stress	Pa	lb/ft^2	ML^1t^2
ψ	Stream function, two dimensions	m/sec	ft/sec	L^2t^{-1}
ψ	Stokes' stream function	m^3/sec	ft^3/sec	L^3t^{-1}
ω	Angular velocity	rad/sec	rad/sec	t^{-1}

Subscripts

c	Critical condition
u	Unit quantities
c.s.	Control surface
c.v.	Control volume
o	Stagnation or standard state condition
1, 2	Inlet and outlet, respectively, of control volume or machine rotor
∞	Upstream condition far away from body, free stream
T	Total pressure
J	Static pressure

Kreith, F.; Boehm, R.F.; et. al. "Heat and Mass Transfer"
Mechanical Engineering Handbook
Ed. Frank Kreith
Boca Raton: CRC Press LLC, 1999

Heat and Mass Transfer

Frank Kreith
University of Colorado
Robert F. Boehm
University of Nevada-Las Vegas
George D. Raithby
University of Waterloo
K. G. T. Hollands
University of Waterloo
N. V. Suryanarayana
Michigan Technological University
Michael F. Modest
Pennsylvania State University
Van P. Carey
University of California at Berkeley
John C. Chen
Lehigh University
Noam Lior
University of Pennsylvania
Ram K. Shah
Delphi Harrison Thermal Systems
Kenneth J. Bell
Oklahoma State University
Robert J. Moffat
Stanford University
Anthony F. Mills
University of California at Los Angeles
Arthur E. Bergles
Rensselaer Polytechnic Institute
Larry W. Swanson
Heat Transfer Research Institute
Vincent W. Antonetti
Poughkeepsie, New York
Thomas F. Irvine, Jr.
State University of New York, Stony Brook
Massimo Capobianchi
State University of New York, Stony Brook

4.1 Conduction Heat Transfer

Robert F. Boehm

Introduction

Conduction heat transfer phenomena are found throughout virtually all of the physical world and the industrial domain. The analytical description of this heat transfer mode is one of the best understood. Some of the bases of understanding of conduction date back to early history. It was recognized that by invoking certain relatively minor simplifications, mathematical solutions resulted directly. Some of these were very easily formulated. What transpired over the years was a very vigorous development of applications to a broad range of processes. Perhaps no single work better summarizes the wealth of these studies than does the book by Carslaw and Jaeger (1959). They gave solutions to a broad range of problems, from topics related to the cooling of the Earth to the current-carrying capacities of wires. The general analyses given there have been applied to a range of modern-day problems, from laser heating to temperature-control systems.

Today conduction heat transfer is still an active area of research and application. A great deal of interest has developed in recent years in topics like contact resistance, where a temperature difference develops between two solids that do not have perfect contact with each other. Additional issues of current interest include non-Fourier conduction, where the processes occur so fast that the equation described below does not apply. Also, the problems related to transport in miniaturized systems are garnering a great deal of interest. Increased interest has also been directed to ways of handling composite materials, where the ability to conduct heat is very directional.

Much of the work in conduction analysis is now accomplished by use of sophisticated computer codes. These tools have given the heat transfer analyst the capability of solving problems in nonhomogenous media, with very complicated geometries, and with very involved boundary conditions. It is still important to understand analytical ways of determining the performance of conducting systems. At the minimum these can be used as calibrations for numerical codes.

Fourier's Law

The basis of conduction heat transfer is **Fourier's Law**. This law involves the idea that the heat flux is proportional to the temperature gradient in any direction n. **Thermal conductivity**, k, a property of materials that is temperature dependent, is the constant of proportionality.

$$q_k = -kA\frac{\partial T}{\partial n} \tag{4.1.1}$$

In many systems the area A is a function of the distance in the direction n. One important extension is that this can be combined with the first law of thermodynamics to yield the **heat conduction equation**. For constant thermal conductivity, this is given as

$$\nabla^2 T + \frac{\dot{q}_G}{k} = \frac{1}{\alpha}\frac{\partial T}{\partial t} \tag{4.1.2}$$

In this equation, α is the thermal diffusivity and $\dot{q}_G$ is the internal heat generation per unit volume. Some problems, typically steady-state, one-dimensional formulations where only the heat flux is desired, can be solved simply from Equation (4.1.1). Most conduction analyses are performed with Equation (4.1.2). In the latter, a more general approach, the temperature distribution is found from this equation and appropriate boundary conditions. Then the heat flux, if desired, is found at any location using Equation (4.1.1). Normally, it is the temperature distribution that is of most importance. For example,

it may be desirable to know through analysis if a material will reach some critical temperature, like its melting point. Less frequently the heat flux is desired.

While there are times when it is simply desired to understand what the temperature response of a structure is, the engineer is often faced with a need to increase or decrease heat transfer to some specific level. Examination of the thermal conductivity of materials gives some insight into the range of possibilities that exist through simple conduction.

Of the more common engineering materials, pure copper exhibits one of the higher abilities to conduct heat with a thermal conductivity approaching 400 W/m^2 K. Aluminum, also considered to be a good conductor, has a thermal conductivity a little over half that of copper. To increase the heat transfer above values possible through simple conduction, more-involved designs are necessary that incorporate a variety of other heat transfer modes like convection and phase change.

Decreasing the heat transfer is accomplished with the use of insulations. A separate discussion of these follows.

Insulations

Insulations are used to decrease heat flow and to decrease surface temperatures. These materials are found in a variety of forms, typically *loose fill*, *batt*, and *rigid*. Even a gas, like air, can be a good insulator if it can be kept from moving when it is heated or cooled. A vacuum is an excellent insulator. Usually, though, the engineering approach to insulation is the addition of a low-conducting material to the surface. While there are many chemical forms, costs, and maximum operating temperatures of common forms of insulations, it seems that when a higher operating temperature is required, many times the thermal conductivity and cost of the insulation will also be higher.

Loose-fill insulations include such materials as milled alumina-silica (maximum operating temperature of 1260°C and thermal conductivities in the range of 0.1 to 0.2 W/m^2 K) and perlite (maximum operating temperature of 980°C and thermal conductivities in the range of 0.05 to 1.5 W/m^2 K). Batt-type insulations include one of the more common types — glass fiber. This type of insulation comes in a variety of densities, which, in turn, have a profound affect on the thermal conductivity. Thermal conductivities for glass fiber insulations can range from about 0.03 to 0.06 W/m^{2}K. Rigid insulations show a very wide range of forms and performance characteristics. For example, a rigid insulation in foam form, polyurethane, is very lightweight, shows a very low thermal conductivity (about 0.02 W/m^2 K), but has a maximum operating temperature only up to about 120°C. Rigid insulations in refractory form show quite different characteristics. For example, high-alumina brick is quite dense, has a thermal conductivity of about 2 W/m^2 K, but can remain operational to temperatures around 1760°C. Many insulations are characterized in the book edited by Guyer (1989).

Often, commercial insulation systems designed for high-temperature operation use a layered approach. Temperature tolerance may be critical. Perhaps a refractory is applied in the highest temperature region, an intermediate-temperature foam insulation is used in the middle section, and a high-performance, low-temperature insulation is used on the outer side near ambient conditions.

Analyses can be performed including the effects of temperature variations of thermal conductivity. However, the most frequent approach is to assume that the thermal conductivity is constant at some temperature between the two extremes experienced by the insulation.

The Plane Wall at Steady State

Consider steady-state heat transfer in a plane wall of thickness L, but of very large extent in both other directions. The wall has temperature T_1 on one side and T_2 on the other. If the thermal conductivity is considered to be constant, then Equation (4.1.1) can be integrated directly to give the following result:

$$q_k = \frac{kA}{L}\left(T_1 - T_2\right) \tag{4.1.3}$$

This can be used to determine the steady-state heat transfer through slabs.

An electrical circuit analog is widely used in conduction analyses. This is realized by considering the temperature difference to be analogous to a voltage difference, the heat flux to be like current flow, and the remainder of Equation (4.1.3) to be like a thermal resistance. The latter is seen to be

$$R_k = \frac{L}{kA} \tag{4.1.4}$$

Heat transfer through walls made of layers of different types of materials can be easily found by summing the resistances in series or parallel form, as appropriate.

In the design of systems, seldom is a surface temperature specified or known. More often, the surface is in contact with a bulk fluid, whose temperature is known at some distance from the surface. Convection from the surface is then represented by Newton's law of cooling:

$$q = \bar{h}_c A\left(T_s - T_\infty\right) \tag{4.1.5}$$

This equation can also be represented as a temperature difference divided by a thermal resistance, which is $1/\bar{h}_{cA}$. It can be shown that a very low surface resistance, as might be represented by phase change phenomena, has the effect of imposing the fluid temperature directly on the surface. Hence, usually a *known* surface temperature results from a fluid temperature being imposed directly on the surface through a very high heat transfer coefficient. For this reason, in the later results given here, particularly those for transient systems, a convective boundary will be assumed. For steady results this is less important because of the ability to add resistances through the circuit analogy.

Long, Cylindrical Systems at Steady State

For long (L) annular systems at steady-state conditions with constant thermal conductivities, the following two equations are the appropriate counterparts to Equations (4.1.3) and (4.1.4). The heat transfer can be expressed as

$$q_k = \frac{2\pi Lk}{\ln\left[r_2/r_1\right]}\left(T_1 - T_2\right) \tag{4.1.6}$$

Here, r_1 and r_2 represent the radii of annular section. A thermal resistance for this case is as shown below.

$$R_k = \frac{\ln\left[r_2/r_1\right]}{2\pi Lk} \tag{4.1.7}$$

The Overall Heat Transfer Coefficient

The **overall heat transfer coefficient** concept is valuable in several aspects of heat transfer. It involves a modified form of Newton's law of cooling, as noted above, and it is written as

$$Q = \bar{U}A\Delta T \tag{4.1.8}$$

In this formulation $\bar{U}$ is the overall heat transfer coefficient based upon the area A. Because the area for heat transfer in a problem can vary (as with a cylindrical geometry), it is important to note that the $\bar{U}$ is dependent upon which area is selected. The overall heat transfer coefficient is usually found from a combination of thermal resistances. Hence, for a common series-combination-circuit analog, the $\bar{U}A$ product is taken as the sum of resistances.

$$\bar{U}A = \frac{1}{\sum_{i=1}^{n} R_i} = \frac{1}{R_{\text{total}}} \tag{4.1.9}$$

To show an example of the use of this concept, consider Figure 4.1.1.

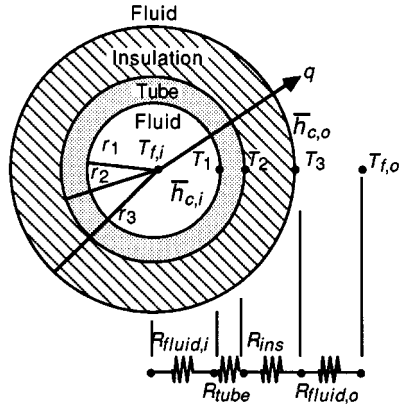


FIGURE 4.1.1. An insulated tube with convective environments on both sides.

For steady-state conditions, the product $\bar{U}A$ remains constant for a given heat transfer and overall temperature difference. This can be written as

$$\bar{U}_1 A_1 = \bar{U}_2 A_2 = \bar{U}_3 A_3 = \bar{U}A \tag{4.1.10}$$

If the inside area, A_1, is chosen as the basis, the overall heat transfer coefficient can then be expressed as

$$\bar{U}_1 = \frac{1}{\frac{1}{\bar{h}_{c,i}} + \frac{r_1 \ln(r_2/r_1)}{k_{\text{pipe}}} + \frac{r_1 \ln(r_3/r_2)}{k_{\text{ins}}} + \frac{r_1}{r_3 \bar{h}_{c,o}}} \tag{4.1.11}$$

Critical Thickness of Insulation

Sometimes insulation can cause an increase in heat transfer. This circumstance should be noted in order to apply it when desired and to design around it when an insulating effect is needed. Consider the circumstance shown in Figure 4.1.1. Assume that the temperature is known on the outside of the tube (inside of the insulation). This could be known if the inner heat transfer coefficient is very large and the thermal conductivity of the tube is large. In this case, the inner fluid temperature will be almost the same as the temperature of the inner surface of the insulation. Alternatively, this could be applied to a coating (say an electrical insulation) on the outside of a wire. By forming the expression for the heat transfer in terms of the variables shown in Equation (4.1.11), and examining the change of heat transfer with variations in r_3 (that is, the thickness of insulation) a maximum heat flow can be found. While simple results are given many texts (showing the critical radius as the ratio of the insulation thermal conductivity to the heat transfer coefficient on the outside), Sparrow (1970) has considered a heat transfer coefficient that varies as $\bar{h}_{c,o} \sim r_3^{-m} |T_3 - T_{f,o}|^n$. For this case, it is found that the heat transfer is maximized at

$$r_3 = r_{crit} = \left[(1-m)/(1+n)\right]\frac{k_{ins}}{\bar{h}_{c,o}} \tag{4.1.12}$$

By examining the order of magnitudes of m, n, k_{ins}, and $\bar{h}_{c,o}$ the critical radius is found to be often on the order of a *few millimeters*. Hence, additional insulation on small-diameter cylinders such as small-gauge electrical wires could actually increase the heat dissipation. On the other hand, the addition of insulation to large-diameter pipes and ducts will almost always decrease the heat transfer rate.

Internal Heat Generation

The analysis of temperature distributions and the resulting heat transfer in the presence of volume heat sources is required in some circumstances. These include phenomena such as nuclear fission processes, joule heating, and microwave deposition. Consider first a slab of material $2L$ thick but otherwise very large, with internal generation. The outside of the slab is kept at temperature T_1. To find the temperature distribution within the slab, the thermal conductivity is assumed to be constant. Equation (4.1.2) reduces to the following:

$$\frac{d^2T}{dx^2} + \frac{\dot{q}_G}{k} = 0 \tag{4.1.13}$$

Solving this equation by separating variables, integrating twice, and applying boundary conditions gives

$$T(x) - T_1 = \frac{\dot{q}_G L^2}{2k}\left[1 - \left(\frac{x}{L}\right)^2\right] \tag{4.1.14}$$

A similar type of analysis for a long, cylindrical element of radius r_1 gives

$$T(r) - T_1 = \frac{\dot{q}_G r_1^2}{4k}\left[1 - \left(\frac{r}{r_1}\right)^2\right] \tag{4.1.15}$$

Two additional cases will be given. Both involve the situation when the heat generation rate is dependent upon the local temperature in a linear way (defined by a slope β), according to the following relationship:

$$\dot{q}_G = \dot{q}_{G,o}\left[1 + \beta\left(T - T_o\right)\right] \tag{4.1.16}$$

For a plane wall of $2L$ thickness and a temperature of T_1 specified on each surface

$$\frac{T(x) - T_o + 1/\beta}{T_1 - T_o + 1/\beta} = \frac{\cos\mu x}{\cos\mu L} \tag{4.1.17}$$

For a similar situation in a long cylinder with a temperature of T_1 specified on the outside radius r_1

$$\frac{T(r) - T_o + 1/\beta}{T_1 - T_o + 1/\beta} = \frac{J_o(\mu r)}{J_o(\mu r_1)} \tag{4.1.18}$$

In Equation (4.1.18), the J_o is the typical notation for the Bessel function. Variations of this function are tabulated in Abramowitz and Stegun (1964). In both cases the following holds:

$$\mu \equiv \sqrt{\frac{\beta \dot{q}_{G,o}}{k}}$$

Fins

Fins are widely used to enhance the heat transfer (usually convective, but it could also be radiative) from a surface. This is particularly true when the surface is in contact with a gas. Fins are used on air-cooled engines, electronic cooling forms, as well as for a number of other applications. Since the heat transfer coefficient tends to be low in gas convection, area is added in the form of fins to the surface to decrease the convective thermal resistance.

The simplest fins to analyze, and which are usually found in practice, can be assumed to be one-dimensional and constant in cross section. In simple terms, to be one-dimensional, the fins have to be long compared with a transverse dimension. Three cases are normally considered for analysis, and these are shown in Figure 4.1.2. They are the insulated tip, the infinitely long fin, and the convecting tip fin.

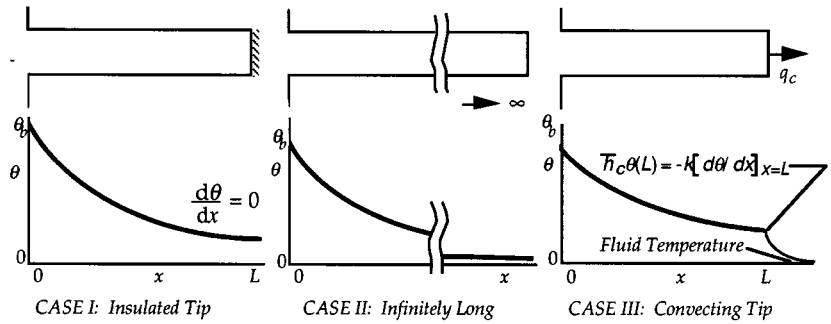


FIGURE 4.1.2. Three typical cases for one-dimensional, constant-cross-section fins are shown.

For Case, I, the solution to the governing equation and the application of the boundary conditions of the known temperature at the base and the insulated tip yields

$$\text{Case I:} \qquad \theta = \theta_b = \frac{\cosh m(L-x)}{\cosh mL} \tag{4.1.19}$$

For the infinitely long case, the following simple form results:

$$\text{Case II:} \qquad \theta(x) = \theta_b e^{-mx} \tag{4.1.20}$$

The final case yields the following result:

$$\text{Case III:} \qquad \theta(x) = \theta_b \frac{mL\cosh m(L-x) + \text{Bi}\sinh m(L-x)}{mL\cosh mL + \text{Bi}\sinh mL} \tag{4.1.21}$$

where

$$\text{Bi} \equiv \bar{h}_c L/k$$

In all three of the cases given, the following definitions apply:

$$\theta \equiv T(x) - T_\infty, \quad \theta_b \equiv T(x=0) - T_\infty, \quad \text{and} \quad m^2 \equiv \frac{\bar{h}_c P}{kA}$$

Here A is the cross section of the fin parallel to the wall; P is the perimeter around that area.

To find the heat removed in any of these cases, the temperature distribution is used in Fourier's law, Equation (4.1.1). For most fins that truly fit the one-dimensional assumption (i.e., long compared with their transverse dimensions), all three equations will yield results that do not differ widely.

Two performance indicators are found in the fin literature. The **fin efficiency** is defined as the ratio of the actual heat transfer to the heat transfer from an ideal fin.

$$\eta \equiv \frac{q_{\text{actual}}}{q_{\text{ideal}}} \tag{4.1.22}$$

The ideal heat transfer is found from convective gain or loss from an area the same size as the fin surface area, all at a temperature T_b. Fin efficiency is normally used to tabulate heat transfer results for various types of fins, including ones with nonconstant area or which do not meet the one-dimensional assumption. An example of the former can be developed from a result given by Arpaci (1966). Consider a straight fin of triangular profile, as shown in Figure 4.1.3. The solution is found in terms of modified Bessel functions of the first kind. Tabulations are given in Abramowitz and Stegun (1964).

$$\eta = \frac{I_1\left(2\tilde{m}L^{1/2}\right)}{\tilde{m}L^{1/2} I_o\left(2\tilde{m}L^{1/2}\right)} \tag{4.1.23}$$

Here, $\tilde{m} \equiv \sqrt{2\bar{h}_c L/kb}$.

The **fin effectiveness**, ε, is defined as the heat transfer from the fin compared with the bare-surface transfer through the same base area.

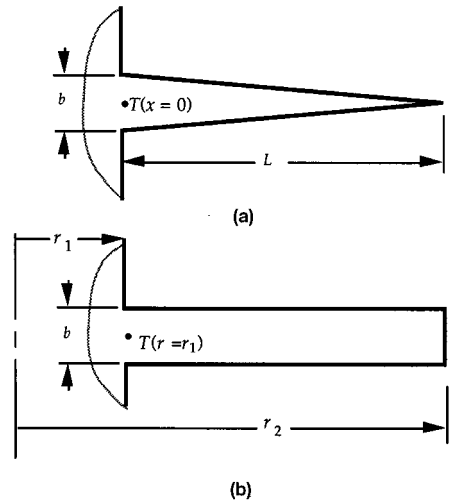


FIGURE 4.1.3. Two examples of fins with a cross-sectional area that varies with distance from the base. (a) Straight triangular fin. (b) Annular fin of constant thickness.

$$\varepsilon = \frac{q_{\text{actual}}}{q_{\text{bare base}}} = \frac{q_f}{\bar{h}_c A\left(T_b - T_\infty\right)} \tag{4.1.24}$$

Carslaw and Jaeger (1959) give an expression for the effectiveness of a fin of constant thickness around a tube (see Figure 4.1.3) This is given as $(\tilde{\mu} \equiv \sqrt{2\bar{h}_c / kb})$.

$$\varepsilon = \frac{2}{\tilde{\mu} b} \frac{I_1\left(\tilde{\mu} r_2\right) K_1\left(\tilde{\mu} r_1\right) - K_1\left(\tilde{\mu} r_2\right) I_1\left(\tilde{\mu} r_1\right)}{I_o\left(\tilde{\mu} r_1\right) K_1\left(\tilde{\mu} r_2\right) + K_o\left(\tilde{\mu} r_1\right) I_1\left(\tilde{\mu} r_2\right)} \tag{4.1.25}$$

Here the notations *I* and *K* denote Bessel functions that are given in Abramowitz and Stegun (1964).

Fin effectiveness can be used as one indication whether or not fins should be added. A rule of thumb indicates that if the effectiveness is less than about three, fins should not be added to the surface.

Transient Systems

Negligible Internal Resistance

Consider the transient cooling or heating of a body with surface area *A* and volume *V*. This is taking place by convection through a heat transfer coefficient $\bar{h}_c$ to an ambient temperature of T_∞. Assume the thermal resistance to conduction inside the body is significantly less than the thermal resistance to convection (as represented by Newton's law of cooling) on the surface of the body. This ratio is denoted by the **Biot number**, Bi.

$$\text{Bi} = \frac{R_k}{R_c} = \frac{\bar{h}_c (V/A)}{k} \tag{4.1.26}$$

The temperature (which will be uniform throughout the body at any time for this situation) response with time for this system is given by the following relationship. Note that the shape of the body is not important — only the ratio of its volume to its area matters.

$$\frac{T(t) - T_\infty}{T_o - T_\infty} = e^{-\bar{h}_c A t / \rho V c} \tag{4.1.27}$$

Typically, this will hold for the Biot number being less than (about) 0.1.

Bodies with Significant Internal Resistance

When a body is being heated or cooled transiently in a convective environment, but the internal thermal resistance of the body cannot be neglected, the analysis becomes more complicated. Only simple geometries (a symmetrical plane wall, a long cylinder, a composite of geometric intersections of these geometries, or a sphere) with an imposed step change in ambient temperature are addressed here.

The first geometry considered is a large slab of minor dimension 2*L*. If the temperature is initially uniform at T_o, and at time 0+ it begins convecting through a heat transfer coefficient to a fluid at T_∞, the temperature response is given by

$$\theta = 2 \sum_{n=1}^{\infty} \left(\frac{\sin \lambda_n L}{\lambda_n L + \sin \lambda_n L \cos \lambda_n L} \right) \exp\left(-\lambda_n^2 L^2 \text{Fo}\right) \cos\left(\lambda_n x\right) \tag{4.1.28}$$

and the λ_n are the roots of the transcendental equation: $\lambda_n L \tan \lambda_n L = \text{Bi}$. The following definitions hold:

$$\text{Bi} \equiv \frac{\bar{h}_c L}{k} \qquad \text{Fo} \equiv \frac{\alpha t}{L^2} \qquad \theta \equiv \frac{T - T_\infty}{T_o - T_\infty}$$

The second geometry considered is a very long cylinder of diameter $2R$. The temperature response for this situation is

$$\theta = 2\text{Bi} \sum_{n=1}^{\infty} \frac{\exp\left(-\lambda_n^2 R^2 \text{Fo}\right) J_o\left(\lambda_n r\right)}{\left(\lambda_n^2 R^2 + \text{Bi}^2\right) J_o\left(\lambda_n R\right)} \tag{4.1.29}$$

Now the λ_n are the roots of $\lambda_n R\, J_1(\lambda_n R) + \text{Bi}\, J_o\,(\lambda_n R) = 0$, and

$$\text{Bi} = \frac{\bar{h}_c R}{k} \qquad \text{Fo} = \frac{\alpha t}{R^2} \qquad \theta = \frac{T - T_\infty}{T_o - T_\infty}$$

The common definition of Bessel's functions applies here.

For the similar situation involving a solid sphere, the following holds:

$$\theta = 2 \sum_{n=1}^{\infty} \frac{\sin\left(\lambda_n R\right) - \lambda_n R \cos\left(\lambda_n R\right)}{\lambda_n R - \sin\left(\lambda_n R\right)\cos\left(\lambda_n R\right)} \exp\left(-\lambda_n^2 R^2 \text{Fo}\right) \frac{\sin\left(\lambda_n r\right)}{\lambda_n r} \tag{4.1.30}$$

and the λ_n are found as the roots of $\lambda_n R \cos \lambda_n R = (1 - \text{Bi}) \sin \lambda_n R$. Otherwise, the same definitions as were given for the cylinder hold.

Solids that can be envisioned as the geometric intersection of the simple shapes described above can be analyzed with a simple product of the individual-shape solutions. For these cases, the solution is found as the product of the dimensionless temperature functions for each of the simple shapes with appropriate distance variables taken in each solution. This is illustrated as the right-hand diagram in Figure 4.1.4. For example, a very long rod of rectangular cross section can be seen as the intersection of two large plates. A short cylinder represents the intersection of an infinitely long cylinder and a plate. The temperature at any location within the short cylinder is

$$\theta_{2R,2L\ \text{Rod}} = \theta_{\text{Infinite}\ 2R\ \text{Rod}}\, \theta_{2L\ \text{Plate}} \tag{4.1.31}$$

Details of the formulation and solution of the partial differential equations in heat conduction are found in the text by Arpaci (1966).

Finite-Difference Analysis of Conduction

Today, numerical solution of conduction problems is the most-used analysis approach. Two general techniques are applied for this: those based upon finite-difference ideas and those based upon finite-element concepts. General numerical formulations are introduced in other sections of this book. In this section, a special, physical formulation of the finite-difference equations to conduction phenomena is briefly outlined.

Attention is drawn to a one-dimensional slab (very large in two directions compared with the thickness). The slab is divided across the thickness into smaller subslabs, and this is shown in Figure 4.1.5. All subslabs are thickness Δx except for the two boundaries where the thickness is $\Delta x/2$. A characteristic temperature for each slab is assumed to be represented by the temperature at the slab center. Of course, this assumption becomes more accurate as the size of the slab becomes smaller. With

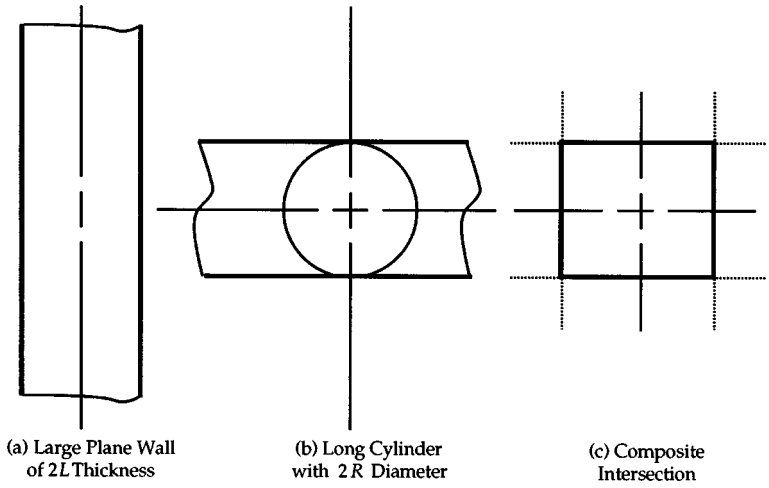


FIGURE 4.1.4. Three types of bodies that can be analyzed with results given in this section. (a) Large plane wall of 2*L* thickness; (b) long cylinder with 2*R* diameter; (c) composite intersection.

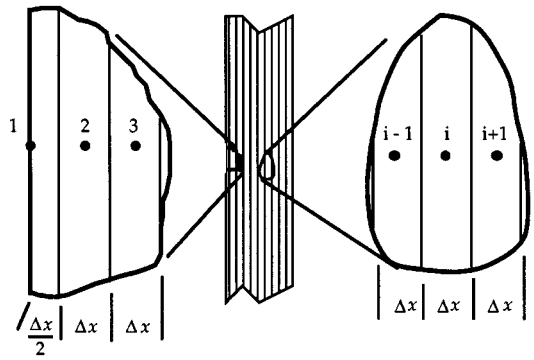


FIGURE 4.1.5. A one-dimensional finite differencing of a slab with a general interior node and one surface node detailed.

the two boundary-node centers located exactly on the boundary, a total of n nodes are used ($n - 2$ full nodes and one half node on each of the two boundaries).

In the analysis, a general interior node i (this applies to all nodes 2 through $n - 1$) is considered for an overall energy balance. Conduction from node $i - 1$ and from node $i + 1$ as well as any heat generation present is assumed to be energy per unit time flowing into the node. This is then equated to the time rate of change of energy within the node. A backward difference on the time derivative is applied here, and the notation $T_i' \equiv T_i(t + \Delta t)$ is used. The balance gives the following on a per-unit-area basis:

$$\frac{T'_{i-1} - T'_i}{\Delta x/k_-} + \frac{T'_{i+1} - T'_i}{\Delta x/k_+} + \dot{q}_{G,i}\Delta x = \rho \Delta x c_p \frac{T'_i - T_i}{\Delta t} \tag{4.1.32}$$

In this equation different thermal conductivities have been used to allow for possible variations in properties throughout the solid.

The analysis of the boundary nodes will depend upon the nature of the conditions there. For the purposes of illustration, convection will be assumed to be occurring off of the boundary at node 1. A balance similar to Equation (4.1.32) but now for node 1 gives the following:

$$\frac{T'_\infty - T'_1}{1/\bar{h}_c} + \frac{T'_2 - T'_1}{\Delta x/k_+} + \dot{q}_{G,1}\frac{\Delta x}{2} = \rho\frac{\Delta x}{2}c_p\frac{T'_1 - T_1}{\Delta t} \tag{4.1.33}$$

After all n equations are written, it can be seen that there are n unknowns represented in these equations: the temperature at all nodes. If one or both of the boundary conditions are in terms of a specified temperature, this will decrease the number of equations and unknowns by one or two, respectively. To determine the temperature as a function of time, the time step is arbitrarily set, and all the temperatures are found by simultaneous solution at $t = 0 + \Delta t$. The time is then advanced by Δt and the temperatures are then found again by simultaneous solution.

The finite difference approach just outlined using the backward difference for the time derivative is termed the *implicit* technique, and it results in an $n \times n$ system of linear simultaneous equations. If the forward difference is used for the time derivative, then only one unknown will exist in each equation. This gives rise to what is called an *explicit* or "marching" solution. While this type of system is more straightforward to solve because it deals with only one equation at a time with one unknown, a *stability criterion* must be considered which limits the time step relative to the distance step.

Two- and three-dimensional problems are handled in conceptually the same manner. One-dimensional heat fluxes between adjoining nodes are again considered. Now there are contributions from each of the dimensions represented. Details are outlined in the book by Jaluria and Torrance (1986).

Defining Terms

Biot number: Ratio of the internal (conductive) resistance to the external (convective) resistance from a solid exchanging heat with a fluid.

Fin: Additions of material to a surface to increase area and thus decrease the external thermal resistance from convecting and/or radiating solids.

Fin effectiveness: Ratio of the actual heat transfer from a fin to the heat transfer from the same cross-sectional area of the wall without the fin.

Fin efficiency: Ratio of the actual heat transfer from a fin to the heat transfer from a fin with the same geometry but completely at the base temperature.

Fourier's law: The fundamental law of heat conduction. Relates the local temperature gradient to the local heat flux, both in the same direction.

Heat conduction equation: A partial differential equation in temperature, spatial variables, time, and properties that, when solved with appropriate boundary and initial conditions, describes the variation of temperature in a conducting medium.

Overall heat transfer coefficient: The analogous quantity to the heat transfer coefficient found in convection (Newton's law of cooling) that represents the overall combination of several thermal resistances, both conductive and convective.

Thermal conductivity: The property of a material that relates a temperature gradient to a heat flux. Dependent upon temperature.

References

Abramowitz, M. and Stegun, I. 1964. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. National Bureau of Standards, Applied Mathematics Series 55.

Arpaci, V. 1966. *Conduction Heat Transfer*, Addison-Wesley, Reading, MA.

Carslaw, H.S. and Jaeger, J.C. 1959. *Conduction of Heat in Solids*, 2nd ed., Oxford University Press, London.

Guyer, E., Ed. 1989. Thermal insulations, in *Handbook of Applied Thermal Design*, McGraw-Hill, New York, Part 3.

Jaluria, Y. and Torrance, K. 1986. *Computational Heat Transfer*, Hemisphere Publishing, New York.

Sparrow, E. 1970. Reexamination and correction of the critical radius for radial heat conduction, *AIChE J.* 16, 1, 149.

Further Information

The references listed above will give the reader an excellent introduction to analytical formulation and solution (Arpaci), material properties (Guyer), and numerical formulation and solution (Jaluria and Torrance). Current developments in conduction heat transfer appear in several publications, including *The Journal of Heat Transfer, The International Journal of Heat and Mass Transfer*, and *Numerical Heat Transfer.*

4.2 Convection Heat Transfer

Natural Convection

George D. Raithby and K.G. Terry Hollands

Introduction

Natural convection heat transfer occurs when the convective fluid motion is induced by density differences that are themselves caused by the heating. An example is shown in Figure 4.2.1(A), where a body at surface temperature T_s transfers heat at a rate q to ambient fluid at temperature $T_\infty < T_s$.

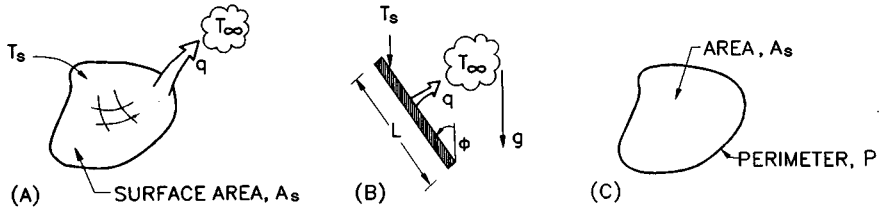


FIGURE 4.2.1 (A) Nomenclature for external heat transfer. (A) General sketch; (B) is for a tilted flat plate, and (C) defines the lengths cal for horizontal surfaces.

In this section, correlations for the average Nusselt number are provided from which the heat transfer rate q from surface area A_s can be estimated. The Nusselt number is defined as

$$\mathrm{Nu} = \frac{\bar{h}_c L}{k} = \frac{qL}{A_s \Delta T k} \tag{4.2.1}$$

where $\Delta T = T_s - T_\infty$ is the temperature difference driving the heat transfer. A dimensional analysis leads to the following functional relation:

$$\mathrm{Nu} = f(\mathrm{Ra},\ \mathrm{Pr},\ \text{geometric shape, boundary conditions}) \tag{4.2.2}$$

For given thermal boundary conditions (e.g., isothermal wall and uniform T_∞), and for a given geometry (e.g., a cube), Equation (4.2.2) states that Nu depends only on the Rayleigh number, Ra, and Prandtl number, Pr. The length scales that appear in Nu and Ra are defined, for each geometry considered, in a separate figure. The fluid properties are generally evaluated at T_f, the average of the wall and ambient temperatures. The exception is that β, the temperature coefficient of volume expansion, is evaluated at T_∞ for external natural convection(Figures 4.2.1 to 4.2.3) in a *gaseous* medium.

The functional dependence on Pr is approximately independent of the geometry, and the following Pr-dependent function will be useful for laminar heat transfer (Churchill and Usagi, 1972):

$$\bar{C}_\ell = 0.671 \Big/ \left(1 + (0.492/\mathrm{Pr})^{9/16}\right)^{4/9} \tag{4.2.3}$$

C_t^V and C_t^H are functions that will be useful for turbulent heat transfer:

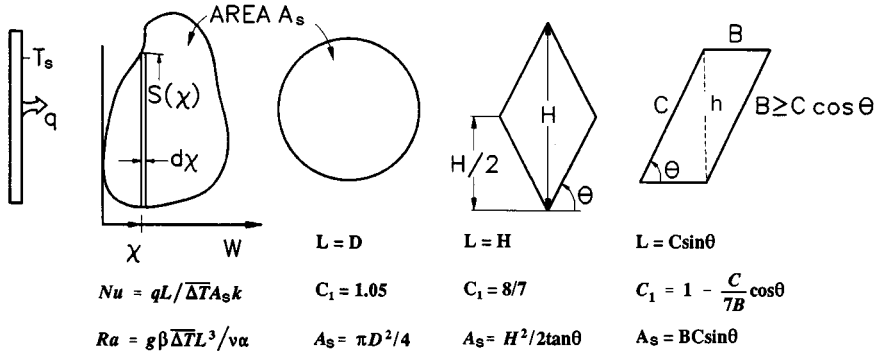


FIGURE 4.2.2 Nomenclature for heat transfer from planar surfaces of different shapes.

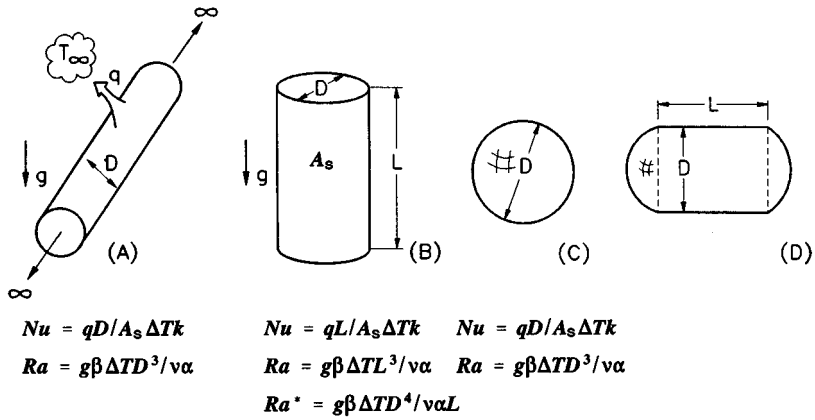


FIGURE 4.2.3 Definitions for computing heat transfer from a long circular cylinder (A), from the lateral surface of a vertical circular cylinder (B), from a sphere (C), and from a compound body (D).

$$C_t^V = 0.13\mathrm{Pr}^{0.22} \Big/ \left(1 + 0.61\mathrm{Pr}^{0.81}\right)^{0.42} \tag{4.2.4}$$

$$C_t^H = 0.14\left(\frac{1 + 0.0107\mathrm{Pr}}{1 + 0.01\mathrm{Pr}}\right) \tag{4.2.5}$$

The superscripts *V* and *H* refer to the vertical and horizontal surface orientation.

The Nusselt numbers for fully laminar and fully turbulent heat transfer are denoted by Nu_ℓ and Nu_t, respectively. Once obtained, these are blended (Churchill and Usagi, 1972) as follows to obtain the equation for Nu:

$$\mathrm{Nu} = \left(\left(\mathrm{Nu}_\ell\right)^m + \left(\mathrm{Nu}_t\right)^m\right)^{1/m} \tag{4.2.6}$$

The blending parameter *m* depends on the body shape and orientation.

The equation for Nu_ℓ in this section is usually expressed in terms of Nu^T, the Nusselt number that would be valid if the thermal boundary layer were thin. The difference between Nu_ℓ and Nu^T accounts for the effect of the large boundary layer thicknesses encountered in natural convection.

It is assumed that the wall temperature of a body exceeds the ambient fluid temperature ($T_s > T_\infty$). For $T_s < T_\infty$ the same correlations apply with $(T_\infty - T_s)$ replacing $(T_s - T_\infty)$ for a geometry that is rotated

180° relative to the gravitational vector; for example, the correlations for a horizontal heated upward-facing flat plate applies to a cooled downward-facing flat plate of the same planform.

Correlations for External Natural Convection

This section deals with problems where the body shapes in Figures 4.2.1 to 4.2.3 are heated while immersed in a quiescent fluid. Different *cases* are enumerated below.

1. *Isothermal Vertical (ϕ = 0) Flat Plate, Figure 4.2.1B.* For heat transfer from a vertical plate (Figure 4.2.1B), for $1 < \mathrm{Ra} < 10^{12}$,

$$\mathrm{Nu}^T = \overline{C}_\ell \mathrm{Ra}^{1/4} \qquad \mathrm{Nu}_\ell = \frac{2.0}{\ln\left(1 + 2.0/\mathrm{Nu}^T\right)} \tag{4.2.7}$$

$$\mathrm{Nu}_t = C_t^V \mathrm{Ra}^{1/3} \Big/ \left(1 + 1.4 \times 10^9\ \mathrm{Pr}/\mathrm{Ra}\right)$$

   $\overline{C}_\ell$ and C_t^V are given by Equations (4.2.3) and (4.2.4). Nu is obtained by substituting Equation (4.2.7) expressions for Nu_ℓ and Nu_t into Equation (4.2.6) with $m = 6$.
2. *Vertical Flat Plate with Uniform Heat Flux, Figure 4.2.1B.* If the plate surface has a constant (known) heat flux, rather than being isothermal, the objective is to calculate the average temperature difference, $\overline{\Delta T}$, between the plate and fluid. For this situation, and for $15 < \mathrm{Ra}^* < 10^5$,

$$\mathrm{Nu}^T = \overline{G}_\ell \left(\mathrm{Ra}^*\right)^{1/5} \qquad \mathrm{Nu}_\ell = \frac{1.83}{\ln\left(1 + 1.83/\mathrm{Nu}^T\right)} \qquad \mathrm{Nu}_t = \left(C_t^V\right)^{3/4} \left(\mathrm{Ra}^*\right)^{1/4} \tag{4.2.8a}$$

$$\overline{G}_\ell = \frac{6}{5}\left(\frac{\mathrm{Pr}}{4 + 9\sqrt{\mathrm{Pr}} + 10\mathrm{Pr}}\right)^{1/5} \tag{4.2.8b}$$

   Ra^* is defined in Figure 4.2.1B and C_t^V is given by Equation (4.2.4). Find Nu by inserting these expressions for Nu_ℓ and Nu_t into Equation (4.2.6) with $m = 6$. The $\overline{G}_\ell$ expression is due to Fujii and Fujii (1976).
3. *Horizontal Upward-Facing (ϕ = 90°) Plates, Figure 4.2.1C.* For horizontal isothermal surfaces of various platforms, correlations are given in terms of a lengthscale L^* (Goldstein et al., 1973), defined in Figure 4.2.1C. For $\mathrm{Ra} \geq 1$,

$$\mathrm{Nu}^T = 0.835\overline{C}_\ell \mathrm{Ra}^{1/4} \qquad \mathrm{Nu}_\ell = \frac{2.0}{\ln\left(1 + 1.4/\mathrm{Nu}^T\right)} \qquad \mathrm{Nu}_t = C_t^H \mathrm{Ra}^{1/3} \tag{4.2.9}$$

   Nu is obtained by substituting Nu_ℓ and Nu_t from Equation 4.2.9 into Equation 4.2.6 with $m = 10$. For non-isothermal surfaces, replace ΔT by $\overline{\Delta T}$.
4. *Horizontal Downward-Facing (ϕ = –90°) Plates, Figure 4.2.1C.* For horizontal downward-facing plates of various planforms, the main buoyancy force is into the plate so that only a very weak force drives the fluid along the plate; for this reason, only laminar flows have been measured. For this case, the following equation applies for $\mathrm{Ra} < 10^{10}$, $\mathrm{Pr} \geq 0.7$:

$$\mathrm{Nu}^T = H_\ell \mathrm{Ra}^{1/5} \qquad \mathrm{H}_\ell = \frac{0.527}{\left[1 + \left(1.9/\mathrm{Pr}\right)^{9/10}\right]^{2/9}} \qquad \mathrm{Nu} = \frac{2.45}{\ln\left(1 + 2.45/\mathrm{Nu}^T\right)} \tag{4.2.10}$$

   H_ℓ fits the analysis of Fujii et al. (1973).

5. *Inclined Plates, Downward Facing ($-90° \le \phi \le 0$), Figure 4.2.1B.* First calculate q from *Case 1* with g replaced by $g \cos \phi$; then calculate q from *Case 4* (horizontal plate) with g replaced by $g \sin(-\phi)$, and use the maximum of these two values of q.
6. *Inclined Plates, Upward Facing ($0 \le \phi \le 90$), Figure 4.2.1B.* First calculate q from *Case 1* with g replaced by $g \cos \phi$; then calculate q from *Case 3* with g replaced by $g \sin \phi$, and use the maximum of these two values of q.
7. *Vertical and Tilted Isothermal Plates of Various Planform, Figure 4.2.2.* The line of constant χ in Figure 4.2.2 is the line of steepest ascent on the plate. Provided all such lines intersect the plate edges just twice, as shown in the figure, the thin-layer (Nu^T) heat transfer can be found by subdividing the body into strips of width $\Delta\chi$, calculating the heat transfer from each strip, and adding. For laminar flow from an isothermal vertical plate, this results in

$$\mathrm{Nu}^T = C_1 \overline{C}_\ell \mathrm{Ra}^{1/4} \qquad C_1 \equiv \left(\frac{L^{1/4}}{A} \int_0^W S^{3/4}\, d\chi \right) \tag{4.2.11}$$

   Symbols are defined in Figure 4.2.2, along with L and calculated C_1 values for some plate shapes. If the plate is vertical, follow the procedure in *Case 1* above (isothermal vertical flat plate) except replace the expression for Nu^T in Equation (4.2.7) by Equation (4.2.11). If the plate is tilted, follow the procedure described in *Case 5* or *6* (as appropriate) but again use Equation (4.2.11) for Nu^T in Equation (4.2.7)
8. *Horizontal Cylinders, Figure 4.2.3A.* For a long, horizontal circular cylinder use the following expressions for Nu_ℓ and Nu_t:

$$\mathrm{Nu}^T = 0.772 \overline{C}_\ell \mathrm{Ra}^{1/4} \qquad \mathrm{Nu}_\ell = \frac{2f}{\left(1 + 2f/\mathrm{Nu}^T\right)} \qquad \mathrm{Nu}_t = \overline{C}_t \mathrm{Ra}^{1/3} \tag{4.2.12}$$

   $\overline{C}_t$ is given in the table below. For $\mathrm{Ra} > 10^{-2}$, $f = 0.8$ can be used, but for $10^{-10} < \mathrm{Ra} < 10^{-2}$ use $f = 1 - 0.13/(\mathrm{Nu}^T)^{0.16}$. To find Nu, the values of Nu_ℓ and Nu_t from Equation (4.2.12) are substituted into Equation (4.2.6) with $m = 15$ (Clemes et al., 1994).

$\overline{C}_t$ for Various Shapes and Prandtl Numbers

Pr→	**0.01**	**0.022**	**0.10**	**0.71**	**2.0**	**6.0**	**50**	**100**	**2000**
Horizontal cylinder	0.077	0.81	0.90	0.103	0.108	0.109	0.100	0.097	0.088
Spheres	0.074	0.078	0.088	0.104	0.110	0.111	0.101	0.97	0.086

9. *Vertical Cylinders ($\phi = 90°$), Figure 4.2.3B.* For high Ra values and large diameter, the heat transfer from a vertical cylinder approaches that for a vertical flat plate. Let the Nu^T and Nu_ℓ equations for a vertical flat plate of height L, Equation (4.2.7), be rewritten here as Nu_p^T and Nu_p, respectively. At smaller Ra and diameter, transverse curvature plays a role which is accounted for in the following equations:

$$\mathrm{Nu}_\ell = \frac{0.9\xi \mathrm{Nu}_p}{\ln\left(1 + 0.9\xi\right)} \qquad \xi = \frac{2L/D}{Nu_p^T} \tag{4.2.13}$$

   These equations are valid for purely laminar flow. To obtain Nu, blend Equation (4.2.13) for Nu_ℓ with Equation (4.2.7) for Nu_t using Equation (4.2.6) with $m = 10$.

10. *Spheres, Figure 4.2.3C.* For spheres use Equation (4.2.6), with $m = 6$, and with

$$\mathrm{Nu}_\ell = 2 + 0.878\overline{C}_\ell \mathrm{Ra}^{1/4} \quad \text{and} \quad \mathrm{Nu}_t = \overline{C}_t \mathrm{Ra}^{1/3} \tag{4.2.14}$$

The table above contains $\overline{C}_t$ values.

11. *Combined Shapes, Figure 4.2.3D.* For combined shapes, such as the cylinder in Figure 4.2.3D with spherical end caps, calculate the heat transfer from the cylinder of length L (*Case 8*), the heat transfer from a sphere of diameter D (*Case 10*) and add to obtain the total transfer. Other shapes can be treated in a similar manner.

Correlations for Open Cavities

Examples of this class of problem are shown in Figure 4.2.4. Walls partially enclose a fluid region (cavity) where boundary openings permit fluid to enter and leave. Upstream from its point of entry, the fluid is at the ambient temperature, T_∞. Since access of the ambient fluid to the heated surfaces is restricted, some of the heated surface is starved of cool ambient to which heat can be transferred. As the sizes of the boundary openings are increased, the previous class of problems is approached; for example, when the plate spacing in Figure 4.2.4A (*Case 12*) becomes very large, the heat transfer from each vertical surface is given by *Case 1*.

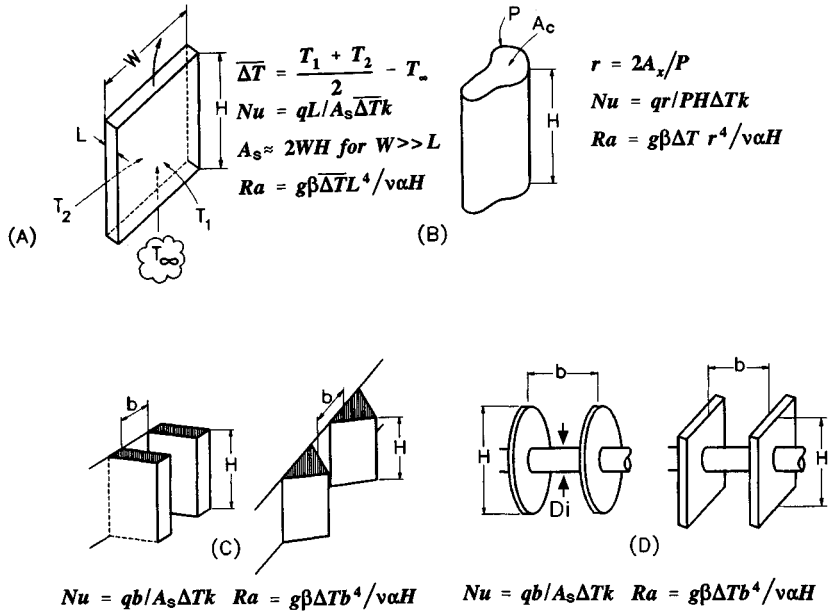


FIGURE 4.2.4 Nomenclature for various open-cavity problems.

12. *Isothermal Vertical Channels, Figure 4.2.4A and B.* Figure 4.2.4A shows an open cavity bounded by vertical walls and open at the top and bottom. The large opposing plates are isothermal, at temperatures T_1 and T_2, respectively, and the spacing between these plates is small. $\overline{\Delta T}$ is the average temperature difference between the plates and T_∞, as shown in Figure 4.2.4A, but T_1 and T_2 must not straddle T_∞. For this case

$$\mathrm{Nu} = \left(\left(\frac{\mathrm{Ra}}{f\,\mathrm{Re}} \right)^m + \left(C_1 \overline{C}_\ell \mathrm{Ra}^{1/4} \right)^m \right)^{1/m} \qquad \mathrm{Ra} \le 10^5 \tag{4.2.15}$$

where *f*Re is the product of friction factor and Reynolds number for fully developed flow through, and C_1 is a constant that accounts for the augmentation of heat transfer, relative to a vertical flat plate (*Case 1*), due to the chimney effect. The *f*Re factor accounts for the cross-sectional shape (Elenbaas, 1942a). Symbols are defined in Figure 4.2.4A and B; in the Nu equation, q is the total heat transferred to the ambient fluid from all heated surfaces.

For the parallel plate channel shown in Figure 4.2.4(A), use f Re = 24, m = –1.9, and for gases $C_1 \approx 1.2$. It should be noted, however, that C_1 must approach 1.0 as Pr increases or as the plate spacing increases. For channels of circular cross section (Figure 4.2.4B) *f*Re - 16, m = –1.03, and for gases $C_1 \approx 1.17$. For other cross-sectional shapes like the square (*f*Re = 14.23), hexagonal (*f*Re = 15.05), or equilateral triangle (*f*Re = 13.3), use Equation (4.2.15) with the appropriate *f*Re, and with m = –1.5, and $C_1 \approx 1.2$ for gases.

The heat transfer per unit cross-sectional area, q/A_c, for a given channel length H and temperature difference, passes through a maximum at approximately Ra_{max}, where

$$\mathrm{Ra}_{max} = \left(\frac{f\,\mathrm{Re} C_1 \overline{C}_\ell}{2^{1/m}} \right)^{4/3} \tag{4.2.16}$$

Ra_{max} provides the value of hydraulic radius $r = 2A_c/P$ at this maximum.

13. *Isothermal Triangular Fins, Figure 4.2.4C.* For a large array of triangular fins (Karagiozis et al., 1994) in air, for $0.4 < \mathrm{Ra} < 5 \times 10^5$

$$\mathrm{Nu} = \overline{C}_\ell \mathrm{Ra}^{1/4} \left[1 + \left(\frac{3.26}{\mathrm{Ra}^{0.21}} \right)^3 \right]^{-1/3} \qquad 0.4 < \mathrm{Ra} < 5 \times 10^5 \tag{4.2.17}$$

In this equation, b is the average fin spacing (Figure 4.2.4C), defined such that bL is the cross-sectional flow area between two adjacent fin surfaces up to the plane of the fin tips. For Ra < 0.4, Equation (4.2.17) underestimates the convective heat transfer. When such fins are mounted horizontally (vertical baseplate, but the fin tips are horizontal), there is a substantial reduction of the convective heat transfer (Karagiozis et al., 1994).

14. *U-Channel Fins, Figure 4.2.4C.* For the fins most often used as heat sinks, there is uncertainty about the heat transfer at low Ra. By using a conservative approximation applying for Ra < 100 (that underestimates the real heat transfer), the following equation may be used:

$$\mathrm{Nu} = \left[\left(\frac{\mathrm{Ra}}{24} \right)^{-2} + \left(C_1 \overline{C}_\ell \mathrm{Ra} \right)^{-2} \right]^{-0.5} \tag{4.2.18}$$

For air C_1 depends on aspect ratio of the fin as follows (Karagiozis, 1991):

$$C_1 = \left[1 + \left(\frac{H}{b} \right), 1.16 \right]_{\min} \tag{4.2.19}$$

Equation (4.2.18) agrees well with measurements for Ra > 200, but for smaller Ra it falls well below data because the leading term does not account for heat transfer from the fin edges and for three-dimensional conduction from the entire array.

15. *Circular Fins on a Horizontal Tube, Figure 4.24D.* For heat transfer from an array of circular fins (Edwards and Chaddock, 1963), for H/D_i = 1.94, $5 < \mathrm{Ra} < 10^4$, and for air,

$$\mathrm{Nu} = 0.125\mathrm{Ra}^{0.55}\left[1-\exp\left(-\frac{137}{\mathrm{Ra}}\right)\right]^{0.294} \tag{4.2.20}$$

A more general, but also more complex, relation is reported by Raithby and Hollands (1985).

16. *Square Fins on a Horizontal Tube, Figure 4.2.4D.* Heat transfer (Elenbaas, 1942b) from the square fins (excluding the cylinder that connects them) is correlated for gases by

$$\mathrm{Nu} = \left[\left(\mathrm{Ra}^{0.89}/18\right)^m + \left(0.62\mathrm{Ra}^{1/4}\right)^m\right]^{1/m} \qquad m = -2.7 \tag{4.2.21}$$

Heat Transfer in Enclosures

This section deals with cavities where the bounding walls are entirely closed, so that no mass can enter or leave the cavity. The fluid motion inside the cavity is driven by natural convection, which enhances the heat transfer among the interior surfaces that bound the cavity.

17. *Extensive Horizontal Layers, Figure 4.2.5A with $\theta = 0°$.* If the heated plate, in a horizontal parallel-plate cavity, is on the top, heat transfer is by conduction alone, so that Nu = 1. For heat transfer from below (Hollands, 1984):

$$\mathrm{Nu} = 1 + \left[1-\frac{1708}{\mathrm{Ra}}\right]^{\bullet}\left[k_1 + 2\left(\frac{\mathrm{Ra}^{1/3}}{k_2}\right)^{1-\ln\left(\mathrm{Ra}^{1/3}/k_2\right)}\right] + \left[\left(\frac{\mathrm{Ra}}{5830}\right)^{1/3} - 1\right]^{\bullet} \tag{4.2.22}$$

where

$$[x]^{\bullet} = (x,0)_{\max} \qquad k_1 = \frac{1.44}{1 + 0.018/\mathrm{Pr} + 0.00136/\mathrm{Pr}^2} \qquad k_2 = 75\exp\left(1.5\mathrm{Pr}^{-1/2}\right) \tag{4.2.23}$$

The equation has been validated for Ra < 10^{11} for water, Ra < 10^8 for air, and over a smaller Ra range for other fluids. Equation (4.2.22) applies to extensive layers: $W/L \geq 5$. Correlations for nonextensive layers are provided by Raithby and Hollands (1985).

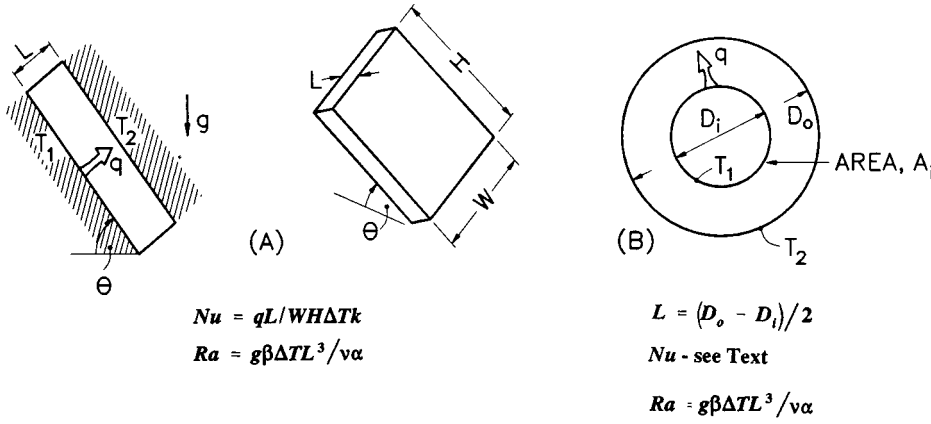


FIGURE 4.2.5 Nomenclature for enclosure problems.

18. *Vertical Layers, Figure 4.2.5(A), with $\theta = 90°$. W/L > 5.* For a vertical, gas-filled (Pr ≈ 0.7) cavity with H/L ≥ 5, the following equation closely fits the data, for example that of Shewen et al. (1996) for $\mathrm{Ra}(H/L)^3 \le 5 \times 10^{10}$ and $H/L \ge 40$.

$$\mathrm{Nu}_1 = \left\{1 + \left[\frac{0.0665\mathrm{Ra}^{1/3}}{1 + \left(\frac{9000}{\mathrm{Ra}}\right)^{1.4}}\right]^2\right\}^{1/2} \qquad \mathrm{Nu}_2 = 0.242\left(\mathrm{Ra}\frac{L}{H}\right)^{0.273} \qquad \mathrm{Nu} = \left[\mathrm{Nu}_1, \mathrm{Nu}_2\right]_{\max} \tag{4.2.24}$$

For Pr ≥ 4, the following equation is recommended (Seki et al., 1978) for $\mathrm{Ra}(H/L)^3 < 4 \times 10^{12}$

$$\mathrm{Nu} = \left[1, 0.36\mathrm{Pr}^{0.051}\left(\frac{L}{H}\right)^{0.36}\mathrm{Ra}^{0.25}, 0.084\mathrm{Pr}^{0.051}\left(\frac{L}{H}\right)^{0.1}\mathrm{Ra}^{0.3}\right]_{\max} \tag{4.2.25a}$$

and for Ra $(H/L)^3 > 4 \times 10^{12}$

$$\mathrm{Nu} = 0.039\mathrm{Ra}^{1/3} \tag{4.2.25b}$$

19. *Tilted Layers, Figure 4.25A, with $0 \le \theta \le 90°$, W/L > 8.* For gases (Pr ≈ 0.7), $0 \le \theta \le 60°$ and $\mathrm{Ra} \le 10^5$ (Hollands et al., 1976), use

$$\mathrm{Nu} = 1 + 1.44\left[1 - \frac{1708}{\mathrm{Ra}\cos\theta}\right]^{\bullet}\left[1 - \frac{1708(\sin 1.8\theta)^{1.6}}{\mathrm{Ra}\cos\theta}\right] + \left[\left(\frac{\mathrm{Ra}\cos\theta}{5830}\right)^{1/3} - 1\right]^{\bullet} \tag{4.2.26}$$

See equation (4.2.23) for definition of $[x]°$. For $60° \le \theta \le 90°$ linear interpolation is recommended using Equations (4.2.24) for $\theta = 90°$ and (4.2.26) for $\theta = 60°$.

20. *Concentric Cylinders, Figure 4.2.5B.* For heat transfer across the gap between horizontal concentric cylinders, the Nusselt number is defined as $\mathrm{Nu} = q'\ln(D_o/D_i)/2\pi\Delta T$ where q' is the heat transfer per unit length of cylinder. For $\mathrm{Ra} \le 8 \times 10^7$, $0.7 \le \mathrm{Pr} \le 6000$, $1.15 \le D/D_i \le 8$ (Raithby and Hollands, 1975)

$$\mathrm{Nu} = \left[0.603\overline{C}_\ell \frac{\ln\left(D_o/D_i\right)\mathrm{Ra}^{1/4}}{\left[\left(L/D_i\right)^{3/5} + \left(L/D_o\right)^{3/5}\right]^{5/4}}, 1\right]_{\max} \tag{4.2.27}$$

For eccentric cylinders, see Raithby and Hollands (1985).

21. *Concentric Spheres, Figure 4.2.5B.* The heat transfer between concentric spheres is given by the following equation (Raithby and Hollands, 1975) for $\mathrm{Ra} \le 6 \times 10^8$, $5 \le \mathrm{Pr} \le 4000$, $1.25 < D_o/D_i \le 2.5$,

$$\mathrm{Nu} = \frac{qL}{D_i D_o k\Delta T} = \left[1.16\overline{C}_\ell\left(\frac{L}{D_i}\right)^{1/4} \frac{\mathrm{Ra}^{1/4}}{\left[\left(D_i/D_o\right)^{3/5} + \left(D_o/D_i\right)^{4/5}\right]^{5/4}}, 1\right]_{\max} \tag{4.2.28}$$

For eccentric spheres, see Raithby and Hollands (1985).

Example Calculations

Problem 1: Heat Transfer from Vertical Plate, Figure 4.2.6A. For the vertical isothermal surface in Figure 4.2.6A with T_s = 40°C, H_1 = 1 m, H_2 = 1 m, W_1 = 1 m, W_2 = 1 m and for an ambient air temperature of T_∞ = 20°C (at 1 atm), find the heat transfer from one side of the plate.

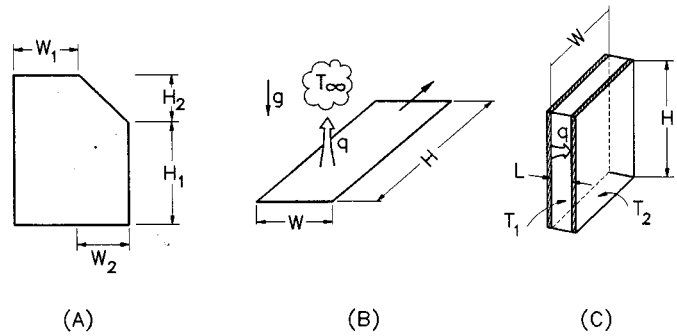


FIGURE 4.2.6 Sketches for example problems.

Properties: At $T_f = (T_w + T_\infty)/2$ = 30°C and atmospheric pressure for air: $\nu = 1.59 \times 10^{-5}$ m²/sec, Pr = 0.71, k = 0.0263 W/mK. At T_∞, $\beta \approx 1/T_\infty = 1(273 + 20) = 0.00341$ K⁻¹.

Solution: For the geometry shown in Figure 4.2.6A:

$$A_s = (H_1 + H_2)W_1 + \left(H_1 + \frac{H_2}{2}\right)W_2 = 3.5 \text{ m}^2 \quad \text{(plate surface area)}$$

$$\int_0^{W_1+W_2} S^{3/4}\, d\chi = (H_1 + H_2)^{3/4} W_1 + \frac{4}{7}\frac{W_2}{H_2}\left[(H_1 + H_2)^{7/4} - H_1^{7/4}\right] = 3.03 \text{ m}^{7/4}$$

$$L^{1/4} = (H_1 + H_2)^{1/4} = 1.19 \text{ m}^{1/4} \quad \text{(see comments below)}$$

$$C_1 = \frac{L^{1/4}\int_0^{W_1+W_2} S^{3/4}\, d\chi}{A_s} = \frac{1.19 \times 3.03}{3.5} = 1.03$$

$$\text{Ra} = \frac{g\beta_\infty L^3 (T_w - T_\infty)}{\nu\alpha} = \frac{9.81 \times 0.00341 \times 2^3 \times (40 - 20)}{1.59 \times 10^{-5} \times 2.25 \times 10^{-5}} = 1.50 \times 10^{10}$$

$\overline{C}_\ell$ = 0.514 from Equation (4.2.3); $C_t = C_t^V$ = 0.103 from Equation (4.2.4). $Nu^T = C_1 \overline{C}_\ell Ra^{1/4}$ = 185 from Equation (4.2.11).

$$\left.\begin{aligned} \text{Nu}_\ell &= \frac{2.0}{\ln\left(1 + 2.0/\text{Nu}^T\right)} = 186 \\ \text{Nu}_t &= C_t^V \text{Ra}^{1/3} \Big/ \left(1 + 1.4 \times 10^9 \text{ Pr/Ra}\right) = 238 \end{aligned}\right\} \text{(from Equation (4.2.7))}$$

$$\text{Nu} = \frac{qL}{A\Delta Tk} = \left(\text{Nu}_\ell^6 + \text{Nu}_t^6\right)^{1/6} = 246$$

from Equation (4.2.6) with $m = 6$.

$$q = \frac{A_s \Delta T k \mathrm{Nu}}{\mathrm{L}} = \frac{3.5 \times 20 \times 0.0263 \times 246}{2} = 226 W$$

Comments on Problem 1: Since $\mathrm{Nu}_\ell < \mathrm{Nu}_t$, the heat transfer is primarily turbulent. Do not neglect radiation. Had the surface been specified to be at constant heat flux, rather than isothermal, the equations in this section can be used to find the approximate average temperature difference between the plate and fluid.

Problem 2: Heat Transfer from Horizontal Strip, Figure 4.2.6B. Find the rate of heat loss per unit length from a very long strip of width $W = 0.1$ m with a surface temperature of $T_s = 70°C$ in water at $T_\infty = 30°C$.

Properties: At $T_f = (T_s + T_\infty)1/2 = 50°C$

$$\nu = 5.35 \times 10^{-7} \text{ m}^2/\text{sec} \qquad \alpha = 1.56 \times 10^{-7} \text{ m}^2/\text{sec} \qquad \mathrm{Pr} = 3.42$$

$$k = 0.645 \text{ W/mK} \qquad \beta = 2.76 \times 10^{-4} \text{ K}^{-1}$$

Solution: This problem corresponds to *Case 3* and Figure 4.2.1C.

$$C_t^H = 0.14$$

from Equation 4.2.5 and $\overline{C}_l = 0.563$ from Equation (4.2.3).

$$L^* = \lim_{H \to \infty} \left(\frac{WH}{2W + 2H} \right) = \frac{W}{2} = 0.05 \text{ m}$$

from Figure 4.2.1C.

$$\mathrm{Ra} = \frac{g \beta \Delta T L^{*3}}{\nu \alpha} = 1.62 \times 10^8 \qquad \mathrm{Nu}^T = 0.835 \overline{C}_\ell \mathrm{Ra}^{1/4} = 53.5$$

$$\mathrm{Nu}_\ell = \frac{1.4}{\ln\left(1 + 1.4/\mathrm{Nu}^T\right)} = 54.2 \qquad \mathrm{Nu}_t = C_t^H \mathrm{Ra}^{1/3} = 76.3$$

$$\mathrm{Nu} = \frac{q}{WH\Delta T} \frac{L^*}{k} = \left(\mathrm{Nu}_\ell^{10} + \mathrm{Nu}_t^{10} \right)^{0.1} = 76.5$$

$$q/H = \frac{W \Delta T k \mathrm{Nu}}{L^*} = 3950 \text{ W/m-length}$$

Comments: Turbulent heat transfer is dominant. Radiation can be ignored (since it lies in the far infrared region where it is not transmitted by the water).

Problem 3: Heat Loss across a Window Cavity, Figure 4.2.6C. The interior glazing is at temperature $T_1 = 10°C$, the exterior glazing at $T_2 = -10°C$, the window dimensions are $W = 1$ m, $H = 1.7$ m, and the air gap between the glazings is $L = 1$ cm and is at atmospheric pressure. Find the heat flux loss across the window.

Properties: At $\bar{T} = T_1 + T_2/2 = 0°C = 273K$

$$\nu = 1.35\times10^{-5}\ \text{m}^2/\text{sec} \qquad \alpha = 1.89\times10^{-5}\ \text{m}^2/\text{sec} \qquad \text{Pr} = 0.71$$

$$k = 0.024\ \text{W/mK} \qquad \beta = 1/273 = 3.66\times10^{-3}\ \text{K}^{-1}$$

Solution: The appropriate correlations are given in *Case 18* and by Equation (4.2.24).

$$\text{Ra} = \frac{g\beta(T_1 - T_2)L^3}{\nu\alpha} = \frac{9.81\times3.66\times10^{-3}\times20\times(0.01)^3}{1.35\times10^{-5}\times1.89\times10^{-5}} = 2.81\times10^3$$

$$\text{Nu}_1 = \left[1+\left\{\frac{0.0665\text{Ra}^{1/3}}{1+\left(\frac{9000}{\text{Ra}}\right)^{1.4}}\right\}^2\right]^{1/2} = 1.01$$

$$\text{Nu}_2 = 0.242\left(\text{Ra}\frac{L}{H}\right)^{0.273} = 0.242\left(2.81\times10^3\times\frac{0.01}{1.7}\right)^{0.273} = 0.520$$

$$\text{Nu} = \frac{qL}{WH(T_1 - T_2)k} = (\text{Nu}_1, \text{Nu}_2)_{\max} = 1.01$$

$$q/WH = \frac{N(T_1 - T_2)k}{L} = \frac{1.01\times20\times0.24}{0.01} = 48.5\ \text{W/m}^2$$

Comments: For pure conduction across the air layer, Nu = 1.0. For the calculated value of Nu = 1.01, convection must play little role. For standard glass, the heat loss by radiation would be roughly double the natural convection value just calculated.

Special Nomenclature

Note that nomenclature for each geometry considered is provided in the figures that are referred to in the text.

$\overline{C}_\ell$ = function of Prandtl number, Equation (4.2.3)
C_t^V = function of Prandtl number, Equation (4.2.4)
C_t^H = function of Prandtl number, Equation (4.2.5)
$\overline{C}_t$ = surface averaged value of C_t, page 4–38
$\overline{\Delta T}$ = surface averaged value of $T_w - T_\infty$

References

Churchill, S.W. 1983. *Heat Exchanger Design Handbook*, Sections 2.5.7 to 2.5.10, E.V. Schlinder, Ed., Hemisphere Publishing, New York.

Churchill S.W. and Usagi, R. 1972. A general expression for the correlation of rates of transfer and other phenomena, *AIChE J.*, 18, 1121–1128.

Clemes, S.B., Hollands, K.G.T., and Brunger, A.P. 1994. Natural convection heat transfer from horizontal isothermal cylinders, *J. Heat Transfer*, 116, 96–104.

Edwards, J.A. and Chaddock, J.B. 1963. An experimental investigation of the radiation and free-convection heat transfer from a cylindrical disk extended surface, *Trans., ASHRAE*, 69, 313–322.
Elenbaas, W. 1942a. The dissipation of heat by free convection: the inner surface of vertical tubes of different shapes of cross-section, *Physica*, 9(8), 865–874.
Elenbaas, W. 1942b. Heat dissipation of parallel plates by free convection, *Physica*, 9(1), 2–28.
Fujii, T. and Fujii, M. 1976. The dependence of local Nusselt number on Prandtl number in the case of free convection along a vertical surface with uniform heat flux, *Int. J. Heat Mass Transfer*, 19, 121–122.
Fujii, T., Honda, H., and Morioka, I. 1973. A theoretical study of natural convection heat transfer from downward-facing horizontal surface with uniform heat flux, *Int. J. Heat Mass Transfer*, 16, 611–627.
Goldstein, R.J., Sparrow, E.M., and Jones, D.C. 1973. Natural convection mass transfer adjacent to horizontal plates, *Int. J. Heat Mass Transfer*, 16, 1025–1035.
Hollands, K.G.T. 1984. Multi-Prandtl number correlations equations for natural convection in layers and enclosures, *Int. J. Heat Mass Transfer*, 27, 466–468.
Hollands, K.G.T., Unny, T.E., Raithby, G.D., and Konicek, K. 1976. Free convection heat transfer across inclined air layers, *J. Heat Transfer*, 98, 189–193.
Incropera, F.P. and DeWitt, D.P. 1990. *Fundamentals of Heat and Mass Transfer*, 3rd ed., John Wiley & Sons, New York.
Karagiozis, A. 1991. An Investigation of Laminar Free Convection Heat Transfer from Isothermal Finned Surfaces, Ph.D. Thesis, Department of Mechanical Engineering, University of Waterloo.
Karagiozis, A., Raithby, G.D., and Hollands, K.G.T. 1994. Natural convection heat transfer from arrays of isothermal triangular fins in air, *J. Heat Transfer*, 116, 105–111.
Kreith, F. and Bohn, M.S. 1993. *Principles of Heat Transfer.* West Publishing, New York.
Raithby, G.D. and Hollands, K.G.T. 1975. A general method of obtaining approximate solutions to laminar and turbulent free convection problems, in *Advances in Heat Transfer*, Irvine, T.F. and Hartnett, J.P., Eds., Vol. 11, Academic Press, New York, 266–315.
Raithby, G.D. and Hollands, K.G.T. 1985. *Handbook Heat Transfer*, Chap. 6: Natural Convection, Rohsenow, W.M., Hartnett, J.P., and Ganic, E.H., Eds., McGraw-Hill, New York.
Seki, N., Fukusako, S., and Inaba, H. 1978. Heat transfer of natural convection in a rectangular cavity with vertical walls of different temperatures, *Bull. JSME.*, 21(152), 246–253.
Shewan, E., Hollands, K.G.T., and Raithby, G.D. 1996. Heat transfer by natural convection across a vertical air cavity of large aspect ratio, *J. Heat Transfer*, 118, 993–995.

Further Information

There are several excellent heat transfer textbooks that provide fundamental information and correlations for natural convection heat transfer (e.g., Kreith and Bohn, 1993; Incropera and DeWitt, 1990). The correlations in this section closely follow the recommendations of Raithby and Hollands (1985), but that reference considers many more problems. Alternative equations are provided by Churchill (1983).

Forced Convection — External Flows

N.V. Suryanarayana

Introduction

In this section we consider heat transfer between a solid surface and an adjacent fluid which is in motion relative to the solid surface. If the surface temperature is different from that of the fluid, heat is transferred as forced convection. If the bulk motion of the fluid results solely from the difference in temperature of the solid surface and the fluid, the mechanism is natural convection. The velocity and temperature of the fluid far away from the solid surface are the free-stream velocity and free-stream temperature. Both

are usually known or specified. We are then required to find the heat flux from or to the surface with specified surface temperature or the surface temperature if the heat flux is specified. The specified temperature or heat flux either may be uniform or may vary. The convective heat transfer coefficient h is defined by

$$q'' = h\left(T_s - T_\infty\right) \tag{4.2.29}$$

In Equation (4.2.29) with the local heat flux, we obtain the local heat transfer coefficient, and with the average heat flux with a uniform surface temperature we get the average heat transfer coefficient. For a specified heat flux the local surface temperature is obtained by employing the local convective heat transfer coefficient.

Many correlations for finding the convective heat transfer coefficient are based on experimental data which have some uncertainty, although the experiments are performed under carefully controlled conditions. The causes of the uncertainty are many. Actual situations rarely conform completely to the experimental situations for which the correlations are applicable. Hence, one should not expect the actual value of the heat transfer coefficient to be within better than ±10% of the predicted value.

Many different correlations to determine the convective heat transfer coefficient have been developed. In this section only one or two correlations are given. For other correlations and more details, refer to the books given in the bibliography at the end of this section.

Flat Plate

With a fluid flowing parallel to a flat plate, changes in velocity and temperature of the fluid are confined to a thin region adjacent to the solid boundary — the boundary layer. Several cases arise:

1. Flows without or with pressure gradient
2. Laminar or turbulent boundary layer
3. Negligible or significant viscous dissipation (effect of frictional heating)
4. $\mathrm{Pr} \geq 0.7$ or $\mathrm{Pr} \ll 1$

Flows with Zero Pressure Gradient and Negligible Viscous Dissipation

When the free-stream pressure is uniform, the free-stream velocity is also uniform. Whether the boundary layer is laminar or turbulent depends on the Reynolds number Re_x ($\rho U_\infty x/\mu$) and the shape of the solid at entrance. With a sharp edge at the leading edge (Figure 4.2.7) the boundary layer is initially laminar but at some distance downstream there is a transition region where the boundary layer is neither totally laminar nor totally turbulent. Farther downstream of the transition region the boundary layer becomes turbulent. For engineering applications the existence of the transition region is usually neglected and it is assumed that the boundary layer becomes turbulent if the Reynolds number, Re_x, is greater than the critical Reynolds number, Re_{cr}. A typical value of 5×10^5 for the critical Reynolds number is generally accepted, but it can be greater if the free-stream turbulence is low and lower if the free-stream turbulence is high, the surface is rough, or the surface does not have a sharp edge at entrance. If the entrance is blunt, the boundary layer may be turbulent from the leading edge.

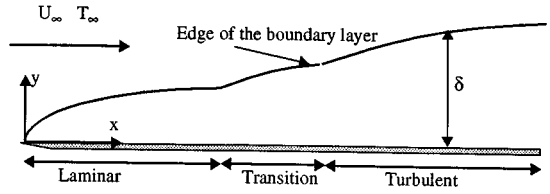


FIGURE 4.2.7 Flow of a fluid over a flat plate with laminar, transition, and turbulent boundary layers.

Temperature Boundary Layer

Analogous to the velocity boundary layer there is a temperature boundary layer adjacent to a heated (or cooled) plate. The temperature of the fluid changes from the surface temperature at the surface to the free-stream temperature at the edge of the temperature boundary layer (Figure 4.2.8).

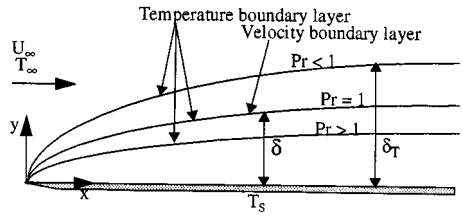


FIGURE 4.2.8 Temperature boundary layer thickness relative to velocity boundary layer thickness.

The velocity boundary layer thickness δ depends on the Reynolds number Re_X. The thermal boundary layer thickness δ_T depends both on Re_X and Pr

$Re_x < Re_{cr}$:

$$\frac{\delta}{x} = \frac{5}{\sqrt{Re_x}} \qquad Pr > 0.7 \qquad \frac{\delta}{\delta_T} = Pr^{1/3}$$
$$Pr \ll 1 \qquad \frac{\delta}{\delta_T} = Pr^{1/2} \tag{4.2.30}$$

$Re_{cr} < Re_x$:

$$\frac{\delta}{x} = \frac{0.37}{Re_x^{0.2}} \qquad \delta \approx \delta_T \tag{4.2.31}$$

Viscous dissipation and high-speed effects can be neglected if $Pr^{1/2}\, Ec/2 \ll 1$. For heat transfer with significant viscous dissipation see the section on flow over flat plate with zero pressure gradient: Effect of High Speed and Viscous Dissipation. The Eckert number Ec is defined as $Ec = U_\infty^2 / C_p(T_s - T_\infty)$.

With a rectangular plate of length L in the direction of the fluid flow the average heat transfer coefficient h_L with uniform surface temperature is given by

$$h_L = \frac{1}{L}\int_0^L h_x\, dx$$

Laminar Boundary Layer *($Re_x < Re_{cr}$, $Re_L < Re_{cr}$)*: With heating or cooling starting from the leading edge the following correlations are recommended. Note: in all equations evaluate fluid properties at the film temperature defined as the arithmetic mean of the surface and free-stream temperatures unless otherwise stated (Figure 4.2.9).

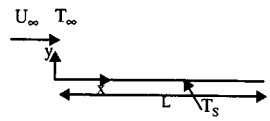


FIGURE 4.2.9 Heated flat plate with heating from the leading edge.

Local Heat Transfer Coefficient (Uniform Surface Temperature)
The Nusselt number based on the local convective heat transfer coefficient is expressed as

$$\mathrm{Nu}_x = f_{\mathrm{Pr}}\mathrm{Re}_x^{1/2} \tag{4.2.32}$$

The classical expression for f_{Pr} is $0.564\,\mathrm{Pr}^{1/2}$ for liquid metals with very low Prandtl numbers, $0.332\mathrm{Pr}^{1/3}$ for $0.7 < \mathrm{Pr} < 50$ and $0.339\mathrm{Pr}^{1/3}$ for very large Prandtl numbers. Correlations valid for all Prandtl numbers developed by Churchill (1976) and Rose (1979) are given below.

$$\mathrm{Nu}_x = \frac{0.3387\mathrm{Re}_x^{1/2}\mathrm{Pr}^{1/3}}{\left[1+\left(\frac{0.0468}{\mathrm{Pr}}\right)^{2/3}\right]^{1/4}} \tag{4.2.33}$$

$$\mathrm{Nu}_x = \frac{\mathrm{Re}^{1/2}\mathrm{Pr}^{1/2}}{\left(27.8+75.9\mathrm{Pr}^{0.306}+657\mathrm{Pr}\right)^{1/6}} \tag{4.2.34}$$

In the range $0.001 < \mathrm{Pr} < 2000$, Equation (4.2.33) is within 1.4% and Equation (4.2.34) is within 0.4% of the exact numerical solution to the boundary layer energy equation.

Average Heat Transfer Coefficient
The average heat transfer coefficient is given by

$$\mathrm{Nu}_L = 2\mathrm{Nu}_{x=L} \tag{4.2.35}$$

From Equation 4.2.35 it is clear that the average heat transfer coefficient over a length L is twice the local heat transfer coefficient at x = L.

Uniform Heat Flux

Local Heat Transfer Coefficient
Churchill and Ozoe (1973) recommend the following single correlation for all Prandtl numbers.

$$\mathrm{Nu}_x = \frac{0.886\mathrm{Re}_x^{1/2}\mathrm{Pr}^{1/2}}{\left[1+\left(\frac{\mathrm{Pr}}{0.0207}\right)^{2/3}\right]^{1/4}} \tag{4.2.36}$$

Note that for surfaces with uniform heat flux the local convective heat transfer coefficient is used to determine the local surface temperature. The total heat transfer rate being known, an average heat transfer coefficient is not needed and not defined.

Turbulent Boundary Layer *($\mathrm{Re}_x > \mathrm{Re}_{cr}$, $\mathrm{Re}_L > \mathrm{Re}_{cr}$):* For turbulent boundary layers with heating or cooling starting from the leading edge use the following correlations:
Local Heat Transfer Coefficient

$\mathrm{Re}_{cr} < \mathrm{Re}_x < 10^7$:

$$\mathrm{Nu}_x = 0.0296\mathrm{Re}_x^{4/5}\mathrm{Pr}^{1/3} \tag{4.2.37}$$

$10^7 < \mathrm{Re}_x$:

$$\mathrm{Nu}_x = 1.596\mathrm{Re}_x\left(\ln\mathrm{Re}_x\right)^{-2.584}\mathrm{Pr}^{1/3} \tag{4.2.38}$$

Equation (4.2.38) is obtained by applying Colburn's *j* factor in conjunction with the friction factor suggested by Schlicting (1979).

In laminar boundary layers, the convective heat transfer coefficient with uniform heat flux is approximately 36% higher than with uniform surface temperature. With turbulent boundary layers, the difference is very small and *the correlations for the local convective heat transfer coefficient can be used for both uniform surface temperature and uniform heat flux.*

Average Heat Transfer Coefficient

If the boundary layer is initially laminar followed by a turbulent boundary layer at $Re_x = Re_{cr}$, the following correlations for $0.7 < Pr < 60$ are suggested:

$Re_{cr} < Re_L < 10^7$:

$$Nu_L = \left[0.664Re_L^{1/2} + 0.037\left(Re_L^{4/5} - Re_{cr}^{4/5}\right)\right]Pr^{1/3} \tag{4.2.39}$$

If $Re_{cr} < Re_L < 10^7$ and $Re_{cr} = 10^5$, Equation 4.2.39 simplifies to

$$Nu_L = \left(0.037Re_L^{4/5} - 871\right)Pr^{1/3} \tag{4.2.40}$$

$10^7 < Re_L$ and $Re_{cr} = 5 \times 10^5$:

$$Nu_L = \left[1.963Re_L\left(\ln Re_L\right)^{-2.584} - 871\right]Pr^{1/3} \tag{4.2.41}$$

Uniform Surface Temperature — Pr > 0.7: Unheated Starting Length

If heating does not start from the leading edge as shown in Figure 4.2.10, the correlations have to be modified. Correlation for the local convective heat transfer coefficient for laminar and turbulent boundary layers are given by Equations (4.2.42) and (4.2.43) (Kays and Crawford, 1993) — the constants in Equations (4.2.42) and (4.2.43) have been modified to be consistent with the friction factors. These correlations are also useful as building blocks for finding the heat transfer rates when the surface temperature varies in a predefined manner. Equations (4.2.44) and (4.2.45), developed by Thomas (1977), provide the average heat transfer coefficients based on Equations (4.2.42) and (4.2.43).

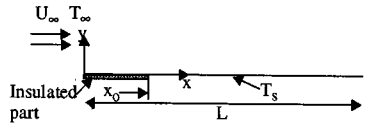


FIGURE 4.2.10 Heated flat plate with unheated starting length.

Local Convective Heat Transfer Coefficient

$Re_x < Re_{cr}$:

$$Nu_x = \frac{0.332Re_x^{1/2}Pr^{1/3}}{\left[1-\left(\frac{x_o}{x}\right)^{3/4}\right]^{1/3}} \tag{4.2.42}$$

$Re_x > Re_{cr}$:

$$Nu_x = \frac{0.0296Re_x^{4/5}Pr^{3/5}}{\left[1-\left(\frac{x_o}{x}\right)^{9/10}\right]^{1/9}} \tag{4.2.43}$$

Average Heat Transfer Coefficient over the Length ($L - x_o$)

$Re_L < Re_{cr}$:

$$h_{L-x_o} = \frac{0.664 Re_L^{1/2} Pr^{1/3}\left[1-\left(\frac{x_o}{L}\right)^{3/4}\right]^{2/3} k}{L - x_o} = 2\frac{1-\left(\frac{x_o}{L}\right)^{3/4}}{1 - x_o/L} h_{x=L} \tag{4.2.44}$$

In Equation (4.2.44) evaluate $h_{x=L}$ from Equation (4.2.42).

$Re_{cr} = 0$:

$$h_{L-x_o} = \frac{0.037 Re_L^{4/5} Pr^{3/5}\left[1-\left(\frac{x_o}{L}\right)^{9/10}\right]^{8/9} k}{L - x_o} = 1.25\frac{1-\left(x_o/L\right)^{9/10}}{1 - x_o/L} h_{x=L} \tag{4.2.45}$$

In Equation (4.2.45) evaluate $h_{x=L}$ from Equation (4.2.43).

Flat Plate with Prescribed Nonuniform Surface Temperature

The linearity of the energy equation permits the use of Equations (4.2.42) through (4.2.45) for uniform surface temperature with unheated starting length to find the local heat flux and the total heat transfer rate by the principle of superposition when the surface temperature is not uniform. Figure 4.2.11 shows the arbitrarily prescribed surface temperature with a uniform free-stream temperature of the fluid. If the surface temperature is a differentiable function of the coordinate *x*, the local heat flux can be determined by an expression that involves integration (refer to Kays and Crawford, 1993). If the surface temperature can be approximated as a series of step changes in the surface temperature, the resulting expression for the local heat flux and the total heat transfer rate is the summation of simple algebraic expressions. Here the method using such an algebraic simplification is presented.

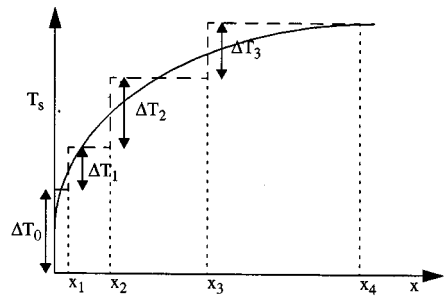


FIGURE 4.2.11 Arbitrary surface temperature approximated as a finite number of step changes.

The local convective heat flux at a distance x from the leading edge is given by

$$q_x'' = \sum_1^n h_{xi}\Delta T_{si} \tag{4.2.46}$$

where h_{xi} denotes the local convective heat transfer coefficient at x due to a single step change in the surface temperature ΔT_{si} at location $x_i (x_i < x)$. Referring to Figure 4.2.11, the local convective heat flux at x ($x_3 < x < x_4$) is given by

$$q_x'' = h_x(x,0)\Delta T_o + h_x(x,x_1)\Delta T_1 + h_x(x,x_2)\Delta T_2 + h_x(x,x_3)\Delta T_3$$

where $h_x(x, x_1)$ is the local convective heat transfer coefficient at x with heating starting from x_1; the local convective heat transfer is determined from Equation (4.2.42) if the boundary layer is laminar and Equation (4.2.43) if the boundary layer is turbulent from the leading edge. For example, $h_x(x, x_2)$ in the third term is given by

$$\mathrm{Re}_x < \mathrm{Re}_{cr} \qquad h_x(x,x_2) = \frac{0.332\left(\dfrac{\rho U_\infty x}{\mu}\right)^{1/2}\mathrm{Pr}^{1/3}}{\left[1-\left(\dfrac{x_2}{x}\right)^{3/4}\right]^{1/3}}\frac{k}{x}$$

$$\mathrm{Re}_{cr} = 0 \qquad h_x(x,x_2) = \frac{0.0296\left(\dfrac{\rho U_\infty x}{\mu}\right)^{4/5}\mathrm{Pr}^{3/5}}{\left[1-\left(\dfrac{x_2}{x}\right)^{9/10}\right]^{1/9}}\frac{k}{x}$$

The procedure for finding the total heat transfer rate from $x = 0$ to $x = L$ is somewhat similar. Denoting the width of the plate by W,

$$\frac{q}{W} = \sum h_{L-x_i}\Delta T_i(L - x_i) \tag{4.2.47}$$

where h_{L-x_i} is the average heat transfer coefficient over the length $L - x_i$ due to a step change ΔT_i in the surface temperature at x_i. For example, the heat transfer coefficient in the third term in Equation (4.2.47) obtained by replacing x_o by x_2 in Equation (4.2.44) or (4.2.45) depending on whether $\mathrm{Re}_L < \mathrm{Re}_{cr}$ or $\mathrm{Re}_{cr} = 0$.

Flows with Pressure Gradient and Negligible Viscous Dissipation

Although correlations for flat plates are for a semi-infinite fluid medium adjacent to the plate, most applications of practical interest deal with fluid flowing between two plates. If the spacing between the plates is significantly greater than the maximum boundary layer thickness, the medium can be assumed to approach a semi-infinite medium. In such a case if the plates are parallel to each other and if the pressure drop is negligible compared with the absolute pressure, the pressure gradient can be assumed to be negligible. If the plates are nonparallel and if the boundary layer thickness is very much smaller than the spacing between the plates at that location, the medium can still be considered as approaching a semi-infinite medium with a non-negligible pressure gradient. In such flows the free-stream velocity (core velocity outside the boundary layer) is related to the pressure variation by the Bernoulli equation:

$$\frac{p}{\rho} + \frac{U_\infty^2}{2} + zg = \text{constant}$$

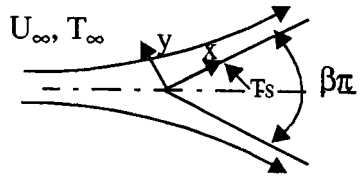


FIGURE 4.2.12 Flow over a wedge. βπ is the wedge angle.

Another situation where the free-stream velocity varies in the direction of flow giving rise to a pressure gradient is flow over a wedge. For the family of flows for which the solutions are applicable, the free-stream velocity at the edge of the boundary layer is related to the x-coordinate by a power law, $U_\infty = cx^m$. Flows over semi-infinite wedges (Figure 4.2.12) satisfy that condition. The exponent m is related to the wedge angle βπ

$$\beta = \frac{2m}{1+m} \qquad m = \frac{\beta}{2-\beta}$$

With laminar boundary layers, the boundary layer thickness, friction factor, and Nusselt numbers are defined by

$$\frac{\delta}{x} = \frac{c_1}{\sqrt{\text{Re}_x}} \qquad \frac{C_{fx}}{2} = \frac{\tau_w}{\rho U_\infty^2} = \frac{c_2}{\sqrt{\text{Re}_x}} \qquad \text{Nu}_x = c_3 \text{Re}_x^{1/2}$$

The values of c_1, c_2, and c_3 are available in Burmeister (1993). For example, for β = 0.5 (wedge angle = 90°), m = 1/3, c_1 = 3.4, c_2 = 0.7575, and c_3 = 0.384 for Pr = 0.7, and c_3 = 0.792 for Pr = 5. Re_x is based on $U_\infty = cx^m$; the free-stream velocity is not uniform.

Uniform Temperature: Flat Plate with Injection or Suction with External Flows of a Fluid Parallel to the Surface

Injection or suction has engineering applications. When the free-stream temperature of the fluid is high, as in gas turbines, a cooling fluid is introduced into the mainstream to cool the surface. If the cooling fluid is introduced at discrete locations (either perpendicular to the surface or at an angle), it is known as film cooling. If a fluid is introduced or withdrawn through a porous medium, it is known as transpiration (Figure 4.2.13). An application of suction is to prevent boundary layer separation (Figure 4.2.13).

Analytical solutions for a laminar boundary layer with transpiration suction or blowing are available if the velocity perpendicular to the surface varies in the following manner:

$$\mathbf{v}_o = \text{constant } x^{(m-1)/2}$$

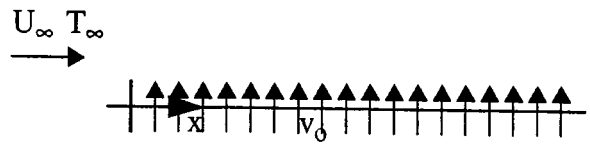


FIGURE 4.2.13 Flat plate with transpiration injection.

Solutions are limited to the cases of the injected fluid being at the same temperature as the surface and the injected fluid being the same as the free-stream fluid. Positive values of v_o indicate blowing and negative values indicate suction. Values of $\mathrm{Nu}_x/\mathrm{Re}_x^{1/2}$ for different values of Pr and for different values of blowing or suction parameter are given in Kays and Crawford (1993).

For example, for a laminar boundary layer over a flat plate with a fluid (Pr = 0.7) the value of $\mathrm{Nu}_x/\mathrm{Re}_x^{1/2}$ is 0.722 for $(v_o/U_\infty)\sqrt{\rho U_\infty x/\mu} = -0.75$ (suction) and 0.166 for $(v_o/U_\infty)\sqrt{\rho U_\infty x/\mu} = 0.25$ (blowing). Heat transfer coefficient increases with suction which leads to a thinning of the boundary layer. Blowing increases the boundary layer thickness and decreases the heat transfer coefficient.

For *turbulent boundary layers* Kays and Crawford (1993) suggest the following procedure for finding the friction factor and convective heat transfer coefficient. Define friction blowing parameter B_f and heat transfer blowing parameter B_h as

$$B_f = \frac{\mathrm{v}_o/U_\infty}{C_f/2} \tag{4.2.48}$$

$$B_h = \frac{\mathrm{v}_o/U_\infty}{\mathrm{St}} = \frac{\dot{m}''/G_\infty}{\mathrm{St}} \tag{4.2.49}$$

where

v_o = velocity normal to the plate
U_∞ = free-stream velocity
$\dot{m}''$ = mass flux of the injected fluid at the surface ($\rho \mathrm{v}_o$)
G_∞ = mass flux in the free stream (ρU_∞)
St = Stanton number = $\mathrm{Nu}_x/\mathrm{Re}_x\mathrm{Pr} = h/\rho U_\infty c_p$

The friction factors and Stanton number with and without blowing or suction are related by

$$\frac{C_f}{C_{fo}} = \frac{\ln(1+B_f)}{B_f} \tag{4.2.50}$$

$$\frac{\mathrm{St}}{\mathrm{St}_o} = \frac{\ln(1+B_h)}{B_h} \tag{4.2.51}$$

In Equations (4.2.50) and (4.2.51) C_{fo} and St_o are the friction factor and Stanton number with $\mathrm{v}_o = 0$ (no blowing or suction), and C_f and St are the corresponding quantities with blowing or suction at the same $\mathrm{Re}_x(\rho U_\infty x/\mu)$.

For the more general case of variable free-stream velocity, temperature difference, and transpiration rate, refer to Kays and Crawford (1993).

Flow over Flat Plate with Zero Pressure Gradient: Effect of High-Speed and Viscous Dissipation

In the boundary layer the velocity of the fluid is reduced from U_∞ to zero at the plate leading to a reduction in the kinetic energy of the fluid. Within the boundary layer there is also the work done by viscous forces; the magnitude of the such viscous work is related to the velocity of the fluid, the velocity gradient, and the viscosity of the fluid. The effect of such a reduction in the kinetic energy and the viscous work is to increase the internal energy of the fluid in the boundary layer. The increase in the internal energy may be expected to lead to an increase in the temperature; but because of the heat transfer to the adjacent fluid the actual increase in the internal energy (and the temperature) will be less than the sum of the decrease in the kinetic energy and viscous work transfer; the actual temperature

increase depends on the decrease in the kinetic energy, the viscous work transfer, and the heat transfer from the fluid. The maximum temperature in the fluid with an adiabatic plate is known as the adiabatic wall temperature (which occurs at the wall) and is given by

$$T_{aw} = T_\infty + r\frac{U_\infty^2}{2C_p} \tag{4.2.52}$$

In Equation (4.2.52) r is the recovery factor and is given by Eckert and Drake (1972).

Laminar boundary layer $0.6 < \text{Pr} < 15$ $r = \text{Pr}^{1/2}$

Turbulent boundary layer $r = \text{Pr}^{1/3}$

Equation (4.2.52) can be recast as

$$\frac{T_{aw} - T_\infty}{T_s - T_\infty} = \frac{r}{2}\frac{U_\infty^2}{C_p(T_s - T_\infty)} \tag{4.2.53}$$

From Equation (4.2.53) the maximum increase in the fluid temperature as a fraction of the difference between the plate and free-stream temperatures is given by r Ec/2. With air flowing over a plate at 500 m/sec, the increase in the temperature of the air can be as high as 105°C. With T_s = 40°C and T_∞ = 20°C, the temperature of the air close to the plate can be higher than the plate temperature. It is thus possible that although the plate temperature is higher than the free-stream temperature, the heat transfer is from the air to the plate. At a Mach number greater than 0.1 for gases, viscous dissipation becomes significant.

The temperature profiles for high-speed flows for different values of T_s are shown in Figure 4.2.14. In high-speed flows, as heat transfer can be to the plate even if the plate temperature is greater than the fluid temperature, the definition of the convective heat transfer coefficient given in Equation (4.2.29) is not adequate. On the other hand, as the heat transfer is always from the plate if $T_s > T_{aw}$, the adiabatic wall temperature is more appropriate as the reference temperature. Thus, in high-speed flows the definition of the convective heat transfer coefficient is given by

$$q'' = h(T_s - T_{aw}) \tag{4.2.54}$$

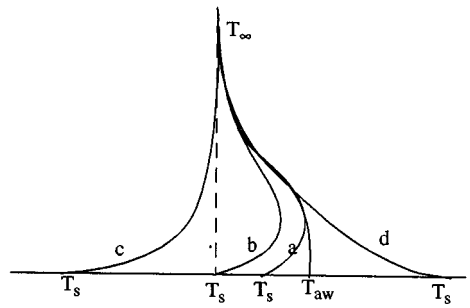


FIGURE 4.2.14 Temperature profiles for high-speed flows: (a) $T_\infty < T_s < T_{aw}$; (b) $T_s = T_\infty$; (c) $T_s \ll T_\infty$; (d) $T_s > T_{aw}$.

Equation (4.2.54) is consistent with Equation (4.2.29) as the adiabatic wall temperature equals the free-stream temperature if the effects of viscous dissipation and reduced kinetic energy in the boundary layer are neglected. With the adiabatic wall temperature as the fluid reference temperature for the definition

of the convective heat transfer coefficient, equations for low speeds can also be used for high-speed flows. Because of the greater variation in the fluid temperature in the boundary layer, the variation of properties due to temperature variation becomes important. It is found that the correlations are best approximated if the properties are evaluated at the reference temperature T^* defined by Eckert (1956):

$$T^* = 0.5\left(T_s + T_\infty\right) + 0.22\left(T_{aw} - T_\infty\right) \tag{4.2.55}$$

With properties evaluated at the reference temperature given by Equation (4.2.55), Equation (4.2.56) through (4.2.61) are applicable to high-speed flows with Prandtl numbers less than 15. It should be noted that the adiabatic wall temperatures in the laminar and turbulent regions are different affecting both the temperature at which the properties are evaluated and the temperature difference for determining the local heat flux. Therefore, when the boundary layer is partly laminar and partly turbulent, an average value of the heat transfer coefficient is not defined as the adiabatic wall temperatures in the two regions are different. In such cases the heat transfer rate in each region is determined separately to find the total heat transfer rate.

Evaluate properties at reference temperature given by Equation (4.2.55):

Laminar	Local: $\mathrm{Re}_x < \mathrm{Re}_{cr}$	$\mathrm{Nu}_x = 0.332\mathrm{Re}_x^{1/2}\mathrm{Pr}^{1/3}$	(4.2.56)
	Average: $\mathrm{Re}_L < \mathrm{Re}_{cr}$	$\mathrm{Nu}_L = 0.664\mathrm{Re}_L^{1/2}\mathrm{Pr}^{1/3}$	(4.2.57)
Turbulent	Local: $10^7 > \mathrm{Re}_x > \mathrm{Re}_{cr}$	$\mathrm{Nu}_x = 0.0296\mathrm{Re}_x^{4/5}\mathrm{Pr}^{1/3}$	(4.2.58)
	Local: $10^7 < \mathrm{Re}_x < 10^9$	$\mathrm{Nu}_x = 1.596\mathrm{Re}_x\left(\ln\mathrm{Re}_x\right)^{-2.584}\mathrm{Pr}^{1/3}$	(4.2.59)
	Average: $\mathrm{Re}_{cr} = 0,\ \mathrm{Re}_L < 10^7$	$\mathrm{Nu}_L = 0.037\mathrm{Re}_L^{4/5}\mathrm{Pr}^{1/3}$	(4.2.60)
	Average: $\mathrm{Re}_{cr} = 0,\ 10^7 < \mathrm{Re}_L < 10^9$	$\mathrm{Nu}_L = 1.967\mathrm{Re}_L\left(\ln\mathrm{Re}_L\right)^{-2.584}\mathrm{Pr}^{1/3}$	(4.2.61)

When the temperature variation in the boundary layer is large, such that the assumption of constant specific heat is not justified, Eckert (1956) suggests that the properties be evaluated at a reference temperature corresponding to the specific enthalpy i^* given by

$$i^* = 0.5\left(i_s + i_\infty\right) + 0.22\left(i_s - i_\infty\right) \tag{4.2.62}$$

where i is the specific enthalpy of the fluid evaluated at the temperature corresponding to the subscript. Equation (4.2.62) gives the same values as Equation (4.2.55) if C_p is constant or varies linearly with temperature.

At very high speeds the gas temperature may reach levels of temperatures that are sufficient to cause disassociation and chemical reaction; these and other effects need to be taken into account in those cases.

Flow over Cylinders, Spheres, and Other Geometries

Flows over a flat plate and wedges were classified as laminar or turbulent, depending on the Reynolds number, and correlations for the local and average convective heat transfer coefficients were developed. But flows over cylinders (perpendicular to the axis) and spheres are more complex. In general, the flow over cylinders and spheres may have a laminar boundary layer followed by a turbulent boundary layer

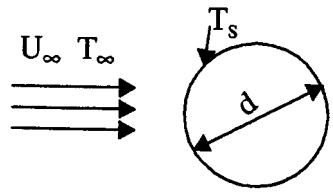


FIGURE 4.2.15 A fluid stream in cross flow over a cylinder.

and a wake region depending on the Reynolds number with the diameter as the characteristic length. Because of the complexity of the flow patterns, only correlations for the average heat transfer coefficients have been developed (Figure 4.2.15).

Cylinders: Use the following correlation proposed by Churchill and Bernstein (1977): $\mathrm{Re}_d\,\mathrm{Pr} > 0.2$. Evaluate properties at $(T_s + T_\infty)/2$:

$$\mathrm{Re}_d > 400{,}000: \qquad \mathrm{Nu}_d = 0.3 + \frac{0.62\mathrm{Re}_d^{1/2}\mathrm{Pr}^{1/3}}{\left[1+(0.4/\mathrm{Pr})^{2/3}\right]^{1/4}}\left[1+\left(\frac{\mathrm{Re}_d}{282{,}000}\right)^{5/8}\right]^{4/5} \tag{4.2.63}$$

$$10{,}000 < \mathrm{Re}_d < 400{,}000: \quad \mathrm{Nu}_d = 0.3 + \frac{0.62\mathrm{Re}_d^{1/2}\mathrm{Pr}^{1/3}}{\left[1+(0.4/\mathrm{Pr})^{2/3}\right]^{1/4}}\left[1+\left(\frac{\mathrm{Re}_d}{282{,}000}\right)^{1/2}\right] \tag{4.2.64}$$

$$\mathrm{Re}_d < 10{,}000: \qquad \mathrm{Nu}_d = 0.3 + \frac{0.62\mathrm{Re}_d^{1/2}\mathrm{Pr}^{1/3}}{\left[1+(0.4/\mathrm{Pr})^{2/3}\right]^{1/4}} \tag{4.2.65}$$

For flow of liquid metals, use the following correlation suggested by Ishiguro et al. (1979):

$$1 < \mathrm{Re}_d\mathrm{Pr} < 100 \qquad \mathrm{Nu}_d = 1.125\left(\mathrm{Re}_d\mathrm{Pr}\right)^{0.413} \tag{4.2.66}$$

For more information on heat transfer with flow over cylinders, refer to Morgan (1975) and Zukauskas (1987).

Spheres: For flows over spheres (Figure 4.2.16) use one of the following two correlations.

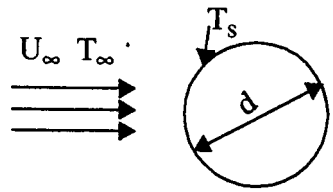


FIGURE 4.2.16 A fluid flowing over a sphere.

1. Whitaker (1972): Evaluate properties at T_∞ except μ_s at T_s.

$$3.5 < \mathrm{Re}_d < 76{,}000 \qquad 0.71 < \mathrm{Pr} < 380 \qquad 1 < \mu/\mu_s < 3.2$$

$$\mathrm{Nu}_d = 2.0 + \left(0.4\mathrm{Re}_d^{1/2} + 0.06\mathrm{Re}_d^{2/3}\right)\mathrm{Pr}^{2/5}\left(\frac{\mu}{\mu_s}\right)^{1/4} \tag{4.2.67}$$

2. Achenbach (1978): Evaluate properties at $(T_s + T_\infty)/2$:

$$100 < \mathrm{Re}_d < 2\times10^5 \qquad \mathrm{Pr} = 0.71$$

$$\mathrm{Nu}_d = 2 + \left(0.25\mathrm{Re}_d + 3\times10^{-4}\mathrm{Re}_d^{1.6}\right)^{1/2} \tag{4.2.68}$$

$$4\times10^5 < \mathrm{Re}_d < 5\times10^6 \qquad \mathrm{Pr} = 0.71$$

$$\mathrm{Nu}_d = 430 + 5\times10^{-3}\mathrm{Re}_d + 0.25\times10^{-9}\mathrm{Re}_d^2 - 3.1\times10^{-17}\mathrm{Re}_d^3 \tag{4.2.69}$$

3. Liquid Metals: From experimental results with liquid sodium, Witte (1968) proposed

$$3.6\times10^4 < \mathrm{Re}_d < 1.5\times10^5 \qquad \mathrm{Nu}_d = 2 + 0.386\left(\mathrm{Re}_d\mathrm{Pr}\right)^{1/2} \tag{4.2.70}$$

Other Geometries: For geometries other than cylinders and spheres, use Equation (4.2.71) with the characteristic dimensions and values of the constants given in the Table 4.2.1.

$$\mathrm{Nu}_D = c\mathrm{Re}_D^m \tag{4.2.71}$$

Although Equation (4.2.71) is based on experimental data with gases, its use can be extended to fluids with moderate Prandtl numbers by multiplying Equation (4.2.71) by $(\mathrm{Pr}/0.7)^{1/3}$.

Heat Transfer across Tube Banks

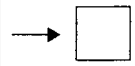
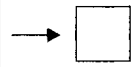
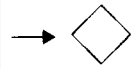
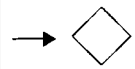
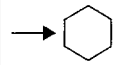
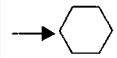
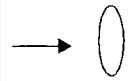
When tube banks are used in heat exchangers, the flow over the tubes in the second subsequent rows of tubes is different from the flow over a single tube. Even in the first row the flow is modified by the presence of the neighboring tubes. The extent of modification depends on the spacing between the tubes. If the spacing is very much greater than the diameter of the tubes, correlations for single tubes can be used. Correlations for flow over tube banks when the spacing between tubes in a row and a column is not much greater than the diameter of the tubes have been developed for use in heat-exchanger applications. Two arrangements of the tubes are considered — aligned and staggered as shown in Figure 4.2.17. The nomenclature used in this section is shown in the figure.

For the average convective heat transfer coefficient with tubes at uniform surface temperature, from experimental results, Zukauskas (1987) recommends correlations of the form:

$$\mathrm{Nu}_d = c\left(\frac{a}{b}\right)^p \mathrm{Re}_d^m \mathrm{Pr}^n \left(\frac{\mathrm{Pr}}{\mathrm{Pr}_s}\right)^{0.25} \tag{4.2.72}$$

In Equation (4.2.72) all properties are evaluated at the arithmetic mean of the inlet and exit temperatures of the fluid, except Pr_s which is evaluated at the surface temperature T_s. The values of the constants c, p, m, and n are given in Table 4.2.2 for in-line arrangement and in Table 4.2.3 for staggered arrangement.

TABLE 4.2.1 Values of *c* and *m* in Equation (4.2.71)

Geometry	Re_D	c	m
	5000-100 000	0.092	0.675
	2500-8000	0.160	0.699
	5000-100 000	0.222	0.588
	2500-7500	0.261	0.624
	5000-19500 19 500-100 000	0.144 0.035	0.638 0.782
	5000-100 000	0.138	0.638
	2500-15 000	0.224	0.612
	3000-15 000	0.085	0.804
	4000-15 000	0.205	0.731
Charcteristic dimension is the equivalent circular diameter = Perimeter/π For example, for a square rod with each side a, D = 4a/π			

From Jakob, 1949. With permission.

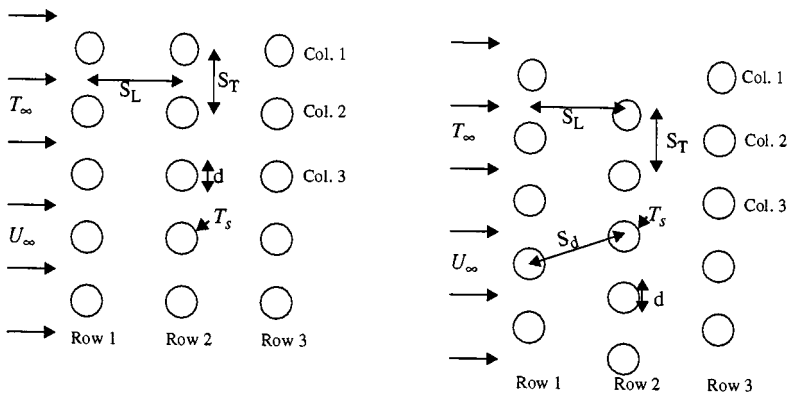


FIGURE 4.2.17 Two arrangements of tube banks. In-line or aligned arrangement on the left and staggered arrangement on the right. ($a = S_T/d$; $b = S_L/d$.)

TABLE 4.2.2 In-Line Arrangement — Values of Constants in Equation (4.2.72) (p = 0 in all cases)

Re_d	c	m	n
1–100	0.9	0.4	0.36
100–1000	0.52	0.5	0.36
10^3–2×10^5	0.27	0.63	0.36
$2 \times 10^5 - 2 \times 10^6$	0.033	0.8	0.4

TABLE 4.2.3 Staggered Arrangement — Values of Constants in Equation (4.2.72)

Re_d	c	p	m	n
1–500	1.04	0	0.4	0.36
500–1000	0.71	0	0.5	0.36
10^3–2×10^5	0.35	0.2	0.6	0.36
$2 \times 10^5 - 2 \times 10^6$	0.031	0.2	0.8	0.36

In computing Re_d, the maximum average velocity between tubes is used. The maximum velocities for the in-line and staggered arrangements are given by

$$\text{In-line:} \qquad U_{\max} = \frac{U_\infty S_T}{S_T - d} \tag{4.2.73}$$

$$\text{Staggered:} \qquad S_d > \frac{S_T + d}{2} \qquad U_{\max} = \frac{U_\infty S_T}{S_T - d} \tag{4.2.74}$$

$$\text{Staggered:} \qquad S_d < \frac{S_T + d}{2} \qquad U_{\max} = \frac{U_\infty S_T}{2\left(S_d - d\right)} \tag{4.2.75}$$

$$S_d = \left[S_L^2 + \left(\frac{S_T}{2} \right)^2 \right]^{1/2}$$

Equation (4.2.72) is for tube banks with 16 or more rows. When there are fewer than 16 rows, the heat transfer coefficient given by Equation (4.2.72) is multiplied by the correction factor c_1 defined by Equation (4.2.76) and given in Table 4.2.4.

$$\frac{h_N}{h_{16}} = c_1 \tag{4.2.76}$$

where

h_N = heat transfer coefficient with N rows (fewer than 16)
h_{16} = heat transfer coefficient with 16 or more rows

TABLE 4.2.4 Correction Factor c_1 to Be Used with Equation (4.2.76)

	Number of Rows (N)							
Tube Arrangement	**1**	**2**	**3**	**4**	**5**	**7**	**10**	**13**
In-line	0.70	0.80	0.86	0.90	0.93	0.96	0.98	0.99
Staggered	0.64	0.76	0.84	0.89	0.93	0.96	0.98	0.99

Pressure Drop: With tube banks, pressure drop is a significant factor, as it determines the fan power required to maintain the fluid flow. Zukauskas (1987) recommends that the pressure drop be computed from the relation

$$\Delta p = p_i - p_e = N\chi \frac{\rho U_{\max}^2}{2} f \tag{4.2.77}$$

where p_i and p_e are the fluid pressures at inlet and exit of the tube banks. The values of χ and f are presented in Figure 4.2.18A. In Figure 4.2.18A the friction factor f for in-line arrangement is presented for different values of b (S_L/d) for $S_L = S_T$. For values of S_L/S_T other than 1, the correction factor χ is given in the inset for different values of $(a - 1)/(b - 1)$. Similarly, the friction factor for staggered arrangement (for equilateral triangle arrangement) and a correction factor for different values of a/b are also given in Figure 4.2.18b. The value of f is for one row of tubes; the total pressure drop is obtained by multiplying the pressure drop for one row by the number of rows, N.

The temperature of the fluid varies in the direction of flow, and, therefore, the value of the convective heat transfer coefficient (which depends on the temperature-dependent properties of the fluid) also varies in the direction of flow. However, it is common practice to compute the total heat transfer rate with the assumption of uniform convective heat transfer coefficient evaluated at the arithmetic mean of the inlet and exit temperatures of the fluid. With such an assumption of uniform convective heat transfer coefficient, uniform surface temperature and constant specific heat (evaluated at the mean fluid temperature), the inlet and exit fluid temperatures are related by

$$\ln\left(\frac{T_s - T_e}{T_s - T_i}\right) = -\frac{hA_s}{\dot{m}c_p} \tag{4.2.78}$$

The heat transfer rate to the fluid is related by the equation

$$q = \dot{m}c_p\left(T_i - T_e\right) \tag{4.2.79}$$

Example

A heat exchanger with aligned tubes is used to heat 40 kg/sec of atmospheric air from 10 to 50°C with the tube surfaces maintained at 100°C. Details of the heat exchanger are

Diameter of tubes	25 mm
Number of columns	20
Length of each tube	3 m
$S_L = S_T$	75 mm

Determine the number of rows required.

Solution: Average air temperature = $(T_i + T_e)/2$ = 30°C. Properties of atmospheric air (from Suryanarayana, 1995):

$\rho = 1.165$ kg/m^3	$c_p = 1007$ J/kg K
$\mu = 1.865 \times 10^{-5}$ N sec/m^3	$k = 0.0264$ W/mK
Pr = 0.712	Pr_s(at 100°C) = 0.705

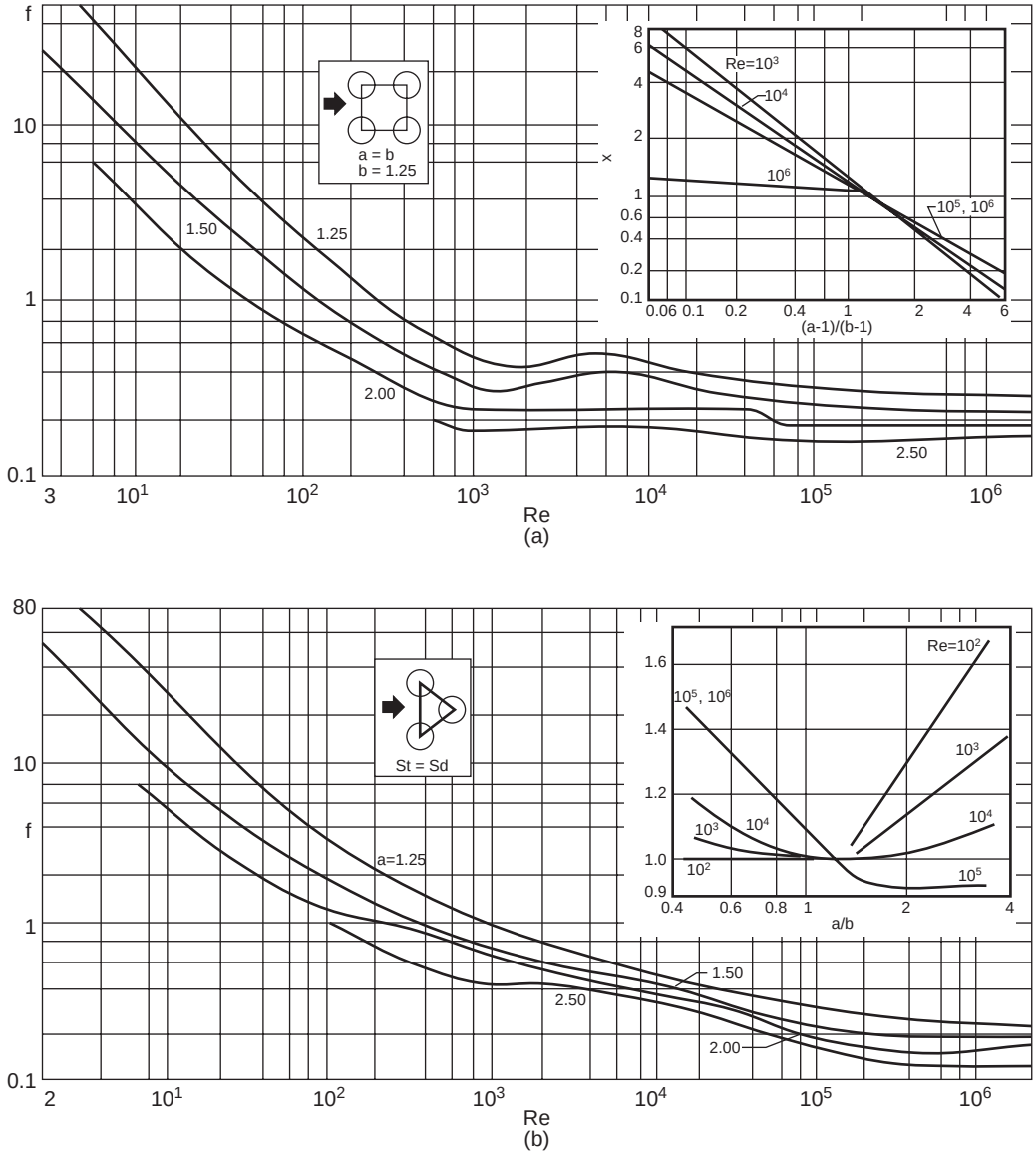


FIGURE 4.2.18 Friction factors for tube banks. (a) In-line arrangement; (b) Staggered arrangement.

To find U_{max} we need the minimum area of cross section for fluid flow (Figure 4.2.19).

$$H = 20 \times 0.075 = 1.5 \text{ m}$$

$$A_{min} = 20(0.075 - 0.025) \times 3 = 3 \text{ m}^2$$

$$U_{max} = \frac{\dot{m}}{\rho A_{min}} = \frac{40}{1.165 \times 3} = 11.44 \text{ m/sec}$$

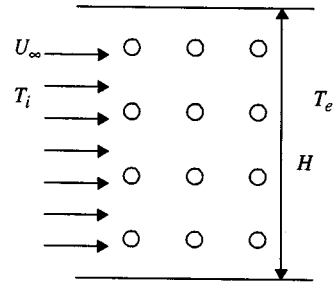


FIGURE 4.2.19 Aligned tube heat exchanger (only a few of the 20 columns and rows are shown).

$$\mathrm{Re}_d = \frac{\rho U_{\max} d}{\mu} = \frac{1.165 \times 11.44 \times 0.025}{1.865 \times 10^{-5}} = 17{,}865$$

With values from Table 4.2.2,

$$\mathrm{Nu}_d = 0.27 \times 17{,}865^{0.63} \times 0.712^{0.36} \left(\frac{0.712}{0.705}\right)^{0.25} = 114.3$$

$$h = \frac{114.3 \times 0.0264}{0.025} = 120.7 \ \mathrm{W/m^2\ K}$$

From Equation 4.2.78,

$$\ln\left(\frac{100-50}{100-10}\right) = -\frac{120.7 \times A_s}{40 \times 1007} \qquad A_s = \pi \times 0.025 \times 3 \times 20 \times N$$

N = number of rows = 42

Fan Power: From the first law of thermodynamics (see Chapter 2), the fan power is

$$\dot{W}_F = \dot{m}\left(\frac{p_i}{\rho_i} + \frac{p_e}{\rho_e} + \frac{\mathbf{v}_e^2}{2}\right)$$

p_i and p_e are the pressures at inlet and exit of the heat exchanger and $\mathbf{v}_e$ is the fluid velocity at exit. Assuming constant density evaluated at $(T_i + T_e)/2$ the pressure drop is found from Figure 4.2.18a.

$\mathrm{Re}_p = 17{,}865$: $\qquad a = b = S_T/d = 75/25 = 3$

In Figure 4.2.18, although the friction factor is available for values of b up to 2.5, we will estimate the value of f for $b = 3$. From Figure 4.2.18, $f \approx 0.11$. The correction factor $c = 1$.

$$p_i - p_e = N\chi \frac{\rho U_{\max}^2}{2} f = 42 \times 1 \frac{1.165 \times 11.44^2}{2} \times 0.11 = 352.2 \ \mathrm{kPa}$$

$$\mathbf{v}_e = \frac{11.44 \times 50}{75} = 7.63 \ \mathrm{m/sec}$$

$$\dot{W}_F = 40\left(352.2 + \frac{7.63^2}{2}\right) = \underline{15{,}250\ \text{W}}$$

Heat Transfer with Jet Impingement

Jet impingement (Figure 4.2.20) on a heated (or cooled) surface results in high heat transfer rates, and is used in annealing of metals, tempering of glass, cooling of electronic equipment, internal combustion engines, and in a wide variety of industries — textiles, paper, wood, and so on. Usually, the jets are circular, issuing from a round nozzle of diameter *d*, or rectangular, issuing from a slot of width *w*. They may be used singly or in an array. The jets may impinge normally to the heated surface or at an angle. If there is no parallel solid surface close to the heated surface, the jet is said to be free; in the presence of a parallel surface close to the heated surface, the jet is termed confined. In this section only single, free jets (round or rectangular) impinging normally to the heated surface are considered.

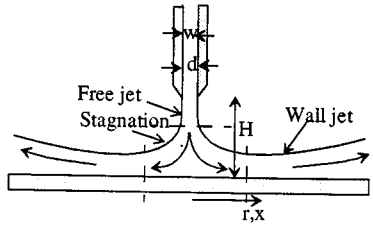


FIGURE 4.2.20 Circular jet of diameter *d* or a rectangular jet of width *w*.

Jets may be submerged with the fluid from the nozzle exiting into a body of a fluid (usually the same fluid), for example, air impinging on a surface surrounded by atmospheric air. In submerged jets entrained fluid (the part of the surrounding fluid dragged by the jet) has a significant effect on the flow and heat transfer characteristics of the jet, but the effect of gravity is usually negligible. In free-surface jets — a liquid jet in an atmosphere of air is a good approximation to a free-surface jet — the entrainment effect is usually negligible, but the effect of gravity may be significant.

A jet is usually divided into three regions, a free-jet region, a stagnation region, and a wall-jet region. In the free-jet region the effect of the target surface on the flow is negligible. In the stagnation region the target surface affects the flow field, and the velocity parallel to the surface increases while the velocity component normal to the surface decreases. At the beginning of the stagnation region, the axial velocity of the fluid is very much greater than the radial component (or the *x*-component) of the velocity. The stagnation region is followed by the wall-jet region where the radial component (or the *x*-component) of the velocity is much greater than the axial velocity.

The heat transfer coefficient is a function of *H/d* (or *H/w*), $\text{Re}_d(\rho v_j d/\mu)$ or $(\rho v_j 2w/\mu)$, and Pr and depends on the region (stagnation or wall jet), whether it is submerged or nonsubmerged and whether the flow adjacent to the plate is laminar or turbulent. Some of the heat transfer correlations suggested by different researchers are given below. All the correlations are for single jets.

Submerged Jets: Single Circular Jets

$$\text{Re}_d = \frac{4\dot{m}}{\pi d\mu} \qquad \text{Nu}_d = \frac{hd}{k} \qquad \dot{m} = \text{mass rate of flow of fluid}$$

Average heat transfer coefficients up to radius r (Martin, 1990):

$$\mathrm{Nu}_d = 2\frac{d}{r}\frac{1-1.1d/r}{1+0.1(H/d-6)d/r}\left[\mathrm{Re}_d\left(1+\frac{\mathrm{Re}_d^{0.55}}{200}\right)\right]^{1/2}\mathrm{Pr}^{0.42} \tag{4.2.80}$$

Range of validity:

$$2000 \le \mathrm{Re}_d \le 400{,}000 \qquad 2.5 \le r/d \le 7.5 \qquad 2 \le H/d \le 12$$

Local convective heat transfer coefficient at radius r (Webb and Ma, 1995):

$$\mathrm{Nu}_d = 1.29\mathrm{Re}_d^{1/2}\mathrm{Pr}^{0.4}\left\{\left[\frac{\tanh(0.88r/d)}{r/d}\right]^{-8.5}+\left[1.69\left(\frac{r}{d}\right)^{-1.07}\right]^{-17}\right\}^{-1/17} \tag{4.2.81}$$

Submerged Jets: Single Rectangular Jet

$$\mathrm{Re}_w = \frac{\rho \mathbf{v}_j 2w}{\mu} = \frac{2\dot{m}}{\mu} \qquad \dot{m} = \text{mass rate of flow per unit length of jet}$$

$$\mathrm{Nu}_w = \frac{h2w}{k}$$

Average heat transfer coefficient (Martin, 1990):

$$\mathrm{Nu}_w = \frac{1.53\mathrm{Pr}^{0.42}\mathrm{Re}_w^m}{\frac{x}{2w}+\frac{H}{2w}+1.39} \tag{4.2.82}$$

$$m = 0.695-\left[\frac{x}{2w}+\left(\frac{H}{2w}\right)^{1.33}+3.06\right]^{-1}$$

Free-Surface Jets: Single Circular Jet. Correlations are given in Table 4.2.5 (Liu et al., 1991 and Webb and Ma, 1995).

For more information on jet impingement heat transfer, refer to Martin (1977) and Webb and Ma (1995) and the references in the two papers.

Bibliography

ASHRAE *Handbook of Fundamentals*, 1993. American Society of Heating, Ventilating and Air Conditioning Engineers, Atlanta, GA.

Hewitt, G.F., Ed. 1990. *Handbook of Heat Exchanger Design*, Hemisphere Publishing, New York.

Incropera, F.P. and Dewitt, D.P. 1990. *Fundamentals of Heat and Mass Transfer*, 3rd ed., John Wiley & Sons, New York.

Kakaç, S., Shah, R.K., and Win Aung, Eds. 1987. *Handbook of Single Phase Convective Heat Transfer*, Wiley-Interscience, New York.

Kreith, F. and Bohn, M.S. 1993. *Principles of Heat Transfer*, 5th ed., PWS, Boston.

Suryanarayana, N.V. 1995. *Engineering Heat Transfer*, PWS, Boston.

TABLE 4.2.5 Correlations for Free-Surface Jets $r_v/d = 0.1773\ \mathrm{Re}_d^{1/3}$

		Nu_d	
$r/d < 0.787$	$0.15 \le \mathrm{Pr} \le 3$	$0.715\mathrm{Re}_d^{1/2}\mathrm{Pr}^{0.4}$	(4.2.83)
	$\mathrm{Pr} > 3$	$0.797\mathrm{Re}_d^{1/2}\mathrm{Pr}^{1/3}$	(4.2.84)
$0.787 < r/d < r_v/d$		$0.632\mathrm{Re}_d^{1/2}\mathrm{Pr}^{1/3}\left(\frac{d}{r}\right)^{1/2}$	(4.2.85)
$r_v/d < r/d < r_t/d$		$\dfrac{0.407\mathrm{Re}_d^{1/3}\mathrm{Pr}^{1/3}(d/r)^{2/3}}{\left[\dfrac{0.1713}{(r/d)^2}+\dfrac{5.147}{\mathrm{Re}_d}\dfrac{r}{d}\right]^{2/3}\left[\dfrac{(r/d)^2}{2}+C\right]^{1/3}}$ where $C=-5.051\times10^{-5}\mathrm{Re}_d^{2/3}$	(4.2.86)
$r > r_t$	$\mathrm{Pr} < 4.86$	$\dfrac{0.25}{\dfrac{1}{\mathrm{Re}_d\mathrm{Pr}}\left[1-\left(\dfrac{r_t}{r}\right)^2\right]\left(\dfrac{r}{d}\right)^2+0.13\dfrac{h}{d}+0.0371\dfrac{h_t}{d}}$ where $h_t = h$ at r_t and $\dfrac{h}{d}=\dfrac{0.1713}{r/d}+\dfrac{5.147}{\mathrm{Re}_d}\left(\dfrac{r}{d}\right)^2$	(4.2.87)

$$\frac{r_t}{d}=\left\{-\frac{s}{2}+\left[\left(\frac{s}{2}\right)^2+\left(\frac{p}{3}\right)^3\right]^{1/2}\right\}^{1/3}+\left\{-\frac{s}{2}+\left[\left(\frac{s}{2}\right)^2-\left(\frac{p}{3}\right)^3\right]^{1/2}\right\}^{1/3}$$

$$p=\frac{-2C}{0.2058\mathrm{Pr}-1}\qquad s=\frac{0.00686\mathrm{Re}_d\mathrm{Pr}}{0.2058\mathrm{Pr}-1}$$

References

Achenbach, E. 1978. *Heat Transfer from Spheres up to Re = 6 × 10⁶*, in *Proc. 6th Int. Heat Transfer Conf.*, Vol. 5, Hemisphere Publishing, Washington, D.C.

Burmeister, L.C. 1993. *Convective Heat Transfer*, Wiley-Interscience, New York.

Churchill, S.W. 1976. A comprehensive correlation equation for forced convection from a flat plate, *AIChE J.* 22(2), 264.

Churchill, S.W. and Bernstein, M. 1977. A correlating equation for forced convection from gases and liquids to a circular cylinder in cross flow, *J. Heat Transfer*, 99, 300.

Churchill, S.W. and Ozoe, H. 1973. Correlations for laminar forced convection with uniform heating in flow over a plate and in developing and fully developed flow in a tube, *J. Heat Transfer*, 18, 78.

Eckert, E.R.G. 1956. Engineering relations for heat transfer and friction in high-velocity laminar and turbulent boundary-layer flow over surfaces with constant pressure and temperature, *Trans. ASME*, 56, 1273.

Eckert, E.R.G. and Drake, M., Jr. 1972. *Analysis of Heat and Mass Transfer*, McGraw-Hill, New York.

Ishiguro, R., Sugiyama, K., and Kumada, T. 1979. Heat transfer around a circular cylinder in a liquid-sodium cross flow, *Int. J. Heat Mass Transfer*, 22, 1041.

Jakob, H., 1949. *Heat Transfer*, John Wiley and Sons, London.

Kays, W.M. and Crawford, M.E. 1993. *Convective Heat and Mass Transfer*, 3rd ed., McGraw-Hill, New York.

Liu, X., Lienhard, v., J.H., and Lombara, J.S. 1991. Convective heat transfer by impingement of circular liquid jets, *J. Heat Transfer*, 113, 571.

Martin, H. 1977. Heat and mass transfer between impinging gas jets and solid surfaces, in *Advances in Heat Transfer*, Hartnett, J.P. and Irvine, T.F., Eds., 13, 1, Academic Press, New York.

Martin, H. 1990. Impinging jets, in *Handbook of Heat Exchanger Design*, Hewitt, G.F., Ed., Hemisphere, New York.

Morgan, Vincent T., 1975. The overall convective heat transfer from smooth circular cylinders, in *Advances in Heat Transfer*, Irvine, T.F. and Hartnett, J.P., Eds., 11, 199, Academic Press, New York.

Rose, J.W. 1979. Boundary layer flow on a flat plate, *Int. J. Heat Mass Transfer*, 22, 969.

Schlichting, H. 1979. *Boundary Layer Theory*, 7th ed., McGraw-Hill, New York.

Suryanarayana, N.V. 1995. *Engineering Heat Transfer*, West Publishing, Minneapolis.

Thomas, W.C. 1977. Note on the heat transfer equation for forced-convection flow over a flat plate with an unheated starting length, *Mech. Eng. News* (*ASEE*), 9(1), 19.

Webb, B.W. and Ma, C.F. 1995. Single-phase liquid jet impingement heat transfer, in *Advances in Heat Transfer*, Hartnett, J.P. and Irvine, T.F., Eds., 26, 105, Academic Press, New York.

Witte, L.C. 1968. An experimental study of forced-convection heat transfer from a sphere to liquid sodium, *J. Heat Transfer*, 90, 9.

Zukauskas, A. 1987. Convective heat transfer in cross flow, in *Handbook of Single-Phase Convective Heat Transfer*, Kakaç, S., Shah, R.K., and Win Aung, Eds., Wiley-Interscience, New York.

Forced Convection — Internal Flows

N.V. Suryanarayana

Introduction

Heat transfer to (or from) a fluid flowing inside a tube or duct is termed *internal forced convection*. The fluid flow may be laminar or turbulent. If the Reynolds number based on the average velocity of the fluid and diameter of the tube ($\rho vd/\mu$) is less than 2100 (Reynolds numbers in the range of 2000 to 2300 are cited in different sources), the flow is laminar. If the Reynolds number is greater than 10,000, the flow is turbulent. The flow with a Reynolds number in the range 2100 to 10,000 is considered to be in the transitional regime. With heating or cooling of the fluid, there may or may not be a change in the phase of the fluid. Here, only heat transfer to or from a single-phase fluid is considered.

Fully Developed Velocity and Temperature Profiles. When a fluid enters a tube from a large reservoir, the velocity profile at the entrance is almost uniform as shown in Figure 4.2.21. The fluid in the immediate vicinity of the tube surface is decelerated and the velocity increases from zero at the surface to u_c at a distance δ from the surface; in the region $r = 0$ to $(R - \delta)$ the velocity is uniform. The value of δ increases in the direction of flow and with constant fluid density the value of the uniform velocity u_c increases. At some location downstream, δ reaches its maximum possible value, equal to the radius of the tube, and from that point onward the velocity profile does not change.

The region where δ increases, i.e., where the velocity profile changes, is known as the entrance region or hydrodynamically developing region. The region downstream from the axial location where δ reaches its maximum value and where the velocity profile does not change is the fully developed velocity profile or hydrodynamically fully developed region. Similarly, downstream of the location where heating or cooling of the fluid starts, the temperature profile changes in the direction of flow. But beyond a certain distance the dimensionless temperature profile does not change in the direction of flow. The region where the dimensionless temperature profile changes is the thermally developing region or the thermal entrance region, and the region where the dimensionless temperature profile does not change is the thermally

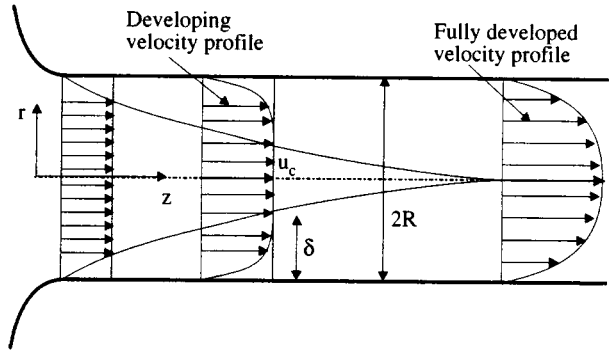


FIGURE 4.2.21 Developing and fully developed velocity profiles.

fully developed region. For simultaneously developing velocity and temperature profiles in laminar flows, the hydrodynamic and thermal entrance lengths are given by

$$\frac{L_e}{d} = 0.0565\mathrm{Re}_d \tag{4.2.88}$$

$$\frac{L_{e,th}}{d} = 0.053\mathrm{Re}_d\mathrm{Pr} \qquad \text{Uniform heat flux} \tag{4.2.89}$$

$$\frac{L_{e,th}}{d} = 0.037\mathrm{RePr} \qquad \text{Uniform surface temperature} \tag{4.2.90}$$

In most engineering applications, with turbulent flows, correlations for fully developed conditions can be used after about 10 diameters from where the heating starts.

Convective Heat Transfer Coefficient and Bulk Temperature. The reference temperature for defining the convective heat transfer coefficient is the bulk temperature T_b and the convective heat flux is given by

$$q'' = h\left(T_s - T_b\right) \tag{4.2.91}$$

The bulk temperature T_b is determined from the relation

$$T_b = \frac{\int_{A_c} \rho v C_p T dA_c}{\int_{A_c} \rho v C_p dA_c} \tag{4.2.92}$$

where A_c is the cross-sectional area perpendicular to the axis of the tube.

If the fluid is drained from the tube at a particular axial location and mixed, the temperature of the mixed fluid is the bulk temperature. It is also know as the mixing cup temperature. With heating or cooling of the fluid the bulk temperature varies in the direction of flow. In some cases we use the term *mean fluid temperature*, T_m, to represent the arithmetic mean of the fluid bulk temperatures at inlet and exit of the tube.

Heat Transfer Correlations

Laminar Flows — Entrance Region. For laminar flows in a tube with uniform surface temperature, in the entrance region the correlation of Sieder and Tate (1936) is

$$\overline{\mathrm{Nu}}_d = 1.86\left(\frac{\mathrm{Re}_d \mathrm{Pr}}{L/d}\right)^{1/3}\left(\frac{\mu}{\mu_s}\right)^{0.14} \tag{4.2.93}$$

valid for

$$\frac{L}{d} < \frac{\mathrm{Re}_d \mathrm{Pr}}{8}\left(\frac{\mu}{\mu_s}\right)^{0.42} \qquad 0.48 < \mathrm{Pr} < 16,700 \qquad 0.0044 < \frac{\mu}{\mu_s} < 9.75$$

The overbar in the Nusselt number indicates that it is formed with the average heat transfer coefficient over the entire length of the tube. Properties of the fluid are evaluated at the arithmetic mean of the inlet and exit bulk temperatures. In Equation (4.2.93) the heat transfer coefficient was determined from

$$q = \bar{h}\pi dL\left(T_s - \frac{T_{bi} + T_{be}}{2}\right) \tag{4.2.94}$$

Therefore, to find the total heat transfer rate with $\bar{h}$ from Equation (4.2.93) employ Equation (4.2.94).

Laminar Flows — Fully Developed Velocity and Temperature Profiles. Evaluate properties at the bulk temperature

Uniform Surface Temperature $\mathrm{Nu}_d = 3.66$ (4.2.95)

Uniform Surface Heat Flux $\mathrm{Nu}_d = 4.36$ (4.2.96)

Turbulent Flows. If the flow is turbulent, the difference between the correlations with uniform surface temperature and uniform surface heat flux is not significant and the correlations can be used for both cases. For turbulent flows, Gnielinsky (1976, 1990) recommends:
Evaluate properties at the bulk temperature.

$0.6 < \mathrm{Pr} < 2000 \qquad 2300 < \mathrm{Re}_d < 10^6 \qquad 0 < d/L < 1$

$$\mathrm{Nu}_d = \frac{(f/2)(\mathrm{Re}_d - 1000)\mathrm{Pr}}{1 + 12.7(f/2)^{1/2}(\mathrm{Pr}^{2/3} - 1)}\left[1 + \left(\frac{d}{L}\right)^{2/3}\right] \tag{4.2.97}$$

$$f = \left[1.58\ln(\mathrm{Re}_d) - 3.28\right]^{-2} \tag{4.2.98}$$

f = friction factor = $2\tau_w/\rho v^2$.

To reflect the effect of variation of fluid properties with temperature, multiply the Nusselt numbers in Equation (4.2.97) by $(T_b/T_s)^{0.45}$ for gases and $(\mathrm{Pr}/\mathrm{Pr}_s)^{0.11}$ for liquids where the temperatures are absolute, and T and Pr with a subscript s are to be evaluated at the surface temperature. The equations can be used to evaluate the heat transfer coefficient in the developing profile region. To determine the heat

transfer coefficient in the fully developed region set $d/L = 0$. A simpler correlation (fully developed region) is the Dittus–Boelter (1930) equation. Evaluate properties at T_b.

$$0.7 \le \mathrm{Pr} \le 160 \qquad \mathrm{Re}_d > 10{,}000 \qquad d/L > 10$$

$$\mathrm{Nu}_d = 0.023\mathrm{Re}_d^{4/5}\mathrm{Pr}^n \tag{4.2.99}$$

where $n = 0.4$ for heating ($T_s > T_b$) and $n = 0.3$ for cooling ($T_s < T_b$).

For liquid metals with $\mathrm{Pr} \ll 1$ the correlations due to Sleicher and Rouse (1976) are
Uniform surface temperature:

$$\mathrm{Nu}_{d,b} = 4.8 + 0.0156\mathrm{Re}_{d,f}^{0.85}\mathrm{Pr}_s^{0.93} \tag{4.2.100}$$

Uniform heat flux:

$$\mathrm{Nu}_{d,b} = 6.3 + 0.0167\mathrm{Re}_{d,f}^{0.85}\mathrm{Pr}_s^{0.93} \tag{4.2.101}$$

Subscripts *b*, *f*, and *s* indicate that the variables are to be evaluated at the bulk temperature, film temperature (arithmetic mean of the bulk and surface temperatures), and surface temperature, respectively.

In the computations of the Nusselt number the properties (evaluated at the bulk temperature) vary in the direction of flow and hence give different values of *h* at different locations. In many cases a representative average value of the convective heat transfer coefficient is needed. Such an average value can be obtained either by taking the arithmetic average of the convective heat transfer coefficients evaluated at the inlet and exit bulk temperatures or the convective heat transfer coefficient evaluated at the arithmetic mean of the inlet and exit bulk temperatures. If the variation of the convective heat transfer coefficient is large, it may be appropriate to divide the tube into shorter lengths with smaller variation in the bulk temperatures and evaluating the average heat transfer coefficient in each section.

Uniform Surface Temperature — Relation between the Convective Heat Transfer Coefficient and the Total Heat Transfer Rate: With a uniform surface temperature, employing an average value of the convective heat transfer coefficient the local convective heat flux varies in the direction of flow. To relate the convective heat transfer coefficient to the temperatures and the surface area, we have, for the elemental length Δz (Figure 4.2.22).

$$\dot{m}C_p \frac{dT_b}{dz} = h\frac{dA_s}{dz}\left(T_s - T_b\right) \tag{4.2.102}$$

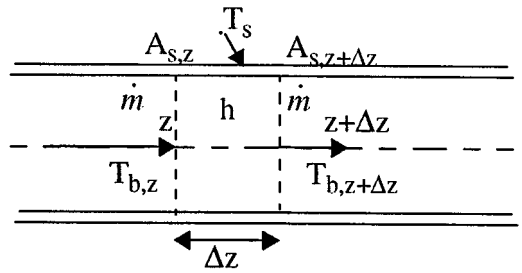


FIGURE 4.2.22 Elemental length of a tube for determining heat transfer rate.

Assuming a suitable average convective heat transfer coefficient over the entire length of the tube, separating the variables, and integrating the equation from $z = 0$ to $z = L$, we obtain

$$\ln\frac{T_s - T_{be}}{T_s - T_{bi}} = -\frac{hA_s}{\dot{m}C_p} \tag{4.2.103}$$

Equation (4.2.103) gives the exit temperature. For a constant-density fluid or an ideal gas, the heat transfer rate is determined from

$$q = \dot{m}C_p\left(T_{be} - T_{bi}\right) \tag{4.2.104}$$

Equation (4.2.103) was derived on the basis of uniform convective heat transfer coefficient. However, if the functional relationship between h and T_b is known, Equation (4.2.102) can be integrated by substituting the relationship. The convective heat transfer coefficient variation with T_b for water in two tubes of different diameters for two different flow rates is shown in Figure 4.2.23. From the figure it is clear that h can be very well approximated as a linear function of T. By substituting such a linear function relationship into Equation (4.2.102), it can be shown that

$$\ln\frac{h_i}{h_e}\frac{T_s - T_{be}}{T_s - T_{bi}} = -\frac{h_s A_s}{\dot{m}C_p} \tag{4.2.105}$$

where h_i, h_e, and h_s are the values of the convective heat transfer coefficient evaluated at bulk temperatures of T_{bi}, T_{be}, and T_s, respectively. Although it has been demonstrated that h varies approximately linearly with the bulk temperature with water as the fluid, the variation of h with air and oil as the fluid is much smaller and is very well approximated by a linear relationship. For other fluids it is suggested that the relationship be verified before employing Equation (4.2.105). [**Note:** It is tempting to determine the heat transfer rate from the relation

$$q = hA_s\frac{\left(T_s - T_{be}\right) + \left(T_s - T_{bi}\right)}{2}$$

Replacing q by Equation (4.2.104) and solving for T_{be} for defined values of the mass flow rate and tube surface area, the second law of thermodynamics will be violated if $hA_s/\dot{m}C_p > 2$. Use of Equation (4.2.103) or (4.2.105) ensures that no violation of the second law occurs however large A_s is.]

Uniform Surface Heat Flux: If the imposed heat flux is known, the total heat transfer rate for a defined length of the tube is also known. From Equation (4.2.104) the exit temperature of the fluid is determined. The fluid temperature at any location in the pipe is known from the heat transfer rate up to that location ($q = q''A_s$) and Equation (4.2.104). The convective heat transfer coefficient is used to find the surface temperature of the tube.

Temperature Variation of the Fluid with Uniform Surface Temperature and Uniform Heat Flux: The fluid temperature variations in the two cases are different. With the assumption of uniform heat transfer coefficient, with a uniform surface temperature the heat flux decreases in the direction of flow leading to a progressively decreasing rate of temperature change in the fluid with axial distance. With uniform heat flux, the surface and fluid temperatures vary linearly except in the entrance region where the higher heat transfer coefficient leads to a smaller difference between the surface and fluid temperatures. The variation of the fluid temperature in the two cases is shown in Figure 4.2.24.

Convective Heat Transfer in Noncircular Tubes

Laminar Flows: The Nusselt numbers for laminar flows have been analytically determined for different noncircular ducts. Some of them can be found in Kakac et al. (1987), Kays and Crawford (1993), and

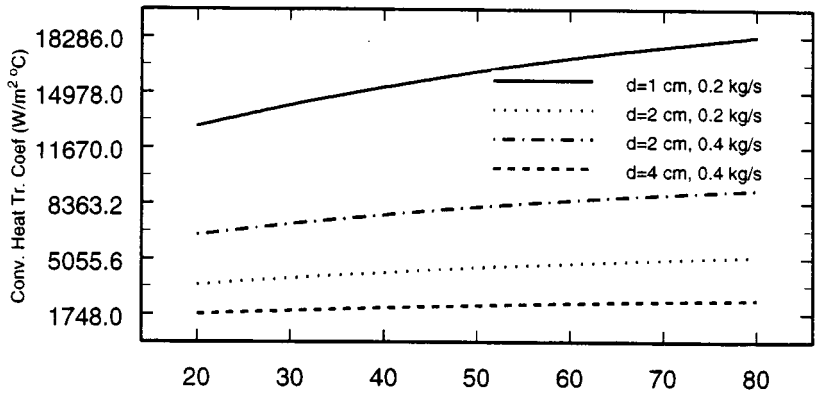


FIGURE 4.2.23 Variation of h with T_b in 1-, 2-, and 4-cm-diameter tubes with water flow rates of 0.2 kg/sec and 0.4 kg/sec with uniform surface temperature.

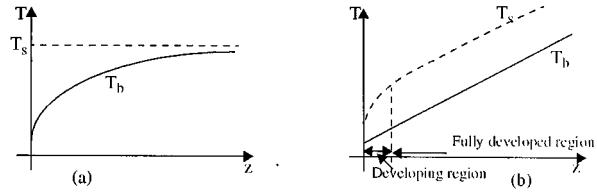


FIGURE 4.2.24 Variation of fluid temperature in a tube with (a) uniform surface temperature and (b) uniform heat flux.

Burmeister (1993). A few of the results are given below. The characteristic length for forming the Reynolds number and Nusselt number is the hydraulic mean diameter defined as

$$d_h = \frac{4 \text{ cross-sectional area}}{\text{wetted perimeter}}$$

Infinite parallel plates: a = spacing between plates, $d_h = 2a$

Both plates maintained at uniform and equal temperatures: Nu = 7.54
Both plates with imposed uniform and equal heat fluxes: Nu = 8.24

Rectangular ducts: a = longer side, b = shorter side, $d_h = 2ab/(a + b)$

b/a	**1**	**0.7**	**0.5**	**0.25**	**0.125**
Uniform surface temperature	2.98	3.08	3.39	4.44	5.6
Uniform heat flux*	3.61	3.73	4.12	5.33	6.49

Equilateral triangle: $d_h = a/3^{1/2}$, a = length of each side

Uniform surface temperature: Nu = 2.35
Uniform surface heat flux:* Nu = 3.0

Coaxial tubes: With coaxial tubes many different cases arise — each tube maintained at uniform but different temperatures, each tube subjected to uniform but different heat fluxes (an insulated

* Uniform axial heat flux but circumferentially uniform surface temperature.

surface is a special case of imposed heat flux being zero), or a combinations of uniform surface temperature of one tube and heat flux on the other. The manner in which the heat transfer coefficient is determined for uniform but different heat fluxes on the two tubes is described below. Define:

$$d_h = 2(r_o - r_i) \qquad r^* = r_i / r_o$$

$$q_i'' = h_i(T_i - T_b) \qquad \mathrm{Nu}_i = \frac{h_i d_h}{k} \qquad q_o'' = h_o(T_o - T_b) \qquad \mathrm{Nu}_o = \frac{h_o d_h}{k}$$

$$q_o'' = 0 \qquad \mathrm{Nu}_{ii} = \frac{h_i d_h}{k} \quad \text{and} \quad q_i'' = 0 \qquad \mathrm{Nu}_{oo} = \frac{h_o d_h}{k}$$

Then

$$\mathrm{Nu}_i = \frac{\mathrm{Nu}_{ii}}{1 - \dfrac{q_o''}{q_i''}\theta_i^*} \quad \text{and} \quad \mathrm{Nu}_o = \frac{\mathrm{Nu}_{oo}}{1 - \dfrac{q_i''}{q_o''}\theta_o^*} \tag{4.2.106}$$

TABLE 4.2.6 Values for Use with Equation (4.2.106)

r^*	Nu_{ii}	Nu_{oo}	θ_i^*	θ_o^*
0.05	17.81	4.792	2.18	0.0294
0.1	11.91	4.834	1.383	0.0562
0.2	8.499	4.883	0.905	0.1041
0.4	6.583	4.979	0.603	0.1823
0.6	5.912	5.099	0.473	0.2455
0.8	5.58	5.24	0.401	0.299
1.0	5.385	5.385	0.346	0.346

Some of the values needed for the computations of Nu_i and Nu_o (taken from Kays and Crawford, 1993) are given in the Table 4.2.6.

For a more detailed information on heat transfer and friction factors for laminar flows in noncircular tubes refer to Kakac et al. (1987).

Turbulent Flows: For noncircular tubes, estimates of the convective heat transfer coefficient can be obtained by employing equations for circular tubes with d_h replacing d in the computations of the Reynolds and Nusselt numbers. To determine the heat transfer coefficients in developing regions and for more-accurate values with turbulent flows in noncircular tubes refer to Kakac et al. (1987) and the references in that book.

Mixed Convection

If the fluid velocity is low, the effect of natural convection becomes significant and the heat transfer rate may be increased or decreased by natural convection. From a review of experimental results, Metais and Eckert (1964) developed maps to delineate the different regimes where one or the other mode is dominant and where both are significant. Figures 4.2.25 and 4.2.26 show the relative significance of natural and forced convection in vertical and horizontal tubes. The maps are applicable for $10^{-2} < \mathrm{Pr}(d/L) < 1$ where d and L are the diameter and the axial length of the tube. The maps show the limits of forced and natural convection regimes. The limits are delineated "in such a way that the actual heat flux under the combined influence of the forces does not deviate by more than 10 percent from the heat flux that

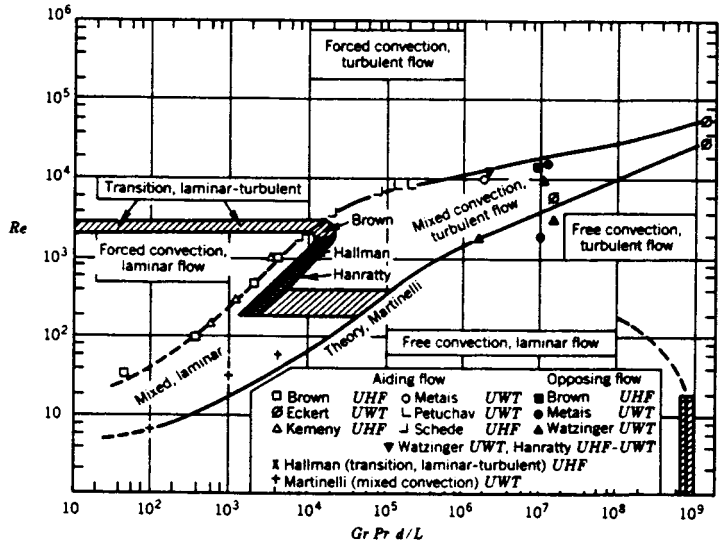


FIGURE 4.2.25 Map delineating forced, mixed, and natural convection — vertical tubes.

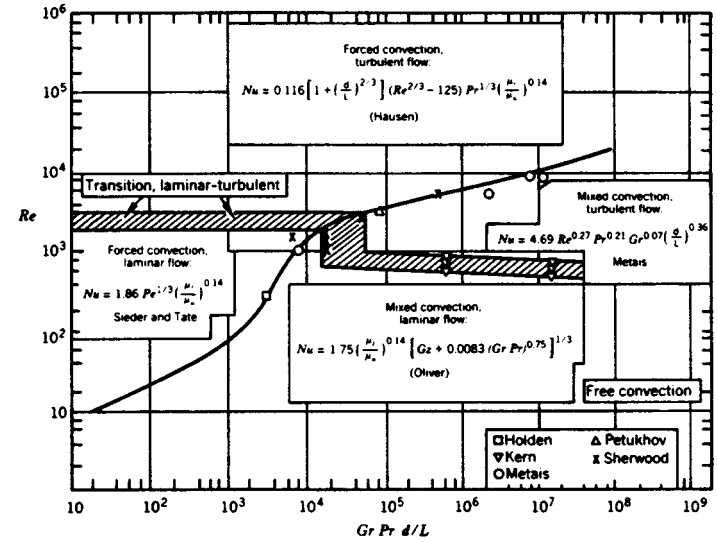


FIGURE 4.2.26 Map delineating forced, mixed, and natural convection — horizontal tubes.

would be caused by the external forces alone or by the body forces alone." The Grashof number is based on the diameter of the tube.

For flows in horizontal tubes, correlations were developed for the mixed convection regime in isothermal tubes by Depew and August (1971) and for uniform heat flux by Morcos and Bergles (1975).

Uniform Surface Temperature. Fully developed velocity profile, developing temperature profile:

$L/d < 28.4$ $\quad 25 < Gz < 712$ $\quad 0.7 \times 10^5 < Gr < 9.9 \times 10^5$

μ_s = dynamic viscosity, evaluated at the wall temperature

All other properties at the average bulk temperature of the fluid

$$\mathrm{Gz} = \frac{\dot{m}C_p}{kL} \qquad \mathrm{Gr} = g\beta\Delta T\, d^3/\nu^2$$

$$\mathrm{Nu}_d = 1.75\left[\mathrm{Gz} + 0.12\left(\mathrm{GzGr}^{1/3}\mathrm{Pr}^{0.36}\right)^{0.88}\right]^{1/3}\left(\mu_b/\mu_s\right)^{0.14} \tag{4.2.107}$$

Uniform Heat Flux. Properties at $(T_s + T_b)/2$: $3 \times 10^4 < \mathrm{Ra} < 10^6$, $4 < \mathrm{Pr} < 175$, $2 < hd^2/(k_w t) < 66$, k_w = tube wall thermal conductivity, t = tube wall thickness.

$$\mathrm{Gr}_d^* = g\beta d^4 q_w''/\left(\nu^2 k\right) \qquad P_w = kd/\left(k_w t\right) \qquad \mathrm{Ra}_d = g\beta\Delta T\, d^3\mathrm{Pr}/\nu^2$$

$$\mathrm{Nu}_d = \left\{4.36^2 + \left[0.145\left(\frac{\mathrm{Gr}_d^*\mathrm{Pr}^{1.35}}{P_w^{0.25}}\right)^{0.265}\right]^2\right\}^{0.5} \tag{4.2.108}$$

In Equation (4.2.107) and (4.2.108) evaluate fluid properties at the arithmetic mean of the bulk and wall temperatures.

Nomenclature

A_s — surface area
d — diameter
d_h — hydraulic mean diameter
f — friction factor
h — convective heat transfer coefficient
k — fluid thermal conductivity
L_e — hydrodynamic entrance length
$L_{e,th}$ — thermal entrance length
Nu_d — Nusselt number
Nu_{ii} — Nusselt number with only inner tube heated
Nu_{oo} — Nusselt number with only outer tube heated
Pr — Prandtl number
q'' — heat flux
q_i'' — heat flux on the inner tube surface
q_o'' — heat flux on the outer tube surface
Re_d — Reynolds number ($\rho vd/\mu$)
T_b — bulk temperature
T_s — surface temperature
v — average fluid velocity
μ — dynamic viscosity
μ_s — dynamic viscosity at surface temperature
ρ — fluid density

References

Burmeister, L.C. 1993. *Convective Heat Transfer*, 2nd ed., Wiley-Interscience, New York.

Depew, C.A. and August, S.E. 1971. Heat transfer due to combined free and forced convection in a horizontal and isothermal tube, *Trans. ASME* 93C, 380.

Dittus, F.W. and Boelter, L.M.K. 1930. Heat transfer in automobile radiators of the tubular type, *Univ. Calif. Pub. Eng.*, 13, 443.

Gnielinsky, V. 1976. New equations for heat and mass transfer in turbulent pipe channel flow, *Int. Chem. Eng.*, 16, 359.

Gnielinsky, V. 1990. Forced convection in ducts, in *Handbook of Heat Exchanger Design*, Hewitt, G.F., Ed., Begell House/Hemisphere, New York.

Kakac, S., Shah, R.K., and Win Aung, Eds. 1987. *Handbook of Single-Phase Convective Heat Transfer*, Wiley-Interscience, New York.

Kays, W.M. and Crawford, M.E. 1993. *Convective Heat and Mass Transfer*, 3rd ed., McGraw-Hill, New York.

Metais, B. and Eckert, E.R.G. 1964. Forced, mixed, and free convection regimes, *Trans. ASME* 86C, 295.

Morcos, S.M. and Bergles, A.E. 1975. Experimental investigation of combined forced and free laminar convection in a horizontal tube, *Trans. ASME* 97C, 212.

Sieder, E.N. and Tate, C.E. 1936. Heat transfer and pressure drop of liquids in tubes, *Ind. Eng. Chem.*, 28, 1429.

Sleicher, C.A. and Rouse, M.W. 1976. A convenient correlation for heat transfer to constant and variable property fluids in turbulent pipe flow, *Int. J. Heat Mass Transfer*, 18, 677.

4.3 Radiation*

Michael F. Modest

Nature of Thermal Radiation

All materials continuously emit and absorb radiative energy by lowering or raising their molecular energy levels. This thermal radiative energy may be viewed as consisting of electromagnetic waves or of massless energy parcels, called **photons**. Electromagnetic waves travel through any medium at the speed of light c, which is $c_0 = 2.998 \times 10^8$ m/sec in vacuum and approximately the same in most gases such as air and combustion products. These are characterized by their wavelength λ or frequency ν, which are related by

$$\nu = c/\lambda \tag{4.3.1}$$

The strength and wavelengths of **emission** and **absorption** depend on the temperature and nature of the material.

The ability of photons to travel unimpeded through vacuum and gases makes thermal radiation the dominant mode of heat transfer in vacuum, low-pressure environments, and outer space applications (due to the near absence of conduction and convection). Its temperature dependence [as given by Equation (4.3.3) below] on the other hand, guarantees that radiative heat transfer is of utmost importance in high-temperature applications (including solar radiation: with the sun being a high-temperature heat source at an effective temperature of $T_{\text{sun}} = 5762$ K).

When an electromagnetic wave traveling through a gas (or vacuum) strikes the surface of a medium, the wave may be partly or totally reflected, and any nonreflected part will penetrate into the medium. If a wave passes through a medium without any attenuation, the material is called **transparent**. A body with partial attenuation is known as **semitransparent**, and a body through which none of the incoming radiation penetrates is called **opaque**. Most gases are rather transparent to radiation (except for narrow spectral regions, called *absorption bands*), while most solids tend to be strong absorbers for most wavelengths, making them opaque over a distance of a few nanometers (electrical conductors, i.e., metals) to a few micrometers (ceramics, semiconductors), or more (dielectrics).

Blackbody Radiation

The total amount of radiative energy emitted from a surface into all directions above it is termed **emissive power**; we distinguish between **spectral** (at a given wavelength λ, per unit wavelength) and total (encompassing all wavelengths) emissive power. The magnitude of emissive power depends on wavelength λ, temperature T, and a surface property, called **emissivity** ε, which relates the ability of a surface to emit radiative energy to that of an ideal surface, which emits the maximum possible energy (at a given wavelength and temperature). Such an ideal surface is known as a "**blackbody**" or "black surface," since it absorbs all incoming radiation; i.e., it reflects no radiation and is, therefore, invisible ("black") to the human eye. The spectral distribution of the emissive power of a black surface is given by **Planck's law**.

$$E_{b\lambda} = \frac{C_1}{\lambda^5\left[e^{C_2/\lambda T} - 1\right]}, \qquad C_1 = 3.7419 \times 10^{-16}\ \text{Wm}^2, \qquad C_2 = 14{,}388\ \mu\text{mK} \tag{4.3.2}$$

where C_1 and C_2 are sometimes called Planck function constants. The total emissive power of a blackbody is given by

*From Modest, M., Radiative Heat Transfer, 1993; reproduced with permission of MacGraw-Hill, Inc.

$$E_b = \int_0^\infty E_{b\lambda}\, d\lambda = \sigma T^4, \qquad \sigma = 5.670 \times 10^{-8}\ \mathrm{W/m^2K^4} \tag{4.3.3}$$

with σ known as the Stefan–Boltzmann constant. Figure 4.3.1 shows the spectral solar irradiation that impinges on Earth, which closely resembles the spectrum of a blackbody at 5762 K. The general behavior of Planck's law is depicted in Figure 4.3.2, together with the fractional emissive power, $f(\lambda T)$, defined as

$$f(\lambda T) = \frac{1}{E_b}\int_0^\lambda E_{b\lambda}(\lambda, T)\, d\lambda \tag{4.3.4}$$

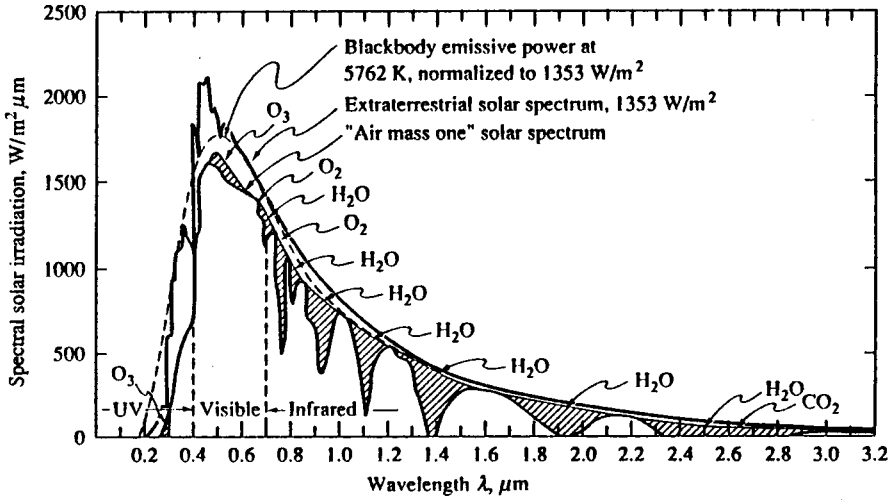


FIGURE 4.3.1 Solar irradiation onto Earth.

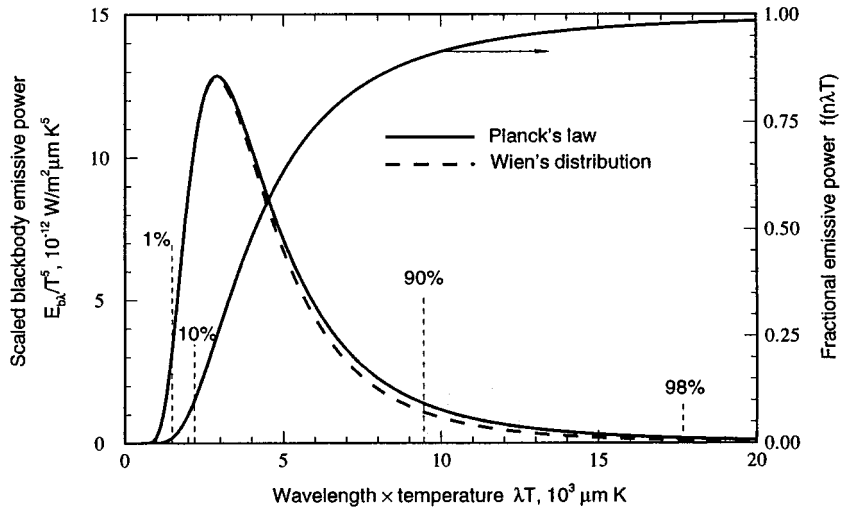


FIGURE 4.3.2 Normalized blackbody emissive power spectrum.

Note that 90% of all blackbody emission takes place at wavelengths of $\lambda T > 2200$ μmK and at all wavelengths $\lambda T < 9400$ μmK. This implies that — for typical high-temperature heat transfer applications in the range between 1000 and 2000 K — infrared wavelengths in the range 1 μm < λ < 10 μm govern the heat transfer rates. For solar applications shorter wavelengths, down to $\lambda \cong 0.4$ μm are also important. Also shown in Figure 4.3.2 is Wien's law:

$$E_{b\lambda} = \frac{C_1}{\lambda^5} e^{-C_2/\lambda T} \tag{4.3.5}$$

which approximates Planck's law accurately over the part of the spectrum that is important to heat transfer, and that is easier to manipulate mathematically.

Example 4.3.1

What fraction of total solar emission falls into the visible spectrum (0.4 to 0.7 μm)?

Solution: With a solar temperature of 5762 K it follows that for

$$\lambda_1 = 0.4\ \mu\text{m}, \qquad \lambda_1 T_{\text{sun}} = 0.4 \times 5762 = 2304\ \mu\text{mK}$$

and for

$$\lambda_2 = 0.7\ \mu\text{m}, \qquad \lambda_2 T_{\text{sun}} = 0.7 \times 5762 = 4033\ \mu\text{mK}$$

From Figure 4.3.2 we can estimate $f(\lambda_1 T_{\text{sun}}) \cong 12\%$ and $f(\lambda_2 T_{\text{sun}}) \cong 48\%$. Thus, the visible fraction of sunlight is 48 – 12 ≅ 36%: with a bandwidth of only 0.3 μm the human eye responds to approximately 36% of all emitted sunlight!

Radiative Exchange between Opaque Surfaces

Radiative Properties of Surfaces

Strictly speaking, the surface of an enclosure wall can only reflect radiative energy and allow a part of it to penetrate into the substrate. A surface cannot absorb or emit photons: attenuation takes place inside the solid, as does emission of radiative energy (with some of the emitted energy escaping through the surface into the enclosure). In practical systems, the thickness of the surface layer over which absorption of **irradiation** from inside the enclosure occurs is very small compared with the overall dimension of an enclosure — usually a few nanometers for metals and a few micrometers for most nonmetals. The same may be said about emission from within the walls that escapes into the enclosure. Thus, in the case of opaque walls it is customary to speak of absorption by and emission from a "surface," although a thin surface layer is implied. Four fundamental radiative properties are defined:

Reflectivity,
$$\rho \equiv \frac{\text{reflected part of incoming radiation}}{\text{total incoming radiation}} \tag{4.3.6a}$$

Absorptivity,
$$\rho \equiv \frac{\text{absorbed part of incoming radiation}}{\text{total incoming radiation}} \tag{4.3.6b}$$

Transmissivity,
$$\tau \equiv \frac{\text{transmitted part of incoming radiation}}{\text{total incoming radiation}} \tag{4.3.6c}$$

Emissivity, $$\varepsilon \equiv \frac{\text{energy emitted from a surface}}{\text{energy emitted by a black surface at same temperature}} \tag{4.3.6d}$$

Since all incoming radiation must be reflected, absorbed, or transmitted, it follows that

$$\rho + \alpha + \tau = 1 \tag{4.37}$$

In most practical applications surface layers are thick enough to be opaque ($\tau = 0$, leading to $\rho + \alpha = 1$. All four properties may be functions of wavelength, temperature, incoming direction (except emissivity), and outgoing direction (except absorptivity).

Directional Behavior. For heat transfer applications, the dependence on incoming direction for absorptivity (as well as ρ and τ) and outgoing direction for emissivity is generally weak and is commonly neglected; i.e., it is assumed that the surface absorbs and emits **diffusely**. Then, for an opaque surface, for any given wavelength

$$\varepsilon_\lambda = \alpha_\lambda = 1 - \rho_\lambda \tag{4.38}$$

Published values of emissivities are generally either "normal emissivities" (the directional value of ε_λ in the direction perpendicular to the surface) or "hemispherical emissivities" (an average value over all outgoing directions). The difference between these two values is often smaller than experimental accuracy and/or repeatability.

Reflected energy (due to a single, distinct incoming direction) may leave the surface into a single direction ("specular" reflection, similar to reflection from a mirror for visible light), or the reflection may spread out over all possible outgoing directions. In the extreme case of equal amounts going into all directions, we talk about "diffuse" reflection. Smooth surfaces (as compared with the wavelength of radiation) tend to be specular reflectors, while rough surfaces tend to be more or less diffusely reflecting. Analysis is vastly simplified if diffuse reflections are assumed. Research has shown that — except for some extreme geometries and irradiation conditions susceptible to beam channeling (irradiated open cavities, channels with large aspect ratios) — radiative heat transfer rates are only weakly affected by the directional distribution of reflections. Therefore, it is common practice to carry out radiative heat transfer calculations assuming only diffuse reflections.

Spectral Dependence. The emissivity of a surface generally varies strongly and in complex ways with wavelength, depending on the material, surface layer composition, and surface structure (roughness). Therefore, unlike bulk material properties (such as thermal conductivity) the surface emissivity may display significant differences between two ostensibly identical samples, and even for one and the same sample measured at different times (due to surface roughness and contamination). Despite these difficulties, surfaces may be loosely grouped into two categories — metals and nonconductors (dielectrics), and some generalizations can be made.

Polished Metals. Smooth, purely metallic surfaces (i.e., without any nonmetallic surface contamination, such as metal oxides) tend to have very low emissivities in the infrared. For many clean metals $\varepsilon_\lambda < 0.1$ for $\lambda > 2$ μm, and spectral as well as temperature dependence are generally well approximated by the proportionality $\varepsilon_\lambda \propto \sqrt{T/\lambda}$ in the infrared. However, for shorter wavelengths ($\lambda < 1$ μm), emissivity values may become quite substantial, and temperature dependence is usually reversed (decreasing, rather than increasing, with temperature). Typical room temperature behavior of several metals is shown in Figure 4.3.3. Caution needs to be exercised when choosing an emissivity value for a metal surface: unless extraordinary care is taken to keep a polished metal clean (i.e., free from oxidation and/or surface contamination), its emissivity may soon become several times the value of the original, polished specimen (for example, consider the formation of aluminum oxide on top of aluminum, Figure 4.3.3).

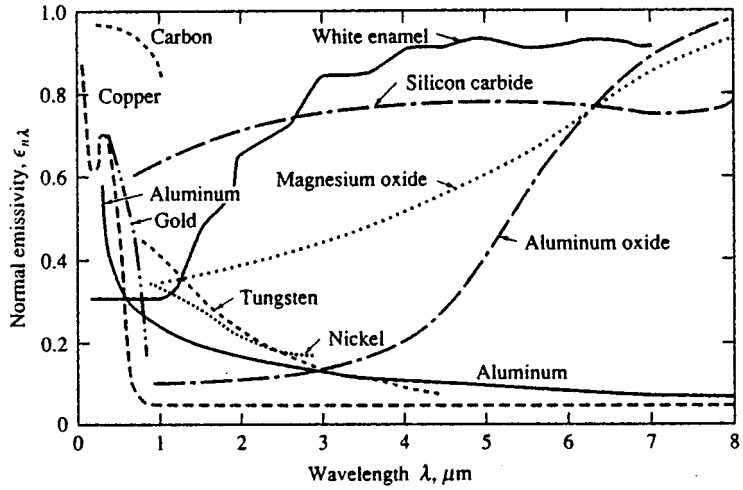


FIGURE 4.3.3 Normal, spectral emissivities for selected materials.

Ceramics and Refractories. Smooth ceramics tend to have fairly constant and intermediate emissivity over the near- to mid-infrared, followed by a sharp increase somewhere between 4 and 10 μm. At short wavelengths these materials display strong decreases in emissivity, so that a number of them may appear white to the human eye even though they are fairly black in the infrared. The temperature dependence of the emissivity of ceramics is rather weak; generally a slight increase with temperature is observed in the infrared. The spectral emissivity of a few ceramics is also shown in Figure 4.3.3.

Other Nonconductors. The behavior of most electrically nonconducting materials is governed by surface structure, nonhomogeneity, dopants, porosity, flaws, surface films, etc. The emissivity may vary irregularly across the spectrum because of various emission bands, influence of flaws, etc., making any generalization impossible. This irregularity may be exploited to obtain surfaces of desired spectral behavior, so-called selective surfaces. Some selective surfaces (as compared with a common steel) are depicted in Figure 4.3.4. For a solar collector it is desirable to have a high spectral emissivity for short wavelengths $\lambda < 2.5$ μm (strong absorption of solar irradiation), and a low value for $\lambda > 2.5$ μm (to minimize re-emission from the collector). The opposite is true for a spacecraft radiator panel used to reject heat into space.

It is clear that (1) values of spectral surface emissivity are subject to great uncertainty and (2) only a relatively small range of infrared wavelengths are of importance. Therefore, it is often assumed that the surfaces are "**gray**", i.e., the emissivity is constant across (the important fraction of) the spectrum, $\varepsilon_\lambda \neq \varepsilon_\lambda(\lambda)$, since this assumption also vastly simplifies analysis. Table 4.3.1 gives a fairly detailed listing of total emissivities of various materials, defined as

$$\varepsilon(T) = \frac{1}{E_b(T)} \int_0^\infty \varepsilon_\lambda(\lambda, T) E_{b\lambda}(T)\, d\lambda \tag{4.3.9}$$

which may be enlisted for a gray analysis.

View Factors

In many engineering applications the exchange of radiative energy between surfaces is virtually unaffected by the medium that separates them. Such (radiatively) *nonparticipating media* include vacuum as well as monatomic and most diatomic gases (including air) at low to moderate temperature levels (i.e., before ionization and dissociation occurs). Examples include spacecraft heat rejection systems, solar collector systems, radiative space heaters, illumination problems, and so on. It is common practice

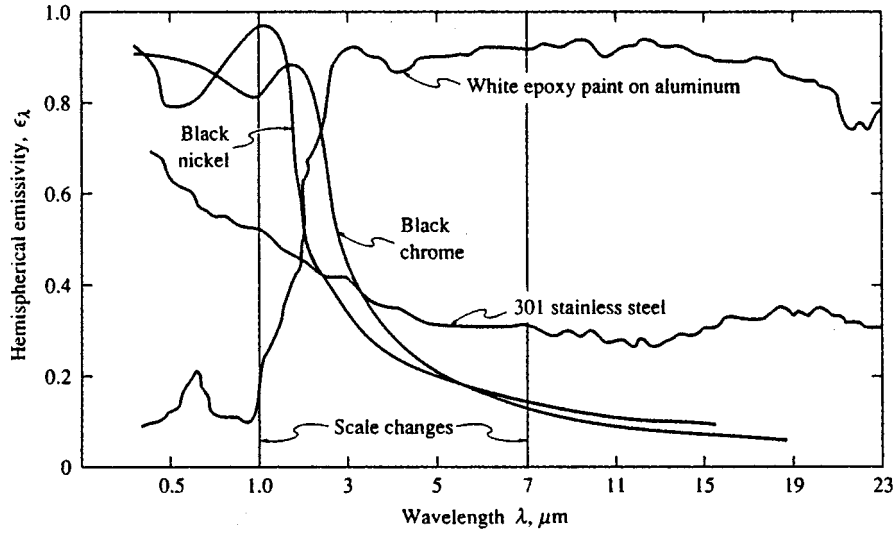


FIGURE 4.3.4 Spectral, hemispherical reflectivities of several spectrally selective surfaces.

to simplify the analysis by making the assumption of an *idealized enclosure* and/or of *ideal surface properties*. The greatest simplification arises if all surfaces are black: for such a situation no reflected radiation needs to be accounted for, and all emitted radiation is diffuse (i.e., the radiative energy leaving a surface does not depend on direction). The next level of difficulty arises if surfaces are assumed to be gray, diffuse emitters (and, thus, absorbers) as well as gray, diffuse reflectors. The vast majority of engineering calculations are limited to such ideal surfaces, since, particularly, the effects of nondiffuse reflections are usually weak (see discussion in previous section).

Thermal radiation is generally a long-range phenomenon. This is *always* the case in the absence of a participating medium, since photons will travel unimpeded from surface to surface. Therefore, performing a thermal radiation analysis for one surface implies that all surfaces, no matter how far removed, that can exchange radiative energy with one another must be considered simultaneously. How much energy any two surfaces exchange depends in part on their size, separation, distance, and orientation, leading to geometric functions known as **view factors**, defined as

$$F_{i-j} = \frac{\text{diffuse energy leaving } A_i \text{ directly toward and intercepted by } A_j}{\text{total diffuse energy leaving } A_i} \tag{4.3.10}$$

In order to make a radiative energy balance we always need to consider an entire *enclosure* rather than and infinitesimal control volume (as is normally done for other modes of heat transfer, i.e., conduction or convection). The enclosure must be closed so that irradiation from all possible directions can be accounted for, and the enclosure surfaces must be *opaque* so that all irradiation is accounted for, for each direction. In practice, an incomplete enclosure may be closed by introducing artificial surfaces. An enclosure may be idealized in two ways, as indicated in Figure 4.3.5: by replacing a complex geometric shape with a few simple surfaces, and by assuming surfaces to be isothermal with constant (i.e., average) heat flux values across them. Obviously, the idealized enclosure approaches the real enclosure for sufficiently small isothermal subsurfaces.

Mathematically, the view factor needs to be determined from a double integral, i.e.,

$$F_{i-j} = \frac{1}{A_i}\int_{A_i}\int_{A_j} \frac{\cos\theta_i \cos\theta_j}{\pi S_{ij}^2} dA_j\, dA_i \tag{4.3.11}$$

TABLE 4.3.1 Total Emissivity and Solar Absorptivity of Selected Surfaces

	Temperature (°C)	Total Normal Emissivity	Extraterrestrial Solar Absorptivity
Alumina, flame-sprayed	–25	0.80	0.28
Aluminum foil			
As received	20	0.04	
Bright dipped	20	0.025	0.10
Aluminum, vacuum-deposited	20	0.025	0.10
Hard-anodized	–25	0.84	0.92
Highly polished plate, 98.3% pure	225–575	0.039–0.057	
Commercial sheet	100	0.09	
Rough polish	100	0.18	
Rough plate	40	0.055–0.07	
Oxidized at 600°C	200–600	0.11–0.19	
Heavily oxidized	95–500	0.20–0.31	
Antimony, polished	35–260	0.28–0.31	
Asbestos	35–370	0.93–0.94	
Beryllium	150	0.18	0.77
	370	0.21	
	600	0.30	
Beryllium, anodized	150	0.90	
	370	0.88	
	600	0.82	
Bismuth, bright	75	0.34	
Black paint			
Parson's optical black	–25	0.95	0.975
Black silicone	–25–750	0.93	0.94
Black epoxy paint	–25	0.89	0.95
Black enamel paint	95–425	0.81–0.80	
Brass, polished	40–315	0.10	
Rolled plate, natural surface	22	0.06	
Dull plate	50–350	0.22	
Oxidized by heating at 600°C	200–600	0.61–0.59	
Carbon, graphitized	100–320	0.76–0.75	
	320–500	0.75–0.71	
Candle soot	95–270	0.952	
Graphite, pressed, filed surface	250–510	0.98	
Chromium, polished	40–1100	0.08–0.36	
Copper, electroplated	20	0.03	0.47
Carefully polished electrolytic copper	80	0.018	
Polished	115	0.023	
Plate heated at 600°C	200–600	0.57	
Cuprous oxide	800–1100	0.66–0.54	
Molten copper	1075–1275	0.16–0.13	
Glass, Pyrex, lead, and soda	260–540	0.95–0.85	
Gypsum	20	0.903	
Gold, pure, highly polished	225–625	0.018–0.035	
Inconel X, oxidized	–25	0.71	0.90
Lead, pure (99.96%), unoxidized	125–225	0.057–0.075	
Gray oxidized	25	0.28	
Oxidized at 150°C	200	0.63	
Magnesium oxide	275–825	0.55–0.20	
	900–1705	0.20	
Magnesium, polished	35–260	0.07–0.13	
Mercury	0–100	0.09–0.12	
Molybdenum, polished	35–260	0.05–0.08	
	540–1370	0.10–0.18	

TABLE 4.3.1 (continued) Total Emissivity and Solar Absorptivity of Selected Surfaces

	Temperature (°C)	Total Normal Emissivity	Extraterrestrial Solar Absorptivity
	2750	0.29	
Nickel, electroplated	20	0.03	0.22
Polished	100	0.072	
Platinum, pure, polished	225–625	0.054–0.104	
Silica, sintered, powdered, fused silica	35	0.84	0.08
Silicon carbide	150–650	0.83–0.96	
Silver, polished, pure	40–625	0.020–0.032	
Stainless steel			
Type 312, heated 300 hr at 260°C	95–425	0.27–0.32	
Type 301 with Armco black oxide	–25	0.75	0.89
Type 410, heated to 700°C in air	35	0.13	0.76
Type 303, sandblasted	95	0.42	0.68
Titanium, 75A	95–425	0.10–0.19	
75A, oxidized 300 hr at 450°C	35–425	0.21–0.25	0.80
Anodized	–25	0.73	0.51
Tungsten, filament, aged	27–3300	0.032–0.35	
Zinc, pure, polished	225–325	0.045–0.053	
Galvanized sheet	100	0.21	

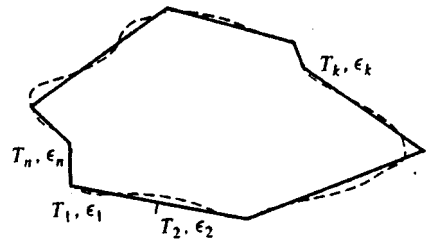


FIGURE 4.3.5 Real and ideal enclosures for radiative transfer calculations.

where θ_i and θ_j are the angles between the surface normals on A_i and A_j, respectively, and the line (of length S_{ij}) connecting two points on the two surfaces. Analytical solutions to Equation (4.3.11) may be found for relatively simple geometries. A few graphical results for important geometries are shown in Figures 4.3.6 to 4.3.8. More-extensive tabulations as well as analytical expressions may be found in textbooks on the subject area (Modest, 1993; Siegel and Howell, 1992) as well as view factor catalogs (Howell, 1982). For nontrivial geometries view factors must be calculated numerically, either (1) by numerical quadrature of the double integral in Equation (4.3.11), or (2) by converting Equation (4.3.11) into a double-line integral, followed by numerical quadrature, or (3) by a Monte Carlo method (statistical sampling and tracing of selected light rays).

View Factor Algebra. For simple geometries analytical values can often be found by expressing the desired view factor in terms of other, known ones. This method is known as *view factor algebra*, by manipulating the two relations,

Reciprocity rule: $$A_i F_{i-j} = A_j F_{j-i} \tag{4.3.12}$$

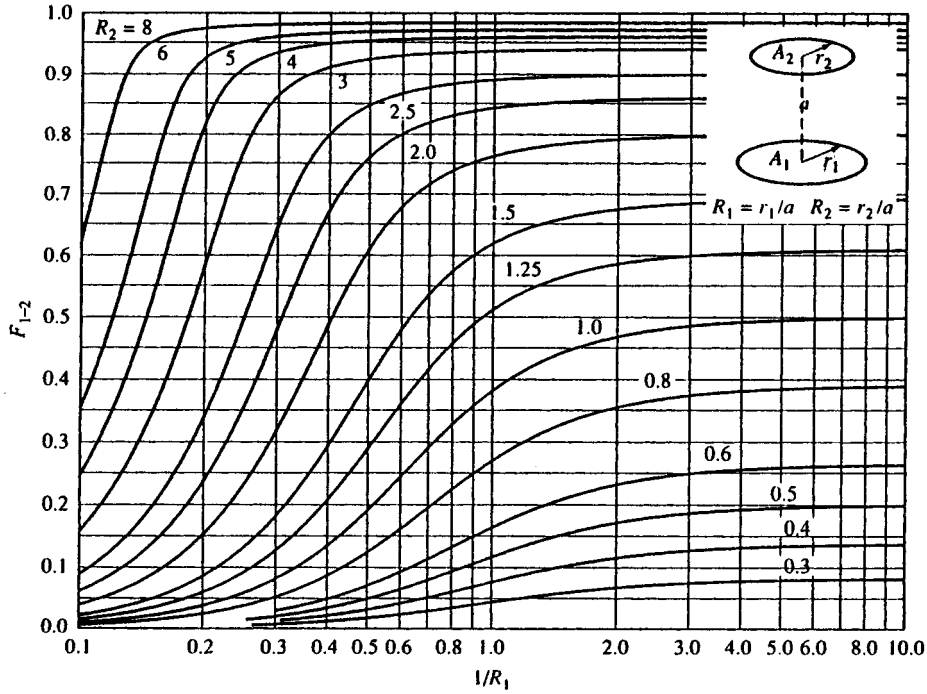


FIGURE 4.3.6 View factor between parallel, coaxial disks of unequal radius.

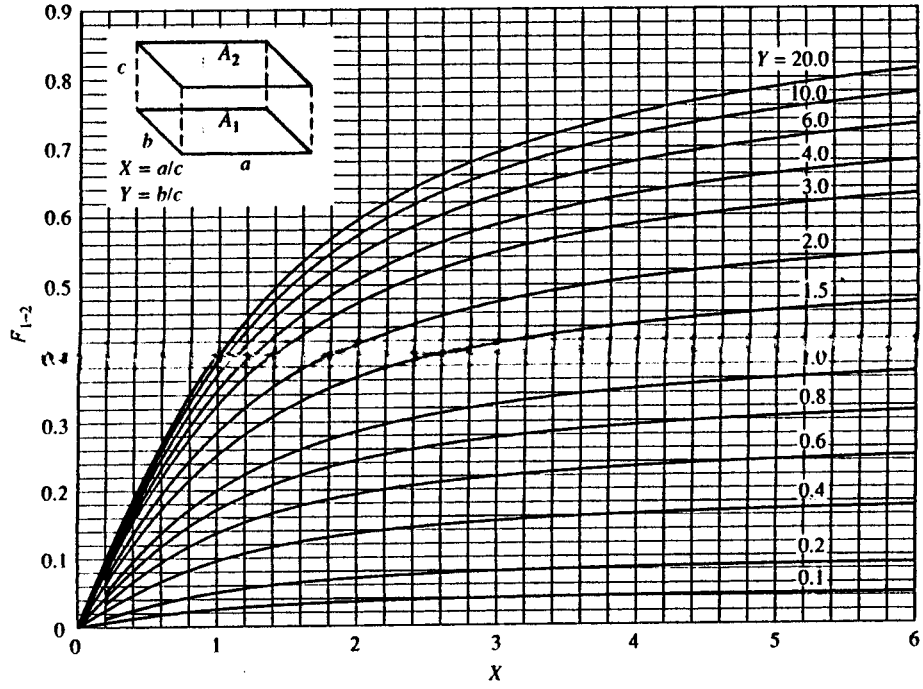


FIGURE 4.3.7 View factor between identical, parallel, directly opposed rectangles.

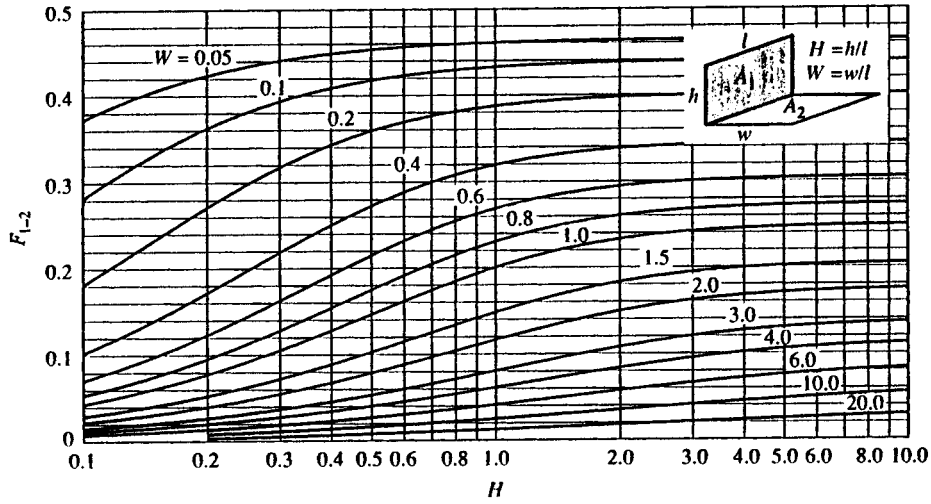


FIGURE 4.3.8 View factor between perpendicular rectangles with common edge.

$$\textit{Summation rule:} \qquad \sum_{j=1}^{N} F_{i-j} = 1, \qquad i = 1, N \tag{4.3.13}$$

assuming that the (closed) configuration consists of N surfaces. The reciprocity rule follows immediately from Equation (4.3.11), while the summation rule simply states that the fractional energies leaving surface A_i must add up to a whole.

Example 4.3.2

Assuming the view factor for a finite corner, as shown in Figure 4.3.8 is known, determine the view factor F_{3-4}, between the two perpendicular strips as shown in Figure 4.3.9.

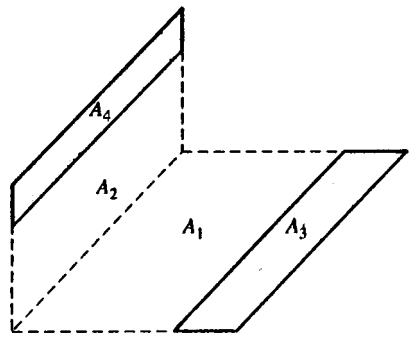


FIGURE 4.3.9 Configuration for Example 4.3.2 (strips on a corner piece).

Solution. From the definition of the view factor, and since the energy traveling to A_4 is the energy going to A_2 and A_4 minus the one going to A_2, it follows that

$$F_{3-4} = F_{3-(2+4)} - F_{3-2}$$

and, using reciprocity,

$$F_{3-4} = \frac{1}{A_3}\left[\left(A_2 + A_4\right)F_{(2+4)-3} - A_2 F_{2-3}\right]$$

Similarly, we find

$$F_{3-4} = \frac{A_2 + A_4}{A_3}\left(F_{(2+4)-(1+3)} - F_{(2+4)-1}\right) - \frac{A_2}{A_3}\left(F_{2-(1+3)} - F_{2-1}\right)$$

All view factors on the right-hand side are corner pieces and, thus, are known from Figure 4.3.8.

Crossed-Strings Method. A special type of view factor algebra may be used to determine all the view factors in long enclosures with constant cross section. The method is called the crossed-strings method since the view factors can be determined experimentally with four pins, a roll of string, and a yardstick. Consider the configuration in Figure 4.3.10, which shows the cross section of an infinitely long enclosure, continuing into and out of the plane of the figure. Repeatedly applying reciprocity and summation rules allows the evaluation of F_{1-2} as

$$F_{1-2} = \frac{\left(A_{bc} + A_{ad}\right) - \left(A_{ac} + A_{bd}\right)}{2A_1} \tag{4.3.14}$$

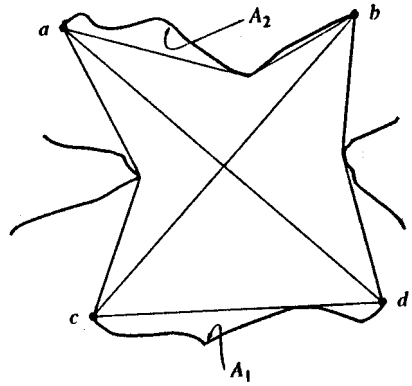


FIGURE 4.3.10 The crossed-strings method for arbitrary two-dimensional configurations.

where A_{ab} is the area (per unit depth) defined by the length of the string between points a and b, etc. This formula is easily memorized by looking at the configuration between any two surfaces as a generalized "rectangle," consisting of A_1, A_2, and the two sides A_{ac} and A_{bd}. Then

$$F_{1-2} = \frac{\text{diagonals} - \text{sides}}{2 \times \text{originating area}} \tag{4.3.15}$$

Example 4.3.3

Calculate F_{1-2} for the configuration shown in Figure 4.3.11.

Solution. From the figure it is obvious that

$$s_1^2 = (c - d\cos\alpha)^2 + d^2\sin^2\alpha = c^2 + d^2 - 2cd\cos\alpha$$

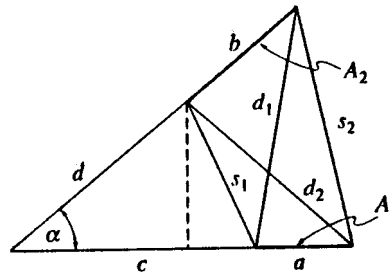


FIGURE 4.3.11 Infinitely long wedge-shaped groove for Example 4.3.3.

Similarly, we have

$$s_2^2 = (a+c)^2 + (b+d)^2 - 2(a+c)(b+d)\cos\alpha$$

$$d_1^2 = (a+c)^2 + d^2 - 2(a+c)d\cos\alpha$$

$$d_2^2 = c^2 + (b+d)^2 - 2c(b+d)\cos\alpha$$

and

$$F_{1-2} = \frac{d_1 + d_2 - (s_1 + s_2)}{2a}$$

Radiative Exchange between Opaque Surfaces (Net Radiation Method)

Consider an enclosure consisting of N opaque surfaces. The enclosure is closed, or, if not, no surface external to the surface reflects or emits radiation into the enclosure (i.e., the open configuration may be artificially closed by replacing openings with cold, black surfaces); any external radiation entering the enclosure is dealt with individually for each surface [see Equation (4.3.17) below]. All surfaces are assumed to be gray, and emit and reflect diffusely. Traditionally, the **radiosity** J of the surfaces is determined, defined as the total diffuse radiative energy leaving a surface (by emission and reflection),

$$J_i = \varepsilon_1 E_{bi} + \rho_i H_i, \qquad i = 1, N \tag{4.3.16}$$

where H_i is the incoming radiative flux (irradiation) onto surface A_i. This leads to N simultaneous equations for the unknown radiosities, specifically,

$$J_i = \varepsilon_i E_{bi} + (1-\varepsilon_i)\left[\sum_{j=1}^{N} J_i F_{i-j} + H_{oi}\right] \tag{4.3.17a}$$

or

$$J_i = q_i + \sum_{j=1}^{N} J_j F_{i-j} + H_{oi} \tag{4.3.17b}$$

depending on whether surface temperature or surface flux are known on surface A_i. In Equation (4.3.17) H_{oi} is irradiation on surface A_i from outside the enclosure, if any; H_{oi} is always zero for closed configurations, but is useful in the presence of external light sources (such as solar energy, lasers, etc.). The

radiosity neither is a useful quantity to determine, nor is there a need to determine it. Eliminating the radiosities from Equations (4.3.17a) and (4.3.17b) leads to N simultaneous equations in temperature (E_{bi}) and heat flux (q_i):

$$\frac{q_i}{\varepsilon_i} - \sum_{j=1}^{N}\left(\frac{1}{\varepsilon_j} - 1\right)F_{i-j}q_j + H_{oi} = E_{bi} - \sum_{j=1}^{N}F_{i-j}E_{bj} \tag{4.3.18}$$

Note that no artificial closing surfaces ($j > N$) appear in Equation (4.3.18), since for these surfaces $\varepsilon_j = 1$ and $E_{bj} = 0$. Thus, such closing surfaces may simply be ignored in the analysis.

Since Equation (4.3.18) is a set of N equations, this requires that N values of emissive power E_{bi} and/or flux q_i must be given as boundary conditions, in order to solve for the remaining N unknowns. For computer calculations Equation (4.3.18) may be recast in matrix form

$$\mathbf{C}\cdot\mathbf{q} = \mathbf{A}\cdot\mathbf{e_b} - \mathbf{h_o} \tag{4.3.19a}$$

where

$$C_{ij} = \frac{\delta_{ij}}{\varepsilon_j} - \left(\frac{1}{\varepsilon_j} - 1\right)F_{i-j} \tag{4.3.19b}$$

$$A_{ij} = \delta_{ij} - F_{i-j} \tag{4.3.19c}$$

δ_{ij} is Kronecker's delta, i.e.,

$$\delta_{ij} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases} \tag{4.3.20}$$

and $\mathbf{q}$, $\mathbf{e_b}$, and $\mathbf{h_o}$ are vectors of the surface heat fluxes q_i, emissive powers E_{bi}, and external irradiations H_{oi} (if any). For example, if the temperatures are given for all the surfaces, and the heat fluxes are to be determined, Equation (4.3.19) is solved by matrix inversion, and

$$\mathbf{q} = \left(\mathbf{C^{-1}}\cdot\mathbf{A}\right)\cdot\mathbf{e_b} - \left(\mathbf{C^{-1}}\cdot\mathbf{h_o}\right) \tag{4.3.21}$$

Example 4.3.4

A right-angled groove, consisting of two long black surfaces of width a, is exposed to solar radiation q_{sol} (Figure 4.3.12). The entire groove surface is kept isothermal at temperature T. Determine the net radiative heat transfer rate from the groove.

Solution. We may employ Equation (4.3.19). However, the enclosure is not closed, and we must close it artificially. We note that any radiation leaving the cavity will not come back (barring any reflection from other surfaces nearby). Thus, our artificial surface should be black. We also assume that, with the exception of the (parallel) solar irradiation, no external radiation enters the cavity. Since the solar irradiation is best treated separately through the external irradiation term H_o, our artificial surface is nonemitting. Both criteria are satisfied by covering the groove with a black surface at 0 K (A_3). Even though we now have three surfaces, the last one does not really appear in Equation (4.3.18) (since $E_{b3} = 0$ and $1/\varepsilon_3 - 1 = 0$):

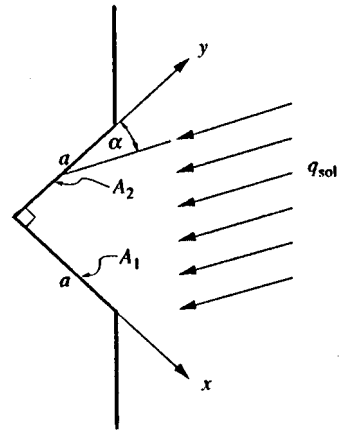


FIGURE 4.3.12 Right-angled groove exposed to solar irradiation, Example 4.3.4.

$$q_1 = E_{b1} - F_{1-2}E_{b2} - H_{o1} = \sigma T^4\left(1 - F_{1-2}\right) - q_{sol}\cos\alpha$$

$$q_2 = E_{b2} - F_{2-1}E_{b1} - H_{o2} = \sigma T^4\left(1 - F_{2-1}\right) - q_{sol}\sin\alpha$$

From the crossed-strings method, Equation (4.3.15), we find

$$F_{1-2} = \frac{a + a - \left(\sqrt{2}a + 0\right)}{2a} = \frac{1}{2}\left(2 - \sqrt{2}\right) = 0.293 = F_{2-1}$$

and

$$Q' = a\left(q_1 + q_2\right) = a\left[\sqrt{2}\sigma T^4 - q_{sol}\left(\cos\alpha + \sin\alpha\right)\right]$$

Example 4.3.5

Consider a very long duct as shown in Figure 4.3.13. The duct is 30 × 40 cm in cross section, and all surfaces are covered with gray, diffuse surface material. Top and bottom walls are at $T_1 = T_3 = 1000$ K with $\varepsilon_1 = \varepsilon_3 = 0.3$, while the side walls are at $T_2 = T_4 = 600$ K with $\varepsilon_2 = \varepsilon_4 = 0.8$ as shown. Determine the net radiative heat transfer rates for each surface.

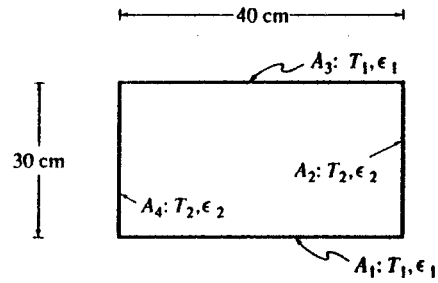


FIGURE 4.3.13 Two-dimensional gray, diffuse duct for Example 4.3.5.

Solution. Using Equation (4.3.18) for $i = 1$ and $i = 2$ and noting that $F_{1-2} = F_{1-4}$ and $F_{2-1} = F_{2-3}$,

$$i = 1: \quad \frac{q_1}{\varepsilon_1} - 2\left(\frac{1}{\varepsilon_2} - 1\right)F_{1-2}q_2 - \left(\frac{1}{\varepsilon_1} - 1\right)F_{1-3}q_1 = 2F_{1-2}\left(E_{b1} - E_{b2}\right)$$

$$i = 2: \quad \frac{q_2}{\varepsilon_2} - 2\left(\frac{1}{\varepsilon_1} - 1\right)F_{2-1}q_1 - \left(\frac{1}{\varepsilon_2} - 1\right)F_{2-4}q_2 = 2F_{2-1}\left(E_{b2} - E_{b1}\right)$$

The view factors are readily evaluated from the crossed-strings method as $F_{1-2} = {}^1/_4$, $F_{1-3} = 1 - 2F_{1-2} = {}^1/_2$, $F_{2-1} = {}^4/_3\, F_{1-2} = {}^1/_3$ and $F_{2-4} = 1 - 2F_{2-1} = {}^1/_3$. Substituting these, as well as emissivity values, into the relations reduces them to the simpler form of

$$\left[\frac{1}{0.3} - \left(\frac{1}{0.3} - 1\right)\frac{1}{2}\right]q_1 - 2\left(\frac{1}{0.8} - 1\right)\frac{1}{4}q_2 = 2 \times \frac{1}{4}\left(E_{b1} - E_{b2}\right)$$

$$-2\left(\frac{1}{0.3} - 1\right)\frac{1}{3}q_1 + \left[\frac{1}{0.8} - \left(\frac{1}{0.8} - 1\right)\right]\frac{1}{3}q_2 = 2 \times \frac{1}{3}\left(E_{b2} - E_{b1}\right)$$

or

$$\frac{13}{6}q_1 - \frac{1}{8}q_2 = \frac{1}{2}\left(E_{b1} - E_{b2}\right)$$

$$-\frac{14}{9}q_1 + \frac{7}{6}q_2 = -\frac{2}{3}\left(E_{b1} - E_{b2}\right)$$

Thus,

$$\left(\frac{13}{6} \times \frac{7}{6} - \frac{14}{9} \times \frac{1}{8}\right)q_1 = \left(\frac{1}{2} \times \frac{7}{6} - \frac{2}{3} \times \frac{1}{8}\right)\left(E_{b1} - E_{b2}\right)$$

$$q_1 = \frac{3}{7} \times \frac{1}{2}\left(E_{b1} - E_{b2}\right) = \frac{3}{14}\sigma\left(T_1^4 - T_2^4\right)$$

and

$$\left(-\frac{1}{8} \times \frac{14}{9} + \frac{7}{6} \times \frac{13}{6}\right)q_2 = \left(\frac{1}{2} \times \frac{14}{9} - \frac{2}{3} \times \frac{13}{6}\right)\left(E_{b1} - E_{b2}\right)$$

$$q_2 = \frac{3}{7} \times \frac{2}{3}\left(E_{b1} - E_{b2}\right) = -\frac{2}{7}\sigma\left(T_1^4 - T_2^4\right)$$

Finally, substituting values for temperatures,

$$Q_1' = 0.4\ \text{m} \times \frac{3}{14} \times 5.670 \times 10^{-8}\ \frac{\text{W}}{\text{m}^2\text{K}^4}\left(1000^4 - 600^4\right)\text{K}^4 = 4230\ \text{W/m}$$

$$Q_2' = -0.3\ \text{m} \times \frac{2}{7} \times 5.670 \times 10^{-8}\ \frac{\text{W}}{\text{m}^2\text{K}^4}\left(1000^4 - 600^4\right)\text{K}^4 = -4230\ \text{W/m}$$

Note that, for conservation of energy, both heat transfer rates must add up to zero.

Small Body Inside Isothermal Enclosure. An especially simple — but important — case occurs if a small, convex body A_1 (i.e., a surface that cannot "see" itself, or $F_{1-1} = 0$) is totally enclosed by an isothermal enclosure A_2. Then, with $N = 2$ and $F_{1-2} = 1$, Equation (4.3.18) reduces to

$$q_1 = \frac{E_{b1} - E_{b2}}{\dfrac{1}{\varepsilon_1} + \dfrac{A_1}{A_2}\left(\dfrac{1}{\varepsilon_2} - 1\right)} = \frac{\sigma\left(T_1^4 - T_2^4\right)}{\dfrac{1}{\varepsilon_1} + \dfrac{A_1}{A_2}\left(\dfrac{1}{\varepsilon_2} - 1\right)} \tag{4.3.22}$$

If the enclosure is large, i.e., $A_1 \ll A_2$, then Equation (4.3.22) simplifies further to

$$q_1 = \varepsilon_1 \sigma\left(T_1^4 - T_2^4\right) \tag{4.3.23}$$

Radiation Shields. If it is desired to minimize radiative heat transfer between two surfaces, it is common practice to place one or more radiation shields between them (usually thin metallic sheets of low emissivity). If two surfaces A_i and A_j are close together, so that $A_i \cong A_j$ and $F_{i-j} \cong 1$, then the radiative exchange between them is, from Equation (4.3.22),

$$q = \frac{E_{bi} - E_{bj}}{R_{ij}}, \qquad R_{ij} = \frac{1}{\varepsilon_i} + \frac{1}{\varepsilon_j} - 1 \tag{4.3.24}$$

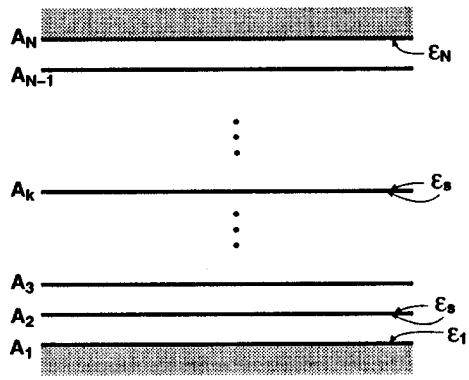


FIGURE 4.3.14 Placement of radiation shields between two large, parallel plates.

where R_{ij} is termed the *radiative resistance*. Equation (4.3.24) is seen to be analogous to an electrical circuit with "current" q and "voltage potential" $E_{bi} - E_{bj}$. Therefore, expressing radiative fluxes in terms of radiative resistances is commonly known as **network analogy**. The network analogy is a very powerful method of solving one-dimensional problems (i.e., whenever only two isothermal surfaces see each other, such as infinite parallel plates, or when one surface totally encloses another). Consider, for example, two large parallel plates, A_1 and A_N, separated by $N - 2$ radiation shields, as shown in Figure 4.3.14. Let each shield have an emissivity ε_S on both sides. Then, by applying Equation (4.3.24) to any two consecutive surfaces and using the fact that q remains constant throughout the gap,

$$q = \frac{E_{b1} - E_{b2}}{R_{12}} = \cdots = \frac{E_{bk-1} - E_{bk}}{R_{k-1,k}} = \cdots = \frac{E_{bN-1} - E_{bN}}{R_{N-1,N}} = \frac{E_{b1} - E_{bN}}{\sum_{j-2}^{N} R_{j-1,j}} \tag{4.3.25}$$

where

$$R_{j-1,j} = \frac{1}{\varepsilon_{j-1}} + \frac{1}{\varepsilon_j} - 1 \tag{4.3.26}$$

and, if $\varepsilon_2 = \varepsilon_3 \cdots = \varepsilon_{N-1} = \varepsilon_S$,

$$\sum_{j=2}^{N} R_{j-1,j} = \frac{1}{\varepsilon_1} + \frac{1}{\varepsilon_N} - 1 + (N-2)\left(\frac{2}{\varepsilon_S} - 1\right) \tag{4.3.27}$$

Equations (4.3.24) to (4.3.27) are also valid for concentric cylinders, concentric spheres, and similar configurations, as long as $r_N - r_1 \ll r_1$. Also, the relations are readily extended to shields with nonidentical emissivities.

While the network analogy can (and has been) applied to configurations with more than two surfaces seeing each other, this leads to very complicated circuits (since there is one resistance between any two surfaces). For such problems the network analogy is not recommended, and the net radiation method, Equation (4.3.18), should be employed.

Radiative Exchange within Participating Media

In many high-temperature applications, when radiative heat transfer is important, the medium between surfaces is not transparent, but is "participating," i.e., it absorbs, emits, and (possibly) scatters radiation. In a typical combustion process this interaction results in (1) continuum radiation due to tiny, burning soot particles (of dimension <1 μm) and also due to larger suspended particles, such as coal particles, oil droplets, fly ash; (2) banded radiation in the infrared due to emission and absorption by molecular gaseous combustion products, mostly water vapor and carbon dioxide; and (3) chemiluminescence due to the combustion reaction itself. While chemiluminescence may normally be neglected, particulates as well as gas radiation generally must be accounted for.

Radiative Properties of Molecular Gases

When a photon (or an electromagnetic wave) interacts with a gas molecule, it may be absorbed, raising the energy level of the molecule. Conversely, a gas molecule may spontaneously lower its energy level by the emission of an appropriate photon. This leads to large numbers of narrow spectral lines, which partially overlap and together form so-called vibration-rotation bands. As such, gases tend to be transparent over most of the spectrum, but may be almost opaque over the spectral range of a band. The **absorption coefficient** κ_λ is defined as a measure of how strongly radiation is absorbed or emitted along a path of length, L, leading to the spectral absorptivity and emissivity for this path, or

$$\alpha_\lambda = \varepsilon_\lambda = 1 - e^{-\kappa_\lambda L} \tag{4.3.28}$$

Although gases are distinctly nongray, for simple heat transfer calculations we often need to determine the total emissivity for an isothermal path (compare Equation (4.3.9))

$$\varepsilon = \frac{1}{E_b}\int_0^\infty \left(1 - e^{-\kappa_\lambda L}\right) E_{b\lambda}\left(T_g\right) d\lambda \tag{4.3.29}$$

For a mixture of gases the total emissivity is a function of path length L, gas temperature T_g, partial pressure(s) of the absorbing gas(es) p_a, and total pressure p. For the — in combustion applications most important — mixture of nitrogen with water vapor and/or carbon dioxide, the total emissivity may be calculated from Leckner.[8] First, the individual emissivities for water vapor and carbon dioxide, respectively, are calculated separately from

$$\varepsilon\left(p_aL,p,T_g\right)=\varepsilon_0\left(p_aL,T_g\right)\left(\frac{\varepsilon}{\varepsilon_0}\right)\left(p_aL,p,T_g\right) \tag{4.3.30a}$$

$$\left(\frac{\varepsilon}{\varepsilon_0}\right)\left(p_aL,p,T_g\right)=\left[1-\frac{(a-1)\left(1-P_E\right)}{a+b-1+P_E}\exp\left(-c\left[\log_{10}\frac{\left(p_aL\right)_m}{p_aL}\right]^2\right)\right] \tag{4.3.30b}$$

$$\varepsilon_0\left(p_aL,T_g\right)=\exp\left[\sum_{i=0}^{N}\sum_{j=0}^{N}c_{ji}\left(\frac{T_g}{T_0}\right)^j\left(\log_{10}\frac{p_aL}{\left(p_aL\right)_0}\right)^i\right] \tag{4.3.30c}$$

Here ε_0 is the total emissivity of a reference state, i.e., for the case of p = 1 bar and $p_a \to 0$ (but $p_aL >$ 0), and the correlation constants a,b,c,c_{ji}, P_E $(p_aL)_0$, $(p_aL)_m$, and T_0 are given in Table 4.3.2 for water vapor and carbon dioxide. (For convenience, plots of ε_0 are given in Figures 4.3.15 for CO_2 and 4.3.16 for H_2O.) The total emissivity of a mixture of nitrogen with both water vapor and carbon dioxide is calculated from

$$\varepsilon_{CO_2+H_2O}=\varepsilon_{CO_2}+\varepsilon_{H_2O}-\Delta\varepsilon \tag{4.3.31}$$

TABLE 4.3.2 Correlation Constants for the Determination of the Total Emissivity for Water Vapor and Carbon Dioxide

Gas	Water Vapor			Carbon Dioxide			
M, N	2,2			2,3			
c_{00} … c_{N1}	–2.2118	–1.1987	0.035596	–3.9893	2.7669	–2.1081	0.39163
⋮ ⋱ ⋮	0.85667	0.93048	–0.14391	1.2710	–1.1090	1.0195	–0.21897
c_{0M} … c_{NM}	–0.10838	–0.17156	0.045915	–0.23678	0.19731	–0.19544	0.044644
P_E	$(p+2.56p_a/\sqrt{t})/p_0$			$(p+0.28p_a)/p_0$			
$(p_aL)_m/(p_aL)_0$	$13.2t^2$			$0.054/t^2$, $t<0.7$ $0.225t^2$, $t>0.7$			
a	2.144, $t<0.75$ $1.88-2.053\log_{10}t$, $t>0.75$			$1+0.1/t^{1.45}$			
b	$1.10/t^{1.4}$			0.23			
c	0.5			1.47			

Note: T_0 = 1000 K, p_0 = 1 bar, $t = T/T_0$, $(p_a L)_0$ = 1 bar cm.

$$\Delta\varepsilon = \left(\frac{\zeta}{10.7+101\zeta} - 0.0089\zeta^{10.4}\right)\left(\log_{10}\frac{\left(p_{H_2O}+p_{CO_2}\right)L}{\left(p_aL\right)_0}\right)^{2.76} \qquad (4.3.32)$$

$$\zeta = \frac{p_{H_2O}}{p_{H_2O}+p_{CO_2}}$$

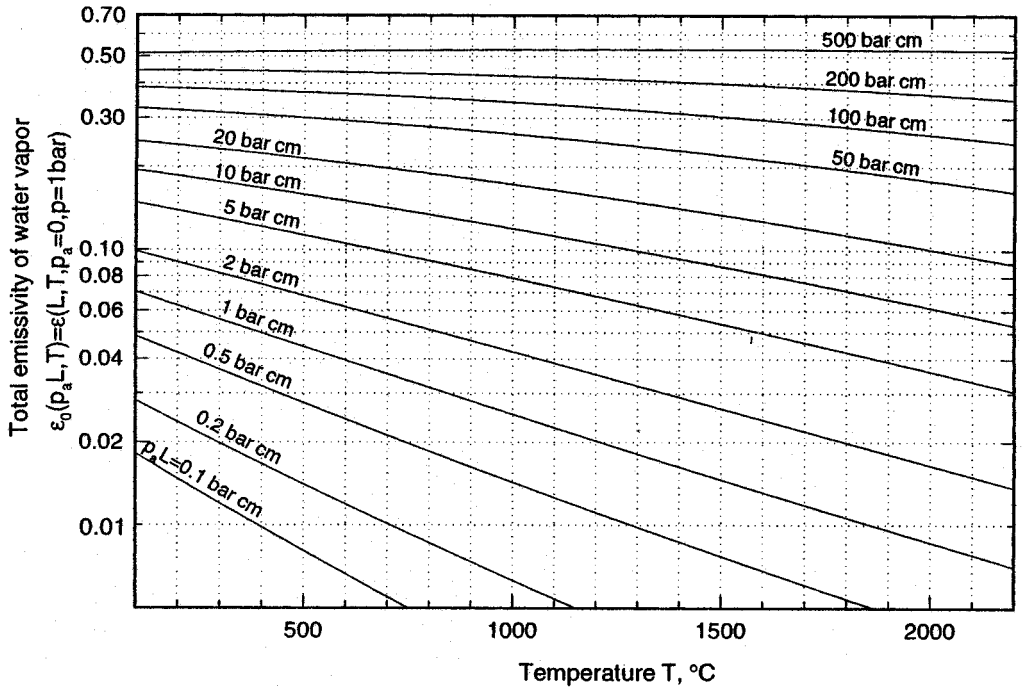


FIGURE 4.3.15 Total emissivity of water vapor at reference state (total gas pressure p = 1 bar, partial pressure of H_2O $p_a \to 0$).

where the $\Delta\varepsilon$ compensates for overlap effects between H_2O and CO_2 bands, and the ε_{CO_2} and ε_{H_2O} are calculated from Equation (4.3.30).

If radiation emitted externally to the gas (for example, by emission from an adjacent wall at temperature T_s) travels through the gas, the total amount absorbed by the gas is of interest. This leads to the absorptivity of a gas path at T_g with a source at T_s:

$$\alpha\left(p_aL,p,T_g,T_s\right) = \frac{1}{E_b\left(T_s\right)}\int_0^\infty\left(1-e^{-\kappa_\lambda\left(T_g\right)L}\right)E_{b\lambda}\left(T_s\right)d\lambda \qquad (4.3.33)$$

which for water vapor or carbon dioxide may be estimated from

$$\alpha\left(p_aL,p,T_g,T_s\right) = \left(\frac{T_g}{T_s}\right)^{1/2}\varepsilon\left(p_aL\frac{T_s}{T_g},p,T_s\right) \qquad (4.3.34)$$

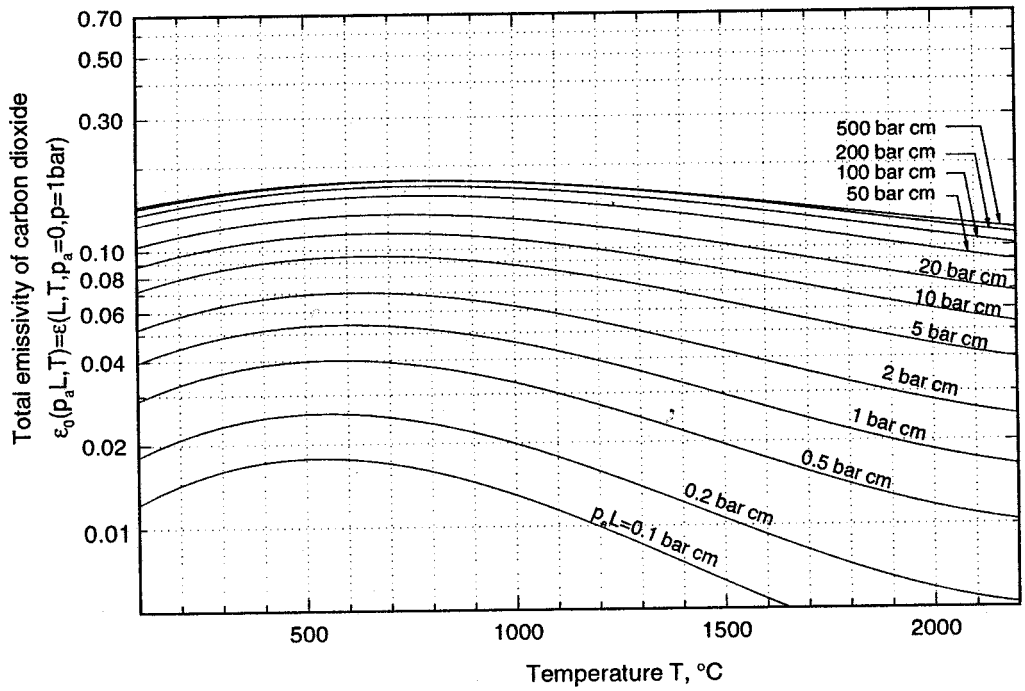


FIGURE 4.3.16 Total emissivity of carbon dioxide at reference state (total gas pressure p = 1 bar, partial pressure of CO_2 $p_a \to 0$).

where ε is the emissivity calculated from Equation (4.3.30) evaluated at the temperature of the surface T_s, and using an adjusted pressure path length, p_aLT_s/T_g. For mixtures of water vapor and carbon dioxide band overlap is again accounted for by taking

$$\alpha_{CO_2+H_2O} = \alpha_{CO_2} + \alpha_{H_2O} - \Delta\varepsilon \tag{4.3.35}$$

with Δε evaluated for a pressure path length of p_aLT_s/T_g.

Example 4.3.6

Consider a layer of a gas mixture at 1000 K and 5 bar that consists of 10% carbon dioxide and 70% nitrogen. What is its emissivity for a path length of 1.76 m, and its absorptivity (for the same path) if the layer is irradiated by a source at 1500 K?

Solution. First we calculate the total emissivity of the CO_2 at the reference state (p = 1 bar, $p_a \to 0$). for a length of 1.76 m from Equation (4.3.30c) or Figure 4.3.15. With

$$T_g = 1000 \text{ K} = 727°\text{C} \quad \text{and} \quad p_aL = 0.1 \times 5 \text{ bar} \times 1.76 \text{ m} = 88 \text{ bar cm}$$

one gets, interpolating Figure 4.3.15, $\varepsilon_0 \cong 0.15$. The correction factor in Equation (4.3.30b) is calculated from Table 4.3.2 with P_E = 5 + 0.28 × 0.5 = 5.14, a = 1.1, b = 0.23, c = 1.47, and $(p_aL)_m$ = 0.225 bar cm. Thus,

$$\frac{\varepsilon}{\varepsilon_0} = 1 - \frac{0.1 \times (-4.14)}{0.33 + 5.14} \exp\left(-1.47\left(\log_{10}\frac{0.225}{88}\right)^2\right) \cong 1$$

and

$$\varepsilon \cong 0.15$$

To calculate the absorptivity ε_0 must be found for a temperature of

$$T_s = 1500 \text{ K} = 1227°\text{C} \quad \text{and} \quad p_a L \frac{T_s}{T_g} = 88 \times 1500/1000 = 132 \text{ bar cm}$$

From Figure 4.3.15 it follows that $\varepsilon_0 \cong 0.15$ again and, with $\varepsilon/\varepsilon_0$ pretty much unchanged, from Equation (4.3.34),

$$\alpha \cong \left(\frac{1000}{1500}\right)^{1/2} \times 0.15 \times 1.00 = 0.122$$

Radiative Properties of Particle Clouds

Nearly all flames are visible to the human eye and are, therefore, called *luminous* (sending out light). Apparently, there is some radiative emission from within the flame at wavelengths where there are no vibration–rotation bands for any combustion gases. This luminous emission is known today to come from tiny *char* (almost pure carbon) particles, call *soot*, which are generated during the combustion process. The "dirtier" the flame is (i.e., the higher the soot content), the more luminous it is.

Radiative Properties of Soot. Soot particles are produced in fuel-rich flames, or fuel-rich parts of flames, as a result of incomplete combustion of hydrocarbon fuels. As shown by electron microscopy, soot particles are generally small and spherical, ranging in size between approximately 50 and 800 Å (0.005 to 0.08 μm), and up to about 3000 Å in extreme cases. While mostly spherical in shape, soot particles may also appear in agglomerated chunks and even as long agglomerated filaments. It has been determined experimentally in typical diffusion flames of hydrocarbon fuels that the volume percentage of soot generally lies in the range between 10^{-4} to 10^{-6}%.

Since soot particles are very small, they are generally at the same temperature as the flame and, therefore, strongly emit thermal radiation in a continuous spectrum over the infrared region. Experiments have shown that soot emission often is considerably stronger than the emission from the combustion gases.

For a simplified heat transfer analysis it is desirable to use suitably defined mean absorption coefficients and emissivities. If the soot volume fraction f_v is known as well as an appropriate spectral average of the complex index of refraction of the soot, $m = n - ik$, one may approximate the spectral absorption coefficient by (Felske and Tien, 1977).

$$\kappa_\lambda = C_0 \frac{f_v}{\lambda} \qquad C_0 = \frac{36\pi nk}{\left(n^2 - k^2 + 2\right)^2 + 4n^2k^2} \tag{4.3.36}$$

and a total, or spectral-average value may be taken as

$$\kappa_m = 3.72\, f_v C_0 T / C_2 \tag{4.3.37}$$

where C_2 = 1.4388 mK is the second Planck function constant. Substituting Equation (4.3.37) into Equation (4.3.29) gives a total soot cloud emissivity of

$$\varepsilon\left(f_v TL\right) = 1 - e^{-\kappa_m L} = 1 - e^{-3.72 C_0 f_v TL/C_2} \tag{4.3.38}$$

Pulverized Coal and Fly Ash Dispersions. To calculate the radiative properties of arbitrary size distributions of coal and ash particles, one must have knowledge of their complex index of refraction as a function of wavelength and temperature. Data for carbon and different types of coal indicate that its real part, n, varies little over the infrared and is relatively insensitive to the type of coal (e.g., anthracite, lignite, bituminous), while the absorptive index, k, may vary strongly over the spectrum and from coal to coal. If the number and sizes of particles are known and if a suitable average value for the complex index of refraction can be found, then the spectral absorption coefficient of the dispersion may be estimated by a correlation given by Buckius and Hwang, 1980. Substitution into Equation (4.3.29) can then provide an estimate of the total emissivity. If both soot as well as larger particles are present in the dispersion, the absorption coefficients of all constituents must be added before applying Equation (4.3.29).

Mixtures of Molecular Gases and Particulates. To determine the total emissivity of a mixture it is generally necessary to find the spectral absorption coefficient κ_λ of the mixture (i.e., the sum of the absorption coefficient of all contributors), followed by numerical integration of Equation (4.3.29). However, since the molecular gases tend to absorb only over a small part of the spectrum, to some degree of accuracy

$$\varepsilon_{\text{mix}} \cong \varepsilon_{\text{gas}} + \varepsilon_{\text{particulates}} \tag{4.3.39}$$

Equation (4.3.39) gives an upper estimate since overlap effects result in lower emissivity (compare Equation (4.3.31) for gas mixtures).

Heat Exchange in the Presence of a Participating Medium

The calculation of radiative heat transfer rates through an enclosure filled with a participating medium is a challenging task, to say the least. High-accuracy calculations are rare and a topic of ongoing research. There are, however, several simplistic models available that allow the estimation of radiative heat transfer rates, and relatively accurate calculations for some simple cases.

Diffusion Approximation. A medium through which a photon can only travel a short distance without being absorbed is known as *optically thick*. Mathematically, this implies that $\kappa_\lambda L \gg 1$ for a characteristic dimension L across which the temperature does not vary substantially. For such an optically thick, nonscattering medium the spectral radiative flux may be calculated from

$$\mathbf{q}_\lambda = -\frac{4}{3\kappa_\lambda}\nabla E_{b\lambda} \tag{4.3.40}$$

similar to Fourier's diffusion law for heat conduction. Note that a medium may be optically thick at some wavelengths, but thin ($\kappa_\lambda L \ll 1$) at others (e.g., molecular gases!). For a medium that is optically thick for all wavelengths, Equation (4.3.40) may be integrated over the spectrum, yielding the total radiative flux

$$\mathbf{q} = -\frac{4}{3\kappa_R}\nabla E_b = -\frac{4}{3\kappa_R}\nabla\left(\sigma T^4\right) = -\frac{16\sigma T^3}{3\kappa_R}\nabla T \tag{4.3.41}$$

where κ_R is the suitably averaged absorption coefficient, termed the *Rosseland-mean absorption coefficient*. For a cloud of soot particles, $\kappa_R \cong \kappa_m$ from Equation (4.3.37) is a reasonable approximation. Equation (4.3.41) may be rewritten by defining a "radiative conductivity" k_R,

$$\mathbf{q} = -k_R \nabla T \qquad k_R = \frac{16\sigma T^3}{3\kappa_R} \tag{4.3.42}$$

This form shows that the diffusion approximation is mathematically equivalent to conductive heat transfer with a (strongly) temperature-dependent conductivity.

Note: More accurate calculations show that, in the absence of other modes of heat transfer (conduction, convection), there is generally a temperature discontinuity near the boundaries ($T_{\text{surface}} \neq T_{\text{adjacent medium}}$), and, unless boundary conditions that allow such temperature discontinuities are chosen, the diffusion approximation will do very poorly in the vicinity of bounding surfaces.

Example 4.3.7

A soot cloud is contained between two walls at T_1 = 1000 K and T_2 = 2000 K, spaced 1 m apart. The effective absorption coefficient of the medium is κ_R = 10 m^{-1} and the effective thermal conductivity is k_c = 0.1 W/mK. Estimate the total heat flux between the plates (ignoring convection effects).

Solution. For simplicity we may want to assume a constant total conductivity $k = k_c + k_R$, leading to

$$q = -k\frac{dT}{dx} = k\frac{T_2 - T_1}{L}$$

where k_R must be evaluated at some effective temperature. Choosing, based on its temperature dependence,

$$k_R \cong \frac{8\sigma}{3\kappa_R}\left(T_1^3 + T_2^3\right) = \frac{8 \times 5.670 \times 10^{-8}\ \text{W/m}^2\text{K}^4}{3 \times 10/\text{m}}\left(1000^3 + 2000^3\right)\text{K}^3 = 136\frac{\text{W}}{\text{mK}}$$

gives

$$q = (0.1 + 136)\frac{2000 - 1000}{1}\frac{\text{W}}{\text{m}^2} = 136\frac{\text{kW}}{\text{m}^2\text{K}}$$

Note that (1) conduction is negligible in this example and (2) the surface emissivities do not enter the diffusion approximation. While a more accurate answer can be obtained by taking the temperature dependence of k_R into account, the method itself should be understood as a relatively crude approximation.

Mean Beam Length Method. Relatively accurate yet simple heat transfer calculations can be carried out if an isothermal, absorbing–emitting, but not scattering medium is contained in an isothermal, black-walled enclosure. While these conditions are, of course, very restrictive, they are met to some degree by conditions inside furnaces. For such cases the local heat flux on a point of the surface may be calculated from

$$q = \left[1 - \alpha\left(L_m\right)\right]E_{bw} - \varepsilon\left(L_m\right)E_{bg} \tag{4.3.43}$$

where E_{bw} and E_{bg} are blackbody emissive powers for the walls and medium (gas and/or particulates), respectively, and $\alpha(L_m)$ and $\varepsilon(L_m)$ are the total absorptivity and emissivity of the medium for a path length L_m through the medium. The length L_m, known as the average *mean beam length*, is a directional average of the thickness of the medium as seen from the point on the surface. On a spectral basis Equation (4.3.43) is exact, provided the above conditions are met and provided an accurate value of the (spectral) mean beam length is known. It has been shown that spectral dependence of the mean beam length is weak (generally less than ±5% from the mean). Consequently, total radiative heat flux at the

surface may be calculated very accurately from Equation (4.3.43), provided the emissivity and absorptivity of the medium are also known accurately. The mean beam lengths for many important geometries have been calculated and are collected in Table 4.3.3. In this table L_o is known as the geometric mean beam length, which is the mean beam length for the optically thin limit ($\kappa_\lambda \to 0$), and L_m is a spectral average of the mean beam length. For geometries not listed in Table 4.3.3, the mean beam length may be estimated from

$$L_o \cong 4\frac{V}{A} \qquad L_m \cong 0.9L_o \cong 3.6\frac{V}{A} \tag{4.3.44}$$

TABLE 4.3.3 Mean Beam Lengths for Radiation from a Gas Volume to a Surface on Its Boundary

Geometry of Gas Volume	Characterizing Dimension L	Geometric Mean Beam Length L_0/L	Average Mean Beam Length L_m/L	L_m/L_0
Sphere radiating to its surface	Diameter, L = D	0.67	0.65	0.97
Infinite circular cylinder to bounding surface	Diameter, L = D	1.00	0.94	0.94
Semi-infinite circular cylinder to:	Diameter, L = D			
Element at center of base		1.00	0.90	0.90
Entire base		0.81	0.65	0.80
Circular cylinder (height/diameter = 1) to:	Diameter, L = D			
Element at center of base		0.76	0.71	0.92
Entire surface		0.67	0.60	0.90
Circular cylinder (height/diameter = 2) to:	Diameter, L = D			
Plane base		0.73	0.60	0.82
Concave surface		0.82	0.76	0.93
Entire surface		0.80	0.73	0.91
Circular cylinder (height/diameter = 0.5) to:	Diameter, L = D			
Plane base		0.48	0.43	0.90
Concave surface		0.53	0.46	0.88
Entire surface		0.50	0.45	0.90
Infinite semicircular cylinder to center of plane rectangular face	Radius, L = R	—	1.26	—
Infinite slab to its surface	Slab thickness, L	2.00	1.76	0.88
Cube to a face	Edge L	0.67	0.6	0.90
Rectangular $1 \times 1 \times 4$ parallelepipeds:	Shortest edge, L			
To 1×4 face		0.90	0.82	0.91
To 1×1 face		0.86	0.71	0.83
To all faces		0.89	0.81	0.91

where V is the volume of the participating medium and A is its entire bounding surface area.

Example 4.3.8

An isothermal mixture of 10% CO_2 and 90% nitrogen at 1000 K and 5 bar is contained between two large, parallel, black plates, which are both isothermal at 1500 K. Estimate the net radiative heat loss from the surfaces.

Solution. The heat loss may be calculated from Equation (4.3.43), after determining the mean beam length, followed by evaluation of $\varepsilon(L_m)$ and $\alpha(L_m)$. From Table 4.3.3 it is clear that $L_m = 1.76 \times$ thickness of slab = 1.76 m. It turns out that the necessary $\varepsilon(L_m) = 0.15$ and $\alpha(L_m) = 0.122$ have already been calculated in Example 4.3.6. Thus, the heat flux is immediately calculated from Equation (4.3.43) as

$$q = (1 - 0.122)5.670 \times 10^{-8} \times 1500^4 - 0.15 \times 5.670 \times 10^{-8} \times 1000^4$$

$$= 2.44 \times 10^5 \frac{\text{W}}{\text{m}^2} = 244 \text{ kW/m}^2$$

Defining Terms

Absorptivity: The ability of a medium to absorb (i.e., trap and convert to other forms of energy) incoming radiation; gives the fraction of incoming radiation that is absorbed by the medium.

Absorption coefficient: The ability of a medium to absorb (i.e., trap and convert to other forms of energy) over a unit path length; the reciprocal of the mean distance a photon travels before being absorbed.

Blackbody: Any material or configuration that absorbs all incoming radiation completely. A blackbody also emits the maximum possible amount of radiation as described by Planck's law.

Diffuse surface: A surface that emits and/or reflects equal amounts of radiative energy (photons) into all directions. Or a surface that absorbs and/or reflects equal amounts of radiation independent of incoming direction.

Emissive power: The rate of radiative energy leaving a surface through emission. The maximum amount of emissive power is emitted by a blackbody with a spectral strength described by Planck's law.

Emissivity: The ability of a medium to emit (i.e., convert internal energy into electromagnetic waves or photons) thermal radiation; gives the fraction of emission as compared with a blackbody.

Gray: A medium whose radiative properties (such as absorptivity, emissivity, reflectivity, absorption coefficient) do not vary with wavelength.

Irradiation: Incoming radiative flux onto a surface from outside it.

Network analogy: Expressing radiative heat exchange between surfaces in terms of an electrical network, with heat flux as "current," differences in emissive power as "potentials," and defining radiative resistances.

Opaque medium: A medium of sufficient thickness that absorbs all nonreflected irradiation; no radiation is transmitted through the medium.

Photon: A massless particle carrying energy in the amount of $h\nu$; the quantum mechanical alternative view of an electromagnetic wave carrying radiative energy.

Planck's law: The law describing the spectral distribution of the radiative energy emitted (emissive power) of a blackbody.

Radiosity: Total radiative flux leaving a surface (diffusely), consisting of emitted as well as reflected radiation.

Reflectivity: The ability of an interface, or of a medium or of a composite with a number of interfaces, to reflect incoming radiation back into the irradiating medium.

Semitransparent: See **transparent**.

Spectral value: The value of a quantity that varies with wavelength at a given wavelength; for dimensional quantities the amount per unit wavelength.

Transmissivity: The ability of a medium to let incoming radiation pass through it; gives the fraction of incoming radiation that is transmitted through the medium.

Transparent: The ability of a medium to let incoming radiation pass through it. A medium that lets all radiation pass through it is called transparent, a medium that only allows a part to pass through it is called **semitransparent**.

View factor: The fraction of diffuse radiant energy leaving one surface that is intercepted by another surface.

References

1. Brewster, M.Q. 1992. *Thermal Radiative Transfer & Properties*, John Wiley & Sons, New York.
2. Buckius, R.O. and Hwang, D.C. 1980. Radiation properties for polydispersions: application to coal, *J. Heat Transfer*, 102, 99–103.
3. Felske, J.D. and Tien, C.L. 1977. The use of the Milne-Eddington absorption coefficient for radiative heat transfer in combustion systems, *J. Heat Transfer*, 99(3), 458–465.
4. Hottel, H.C. and Sarofim, A.F. 1967. *Radiation Transfer*, McGraw-Hill, New York.
5. Howell, J.R. 1982. *Catalog of Radiation Configuration Factors*, McGraw-Hill, New York.
6. Leckner, B. 1972. Spectral and total emissivity of water vapor and carbon dioxide, *Combust. Flame*, 19, 33–48.
7. Modest, M.F. 1993. *Radiative Heat Transfer*, McGraw-Hill, New York.
8. Ozisik, M.N. 1973. *Radiative Transfer and Interactions with Conduction and Convection*, John Wiley & Sons, New York.
9. Siegel, R. and Howell, J.R. 1992. *Thermal Radiation Heat Transfer*, 3rd ed., Hemisphere Publishing, New York.
10. Sparrow, E.M. and Cess, R.D. 1978. *Radiation Heat Transfer*, Hemisphere, New York.

4.4 Phase-Change

Boiling and Condensation

Van P. Carey

Introduction

Liquid-vapor phase-change processes play an important role in many technological applications. The virtually isothermal heat transfer associated with boiling and condensation processes makes their inclusion in power and refrigeration processes highly advantageous from a thermodynamic efficiency standpoint. In addition, the high heat transfer coefficients associated with boiling and condensation have made the use of these processes increasingly attractive in the thermal control of compact devices that have high heat dissipation rates. Applications of this type include the use of boiling heat transfer to cool electronic components in computers and the use of compact evaporators and condensers for thermal control of aircraft avionics and spacecraft environments. Liquid-vapor phase-change processes are also of critical importance to nuclear power plant design, both because they are important in normal operating circumstances and because they dominate many of the accident scenarios that are studied as part of design evaluation.

The heat transfer and fluid flow associated with liquid-vapor phase-change processes are typically among the more complex transport circumstances encountered in engineering applications. These processes have all the complexity of single-phase convective transport, plus additional elements resulting from motion of the interface, nonequilibrium effects, and dynamic interactions between the phases. Due to the highly complex nature of these processes, development of methods to accurately predict the associated heat and mass transfer is often a formidable task.

In this section, commonly used variables not defined in the nomenclature are as follows: q'' = surface heat flux, μ_l = liquid viscosity, μ_v = vapor viscosity, Pr_l = liquid Prandtl number, T_w = wall surface temperature, T_{sat} = saturation temperature, c_{pl} = liquid specific heat, k_v = vapor thermal conductivity, g = gravitational acceleration, and x = mass quality.

Boiling

Three mechanisms that play important roles in boiling processes are (1) surface tension effects, (2) surface wetting characteristics of the liquid, and (3) metastable phase stability.

Anyone who has watched small bubbles rise in a carbonated beverage or a pot of boiling water has undoubtedly noticed that the bubbles are almost perfectly spherical, as if an elastic membrane were present at the interface to pull the vapor into a spherical shape. This apparent interfacial tension or *surface tension* σ is equivalent to an energy stored in the interface region per unit area. The energy excess in this region is due to the slightly larger separation of the liquid phase molecules adjacent to the gas phase.

The magnitude of the surface tension for a substance is directly linked to the strength of intermolecular forces in the material. Nonpolar liquids typically have the lowest surface tension. Water and other **polar molecules** have somewhat higher surface tension, and liquid metals, which exhibit metallic bond attraction, have very high surface tension. The surface tension of water at 20°C is 0.0728 N/m, whereas liquid mercury has a surface tension of 0.484 N/m at the same temperature. The surface tension for any pure liquid varies with temperature. It decreases almost linearly with increasing temperature, vanishing altogether at the critical point where the distinction between the phases disappears.

As a result of the surface tension at the interface, the pressure inside a spherical bubble of radius r must exceed that in the surrounding liquid by $2\sigma/r$:

$$P_v = P_l + \frac{2\sigma}{r} \tag{4.4.1}$$

By using the relation (1) between the pressure in the two phases it can be shown that for the bubble to be in equilibrium with the surrounding liquid, the liquid must actually be superheated above the saturation temperature for the ambient liquid pressure. The amount of required superheating increases as the radius of curvature of the bubble interface decreases.

The wetting characteristics of the liquid are generally quantified in terms of a *contact angle* between the solid surface and the tangent to the interface at the point where it contacts the solid. This angle is measured through the liquid phase, as shown in Figure 4.4.1. In some systems, the wetting angle established at equilibrium may depend on the fluid motion history. For some systems the contact angle established by liquid advancing over a solid surface is larger than that established when a liquid front recedes over the surface. This behavior is referred to as *contact angle hysteresis*. Contact angle hysteresis can have an important effect on boiling and condensation processes, particularly those involving water.

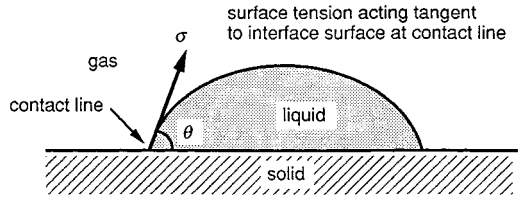


FIGURE 4.4.1 Definition of the contact angle θ.

For a bubble with a specified vapor volume, the contact angle will dictate the radius of curvature of the bubble interface. The wetting behavior in combination with the surface tension effect, thus, determines the level of superheat required for the bubble to be in equilibrium with the surrounding liquid. The liquid must be heated above this superheat level for the bubble to grow. A steady boiling process can be sustained only if the liquid is heated above this threshold superheat level.

It can be shown from basic thermodynamic analysis that a necessary and sufficient condition for phase stability is that

$$\left(\frac{\partial P}{\partial v}\right)_T < 0 \tag{4.4.2}$$

where v is the specific volume. Below the critical temperature, extrapolation of the isotherms for the liquid and vapor phases consistent with an equation of state like the van de Waals equation results in an isotherm shape similar to that shown in Figure 4.4.2.

The locus of points where $(\partial P/\partial v)_T = 0$ are termed *spinodal curves*. Regions of metastable vapor and liquid exist between the saturation curve and the spinodal curves. The effects of surface tension discussed above require that fluid surrounding a vapor bubble be in the metastable superheated liquid region. Predictions of statistical thermodynamics imply that as $(\partial P/\partial v)_T$ approaches zero, the level of fluctuations in a fluid system increases. This, in turn, increases the probability that an embryonic new phase will form as a result of density fluctuations. Initiation of a phase change in this manner is termed *homogeneous nucleation*. Generally, a pure liquid must be heated to nearly 90% of its absolute critical temperature before homogeneous nucleation of vapor bubbles occurs.

In most physical systems of engineering interest, the bulk phase is in contact with solid walls of the containing structures, or solid particulate contaminants. These solid phases may provide nucleation sites where phase change may occur if the system state is driven into the metastable range. Nucleation of vapor bubbles may preferentially occur at low liquid superheat levels in crevices in the solid surface where gas is trapped. This type of nucleation at the solid surface of a containment wall is categorized as *heterogeneous nucleation*. Because solid containment walls usually contain microscopic crevice-type

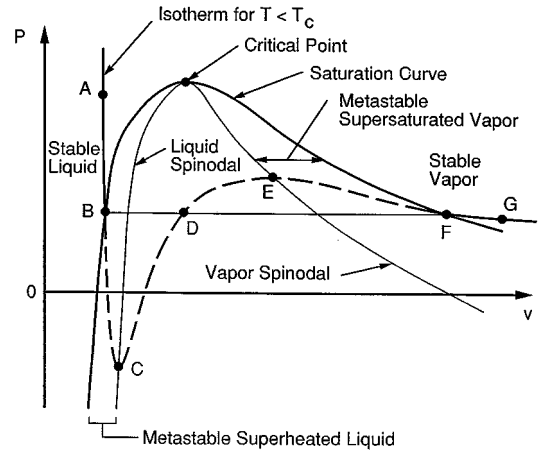


FIGURE 4.4.2 Spinodal lines and metastable regions on a *P–v* diagram.

imperfections, heterogeneous nucleation is more common than homogeneous nucleation in systems where boiling occurs.

Vapor entrapment in crevices of the heated walls of evaporator heat exchangers usually makes it easier to initiate the vaporization process. Vapor bubbles grow from these crevices until buoyancy or drag on the bubbles exceeds the surface tension force holding the droplet to the solid surface. The bubble then releases into the bulk liquid. A small remnant of vapor remains in the crevice after a bubble releases, and this remnant grows in size as further vaporization occurs until the bubble grows out of the crevice again. The result is a cyclic process of bubble growth and release known as the *ebullition cycle*. Crevices at which the ebullition cycle is sustained are said to be active nucleation sites. When the ebullition process occurs at many sites over a heated surface, the overall process is referred to as **nucleate boiling**, which is one possible mode of **pool boiling**.

Pool Boiling

Vaporization of liquid at the surface of a body immersed in an extensive pool of motionless liquid is generally referred to as pool boiling. The nature of the pool boiling process varies considerably depending on the conditions at which boiling occurs. The level of heat flux, the thermophysical properties of the liquid and vapor, the surface material and finish, and the physical size of the heated surface all may have an effect on the boiling process.

The regimes of pool boiling are most easily understood in terms of the so-called boiling curve: a plot of heat flux q'' vs. wall superheat $T_w - T_{sat}$ for the circumstances of interest. Many of the features of the classic pool boiling curve were determined in the early investigations of pool boiling conducted by Nukiyama (1934). Strictly speaking, the classic pool boiling curve defined by the work of this investigator and others applies to well-wetted surfaces for which the characteristic physical dimension L is large compared to the bubble or capillary length scale L_b defined as

$$L_b = \sqrt{\frac{\sigma}{g(\rho_l - \rho_v)}} \tag{4.4.3}$$

The discussion in this section is limited to pool boiling of wetting liquids on surfaces with dimensions large compared with L_b. Additional information on features of the boiling curve when the liquid poorly wets the surface or when L/L_b is small can be found in Carey (1992). To make this discussion concrete, we will assume that the ambient liquid surrounding the immersed body is at the saturation temperature for the ambient pressure. If the surface temperature of the immersed body is controlled and slowly

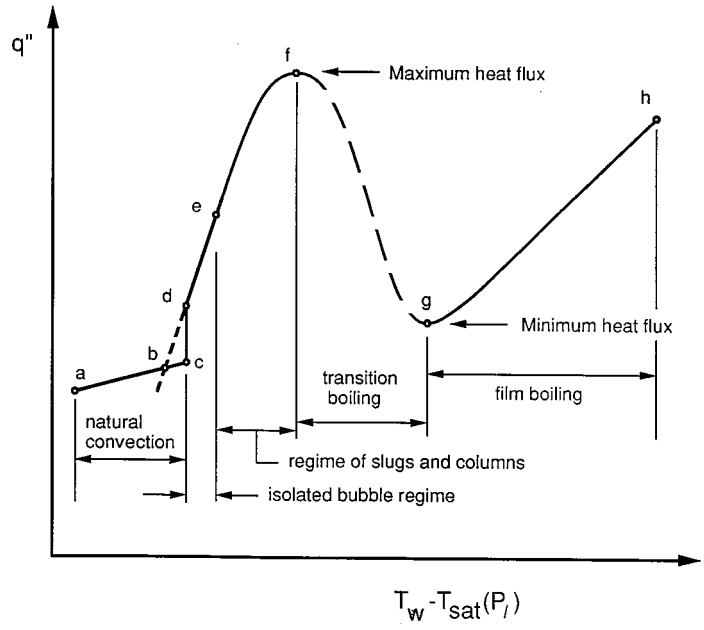


FIGURE 4.4.3 Pool boiling regimes for an independently controlled surface temperature.

increased, the boiling curve will look similar to that shown in Figure 4.4.3. The axes in this plot are logarithmic scales. The regimes of pool boiling encountered for an upward-facing horizontal flat surface as its temperature is increased are also indicated in Figure 4.4.3. The lateral extent of the surface is presumed to be much larger than L_b. At very low wall superheat levels, no nucleation sites may be active and heat may be transferred from the surface to the ambient liquid by natural convection alone and q'' increases slowly with $T_w - T_{sat}$.

Eventually, the superheat becomes large enough to initiate nucleation at some of the cavities on the surface. This *onset of nucleate boiling* (ONB) condition occurs at point *c* in Figure 4.4.3. Once nucleate boiling is initiated, any further increase in wall temperature causes the system operating point to move upward along section *d-f* of the curve in Figure 4.4.3. This portion of the curve corresponds to the nucleate boiling regime. The active sites are few and widely separated at low wall superheat levels. This range of conditions, corresponding to segment *d-e* of the curve, is sometimes referred to as the *isolated bubble regime.*

With increasing surface superheat, more sites become active, and the bubble frequency at each site generally increases. Eventually, the active sites are spaced so closely that bubbles from adjacent sites merge together during the final stages of growth and release. Vapor is being produced so rapidly that bubbles merging together form columns of vapor slugs that rise upward in the liquid pool toward its free surface. This higher range of wall superheat, corresponding to segment *e-f* of the boiling curve in Figure 4.4.3, is referred to as the *regime of slugs and columns.*

Increasing the wall superheat and heat flux within the regime of slugs and columns produces an increase in the flow rate of vapor away from the surface. Eventually, the resulting vapor drag on the liquid moving toward the surface becomes so severe that liquid is unable to reach the surface fast enough to keep the surface completely wetted with liquid. Vapor patches accumulate at some locations and evaporation of the liquid between the surface and some of these patches dries out portions of the surface.

If the surface temperature is held constant and uniform, dry portions of the surface covered with a vapor film will locally transfer a much lower heat flux than wetted portions of the surface where nucleate boiling is occurring. Because of the reduction in heat flux from intermittently dry portions of the surface, the mean overall heat flux from the surface is reduced. Thus, increasing the wall temperature within the

slugs and columns region ultimately results in a peaking and rollover of the heat flux. The peak value of heat flux is called the **critical heat flux** (CHF), designated as point *f* in Figure 4.4.3

If the wall temperature is increased beyond the critical heat flux condition, a regime is encountered in which the mean overall heat flux decreases as the wall superheat increases. This regime, which is usually referred to as the **transition boiling** regime, corresponds to segment *f-g* on the boiling curve shown in Figure 4.4.3. The transition boiling regime is typically characterized by rapid and severe fluctuations in the local surface heat flux and/or temperature values (depending on the imposed boundary condition). These fluctuations occur because the dry regions are generally unstable, existing momentarily at a given location before collapsing and allowing the surface to be rewetted.

The vapor film generated during transition boiling can be sustained for longer intervals at higher wall temperatures. Because the intermittent insulating effect of the vapor blanketing is maintained longer, the time-averaged contributions of the blanketed locations to the overall mean heat flux are reduced. The mean heat flux from the surface thus decreases as the wall superheat is increased in the transition regime. As this trend continues, eventually a point is reached at which the surface is hot enough to sustain a stable vapor film on the surface for an indefinite period of time. The entire surface then becomes blanketed with a vapor film, thus making the transition to the **film boiling** regime. This transition occurs at point *g* in Figure 4.4.3.

Within the film boiling regime, the heat flux monotonically increases as the superheat increases. This trend is a consequence of the increased conduction and/or convection transport due to the increased driving temperature difference across the vapor film. Radiative transport across the vapor layer may also become important at higher wall temperatures.

Once a surface is heated to a superheat level in the film boiling regime, if the surface temperature is slowly decreased, in general the system will progress through each of the regimes described above in reverse order. However, the path of the boiling curve may differ significantly from that observed for increasing wall superheat, depending on whether the surface heat flux or temperature is controlled.

Experimental evidence summarized by Witte and Lienhard (1982) implies that the path of the transition boiling curve is determined, to a large degree, by the wetting characteristics of the liquid on the solid surface. For a given wall superheat level in the transition boiling regime, a higher heat flux is generally obtained if the liquid wets the surface than if it poorly wets the surface. For systems that exhibit contact angle hysteresis, the transition boiling curves obtained for decreasing and increasing wall superheat may therefore be somewhat different. The transition boiling curve for decreasing wall superheat may be significantly below that for increasing superheat for such circumstances, as indicated in Figure 4.4.4.

For an electrically heated surface, the rise in temperature associated with the jump from nucleate to film boiling at the critical heat flux is very often large enough to melt component materials and burn out the component. As a result, the critical heat flux is often referred to as the *burnout heat flux* to acknowledge the potentially damaging effects of applying this heat flux level to components cooled by nucleate boiling. Once the jump to film boiling has been made, any further increase in applied heat flux increases the wall superheat, and the system follows basically the same film boiling curve as in the temperature-controlled case.

Correlations of nucleate pool boiling heat transfer data have typically been used as tools to predict nucleate boiling heat transfer in engineering systems and heat exchangers. Many investigators have proposed methods of correlating data of this type; so many, in fact, that a complete discussion of them all could easily fill a major portion of this section. In this section, three of the more commonly used correlation methods will be mentioned. However, before proceeding, two aspects of the interpretation of such correlations are worth noting. First, experimental data indicate that the subcooling of the liquid pool has a negligible effect on the nucleate boiling heat transfer rate. Consequently, the pool boiling correlations are generally regarded as being valid for both subcooled and saturated nucleate boiling. Second, it has also been observed that at moderate to high heat flux levels, a pool boiling heat transfer correlation developed for one heated surface geometry in one specific orientation often works reasonably well for other geometries and/or other orientations. Hence, although a correlation was developed for a

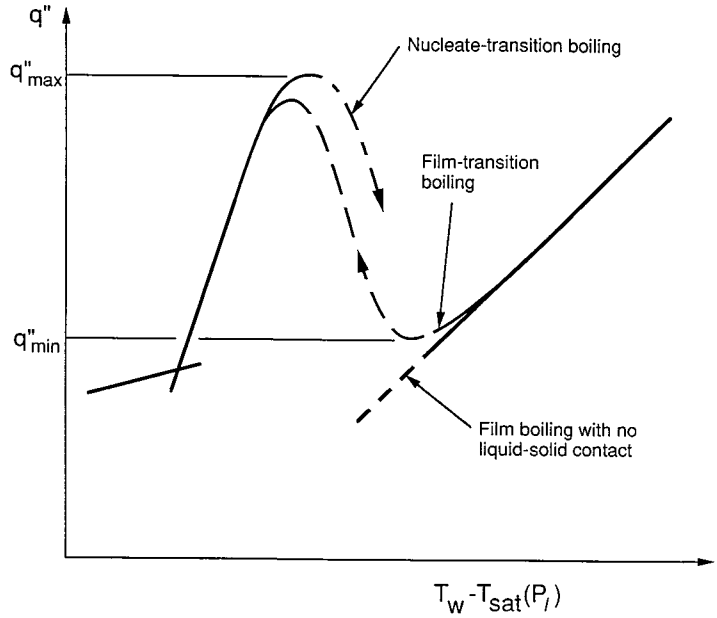


FIGURE 4.4.4 Relative locations of the nucleate transition and film transition portions of the pool boiling curve.

specific geometry and orientation, it may often be used for other geometries, at least at moderate to high heat flux levels.

Having taken note of the above points, a commonly used correlation for nucleate boiling heat transfer developed by Rohsenow (1962) is

$$\frac{q''}{\mu_l h_{fg}}\left[\frac{\sigma}{g(\rho_l-\rho_v)}\right]^{1/2}=\left(\frac{1}{C_{sf}}\right)^{1/r}\mathrm{Pr}_l^{-s/r}\left[\frac{c_{pl}\left[T_w-T_{sat}(P_l)\right]}{h_{fg}}\right]^{1/r} \tag{4.4.4}$$

Values of $r = 0.33$ and $s = 1.7$ are recommended for this correlation, but for water s should be changed to 1.0. The values of C_{sf} in this correlation vary with the type of solid surface and the type of fluid in the system. This empirically accounts for material property and/or wetting angle effects. Recommended values of C_{sf} for specific liquid–solid combinations are given by Rohsenow (1962), but whenever possible, an experiment should be conducted to determine the appropriate value of C_{sf} for the particular solid–liquid combination of interest. If this is not possible, a value of $C_{sf} = 0.013$ is recommended as a first approximation.

As noted previously, the pool boiling curve generally exhibits a maximum heat flux or CHF at the transition between nucleate and transition boiling. This peak value is the maximum level of heat flux from the surface which the system can provide in a nonfilm-boiling mode at a given pressure. The mechanism responsible for the CHF has been the subject of considerable investigation and debate over the past five decades. As the heat flux increases, bubbles generated at the surface coalesce to form vapor columns or jets. Perhaps the most widely cited CHF model postulates that the CHF condition occurs when Helmholtz instability of the large vapor jets leaving the surface distorts the jets, blocking liquid flow to portions of the heated surface. Continued vaporization of liquid at locations on the surface which are starved of replacement liquid than leads to formation of a vapor blanket over part or all of the surface. According to Zuber (1959) for a flat horizontal surface, the predicted maximum heat flux $q''_{\max}$ is

$$q''_{\max} = 0.131\rho_v h_{fg}\left[\frac{\sigma_g(\rho_l - \rho_v)}{\rho_v^2}\right]^{1/4} \tag{4.4.5}$$

but Lienhard and Dhir (1973) recommend that the constant 0.131 in the above relation be replaced with 0.141. Other geometries are treated by Lienhard et al. (1973) and Lienhard and Dhir (1973b). An alternative model has been proposed by Haramura and Katto (1983).

Lienhard and Witte (1985) discuss the development and limitations of hydrodynamic CHF theories.

As shown in the Figure 4.4.3, the boundary between the transition boiling regime and the film boiling regime corresponds to a minimum in the heat flux vs. superheat curve. This condition is referred to as the **minimum heat flux** condition, referred to as the *Leidenfront point*. The minimum heat flux corresponds approximately to the lowest heat flux which will sustain stable film boiling.

For an infinite flat (upward-facing) heated surface, vapor generated at the interface during stable film boiling is released as bubbles at the nodes of a standing two-dimensional Taylor wave pattern. The following relation for the minimum heat flux $q''_{\min}$ derived by Zuber (1959) and Berenson (1961).

$$q''_{\min} = 0.09\rho_v h_{fg}\left[\frac{\sigma_g(\rho_l - \rho_v)}{(\rho_l + \rho_v)^2}\right]^{1/4} \tag{4.4.6}$$

$q''_{\min}$ correlations have been developed by Lienhard and Wong (1964) for horizontal cylinders and Gunnerson and Cronenberg (1980) for spheres.

In film boiling, transport of heat across the vapor film from the wall to the interface may occur by convection, conduction, and radiation. The radiation contribution depends on the nature of the solid surface, but when the radiation effect is small, the heat transfer for film boiling is independent of the material properties and finish of the surface. For buoyancy-driven laminar film boiling over a vertical flat isothermal surface in a pool of saturated liquid, the local heat transfer coefficient from the surface can be obtained from the following relation:

$$h = \left[\frac{k_v^3 g\rho_v(\rho_l - \rho_v)h_{fg}}{4\mu_v(T_w - T_{\text{sat}})x}\right]^{1/4} \tag{4.4.7}$$

At low surface temperatures, radiation effects are negligible and consideration of convective transport alone is sufficient to predict the heat transfer. At higher temperatures radiation effects must also be included. If the vapor in the film absorbs and emits radiation at infrared wavelengths, a detailed treatment of the radiation interaction with the vapor may be necessary to accurately predict the film boiling heat transfer.

Additional information mechanisms such as interfacial waves, turbulence, and variable properties is summarized in Carey (1992).

Transition pool boiling has traditionally been interpreted as a combination of nucleate and film boiling alternately occurring over the heated surface, and a model of transition boiling that accounts for contact angle effects has been proposed by Ramilison and Lienhard (1987).

Internal Convective Boiling

Flow boiling in tubes is perhaps the most complex convective process encountered in applications. In most evaporator and boiler applications, the flow is either horizontal or vertically upward. Figure 4.4.5 schematically depicts a typical low-flux vaporization process in a horizontal round tube. In this example liquid enters as subcooled liquid and leaves as superheated vapor. As indicated in Figure 4.4.5, the flow undergoes transitions in the boiling regime and the two-phase flow regime as it proceeds down the tubes. The regimes encountered depend on the entrance conditions and the thermal boundary conditions at the

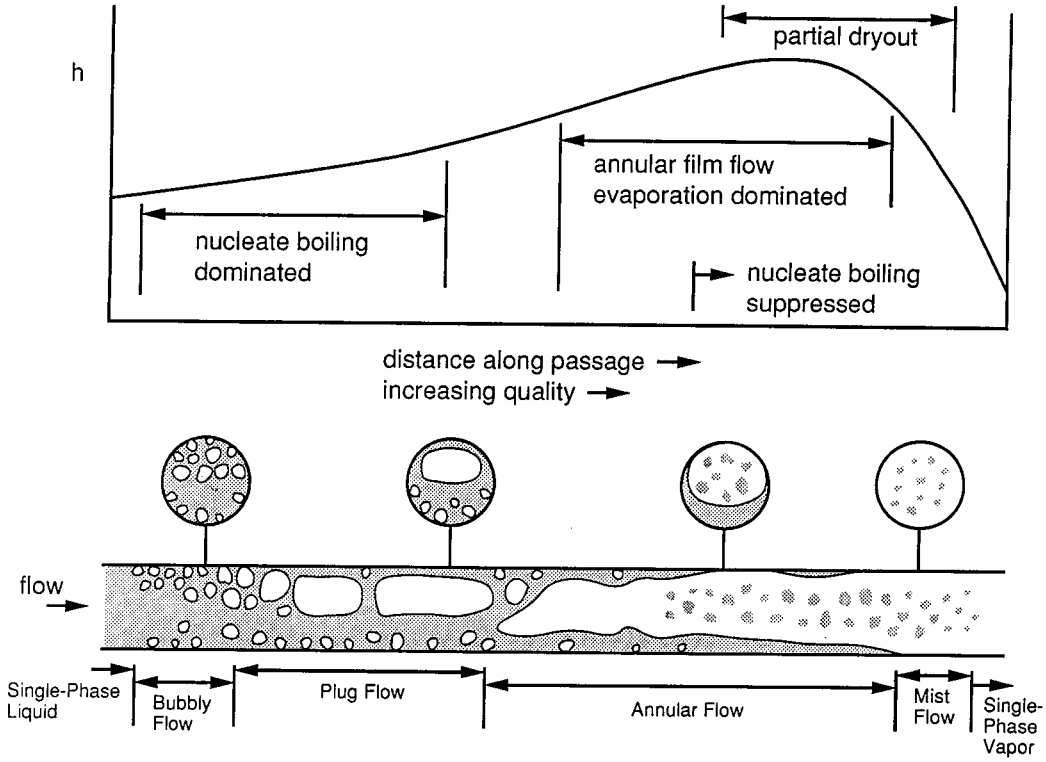


FIGURE 4.4.5 Qualitative variation of the heat transfer coefficient *h* and flow regime with quality for internal convective boiling in a horizontal tube at moderate wall superheat.

tube wall. At low quality the vaporization process is dominated by nucleate boiling, with convective effects being relatively weak. As the quality increases, the flow quickly enters the annular film flow regime in which convective evaporation of the annular liquid film is the dominant heat transfer mechanism. Often the conditions are such that liquid droplets are often entrained in the core vapor flow during annular flow evaporation. Eventually, the annular film evaporates away, leaving the wall dry. Mist-flow evaporation of entrained liquid droplets continues in the post-dryout regime until only vapor remains. Similar, sequences of flow and boiling regimes occurring in vertical upward flow, as indicated in Figure 4.4.6.

The boiling regime trends shown in Figures 4.4.5 and 4.4.6 are typical for low heat flux vaporization processes. At high wall superheat levels, transition boiling or film boiling can also occur. The transition from nucleate boiling to one of these regimes is termed a *departure from nucleate boiling* (DNB) or the CHF condition. However, the heat transfer performance of an evaporator under transition or film boiling conditions is so poor that equipment is not usually designed to operate under such conditions.

Because low-quality or subcooled flow boiling are nucleate boiling dominated, the heat transfer coefficient in these regimes is often predicted using a nucleate boiling correlation developed as a fit to pool boiling data. The usefulness of such an approach is a consequence of the fact that for most conditions of practical interest, nucleate boiling heat transfer is only weakly affected by liquid subcooling or liquid bulk convection.

For saturated convective boiling prior to dryout, relations to predict the heat transfer coefficient have typically been formulated to impose a gradual suppression of nucleate boiling and gradual increase in liquid film evaporation heat transfer as the quality increases. A number of correlations based on such an approach have been developed. An early correlation of this type developed by Chen (1966) has been widely used. One of the better methods of this type is the recently developed correlation of Kandlikar

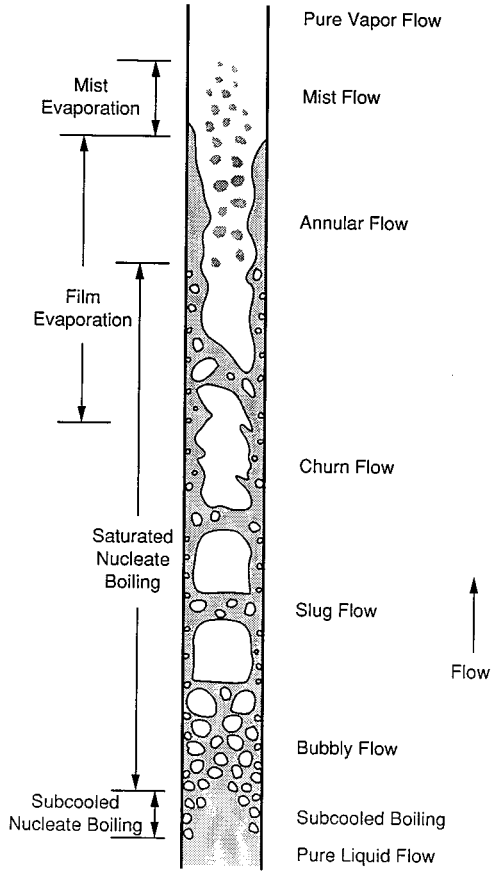


FIGURE 4.4.6 Flow regimes and boiling mechanisms for upflow convective boiling in a vertical tube at moderate wall superheat.

(1989) which has been fit to a broad spectrum of data for both horizontal and vertical tubes. For this method the heat transfer coefficient for a tube of diameter D is given by

$$h = h_l\left[C_1 Co^{C_2}\left(25\mathrm{Fr}_{le}\right)^{C_5} + C_3 Bo^{C_4} F_K\right] \tag{4.4.8}$$

where

$$Co = \left(\frac{1-x}{x}\right)^{0.8}\left(\frac{\rho_v}{\rho_l}\right)^{0.5} \tag{4.4.9}$$

$$Bo = q''/Gh_{fg} \tag{4.4.10}$$

$$Fr_{le} = G^2/\rho_l^2 gD \tag{4.4.11}$$

and h_l is the single-phase heat transfer coefficient for the liquid phase flowing alone in the tube computed as

$$h_l = 0.023\left(\frac{k_l}{D}\right)\left(\frac{G(1-x)D}{\mu_l}\right)^{0.8} \mathrm{Pr}_l^{0.4} \tag{4.4.12}$$

The constants $C_1 - C_5$ are given in Table 4.4.1. The factor F_K is a fluid-dependent parameter. For water, $F_K = 1$. For R-12 and nitrogen, recommended values of F_K are 1.50 and 4.70, respectively. Values of F_K for a variety of other fluids can be obtained from Kandlikar (1989).

TABLE 4.4.1 Constants for the Correlation of Kandlikar (1987)

	Co **< 0.65 (Convective Region)**	***Co*** **≥ 0.65 (Nucleate Boiling Region)**
C_1	1.1360	0.6683
C_2	–0.9	–0.2
C_3	667.2	1058.0
C_4	0.7	0.7
C_5[a]	0.3	0.3

[a] $C_5 = 0$ for vertical tubes and horizontal tubes with $\mathrm{Fr}_{le} > 0.04$.

Methods for predicting the conditions at which dryout or a DNB transition occurs have typically been empirical in nature. Based on fits to extensive data, Levitan and Lantsman (1975) recommended the following relations for the DNB heat flux and the quality at which dryout occurs during flow boiling of water in a tube with an 8-mm diameter.

$$q''_{\mathrm{crit}} = \left[10.3 - 7.8\left(\frac{P}{98}\right) + 1.6\left(\frac{P}{98}\right)^2\right]\left(\frac{G}{1000}\right)^{1.2\{[0.25(P-98)/98]-x\}} e^{-1.5x} \tag{4.4.13}$$

$$x_{\mathrm{crit}} = \left[0.39 + 1.57\left(\frac{P}{98}\right) - 2.04\left(\frac{P}{98}\right)^2 + 0.68\left(\frac{P}{98}\right)^3\right]\left(\frac{G}{1000}\right)^{-0.5} \tag{4.4.14}$$

In the above relations, q''_{crit} is in MW/m^2, P is the pressure in bar, and G is in kg/m^{2}sec. To obtain values of q''_{crit} and x_{crit} for diameters other than 8 mm, Levitan and Lantsman (1975) recommended that the 8-mm values from the above relations be corrected as follows:

$$q''_{\mathrm{crit}} = \left(q''_{\mathrm{crit}}\right)_{8\mathrm{mm}}\left(\frac{8}{D}\right)^{1/2} \tag{4.4.15}$$

$$x_{\mathrm{crit}} = \left(x_{\mathrm{crit}}\right)_{8\mathrm{mm}}\left(\frac{8}{D}\right)^{0.15} \tag{4.4.16}$$

where D is the diameter in millimeters. A good generalized empirical correlation for predicting dryout or CHF conditions in vertical uniformly heated tubes is that recently proposed by Katto and Ohno (1984).

In many cases, post-dryout mist flow evaporation is driven primarily by convective transport from the tube wall to the gas and then to the entrained droplets. In some circumstances, impingement of droplets onto the heat surface and radiation interactions may also be important. In cases where convection is dominant, predictions of the heat transfer coefficient have been developed by modifying a single-phase correlation for the entire flow as vapor with a correction factor which accounts for the presence of the entrained droplets. Often this correction factor has been presumed to be a function of property ratios. An example of such an approach is the correlation of Dougall and Rohsenow (1963) for which the heat transfer coefficient h is given by

$$\frac{hD}{k_v} = 0.023\left[\left(\frac{GD}{\mu_v}\right)\left(x + \frac{\rho_v}{\rho_l}(1-x)\right)\right]^{0.8} \mathrm{Pr}_{v,\mathrm{sat}}^{0.4} \tag{4.4.17}$$

For further information on mechanisms of convective boiling, see the texts of Collier (1981), Stephan (1992), and Carey (1992).

Condensation

As in the case of boiling, surface tension effects, surface wetting characteristics, and metastable phase stability also can play important roles in condensation processes. As a result of interfacial tension, the pressure inside a spherical liquid droplet of radius r must exceed that in the surrounding liquid by $2\sigma/r$. A consequence of this and basic thermodynamics is that at equilibrium the surrounding vapor must actually be slightly supersaturated. The amount of supersaturation required at equilibrium increases as the radius of curvature of the bubble interface decreases.

For a liquid droplet on a solid surface with a specified volume, the wetting contact angle dictates the radius of curvature of the droplet interface. Because of the linkage between the interface curvature and the required equilibrium supersaturation, the wetting behavior thus determines the level above which the vapor supersaturation must be raised for the droplet to grow. Steady condensation on the droplet interface can be sustained only if the vapor is driven beyond this supersaturation level by cooling or depressurization. For such conditions, the vapor is in the metastable supersaturated range indicated in Figure 4.4.2.

Condensation on external surfaces of a body immersed in a gas phase generally falls into one or two categories: **dropwise condensation** or **film condensation**. In dropwise condensation, the liquid-phase condensate collects as individual droplets which grow in size with time on the cold surface. This mode of condensation is most likely when the liquid poorly wets the solid surface. When the condensation rate is high or the liquid readily wets the surface, a film of liquid condensate covers the solid surface, and the process is referred to as film condensation.

Dropwise Condensation. Dropwise condensation may occur on a solid surface cooled below the saturation temperature of a surrounding vapor when the surface is poorly wetted except at locations where well-wetted contaminant nuclei exist. The poorly wetted surface condition can result from contamination or coating of the surface with a substance which is poorly wetted by the liquid phase of the surrounding vapor. In practice, this can be achieved for steam condensation by (1) injecting a nonwetting chemical into the vapor which subsequently deposits on the surface, (2) introducing a substance such as a fatty (i.e., oleic) acid or wax onto the solid surface, or (3) by permanently coating the surface with a low-surface-energy polymer or a noble metal. The effects of the first two methods are generally temporary, since the resulting surface films eventually are dissolved or eroded away.

During dropwise condensation, the condensate is usually observed to appear in the form of droplets which grow on the surface and coalesce with adjacent droplets. When droplets become large enough, they are generally removed from the surface by the action of gravity or drag forces resulting from the motion of the surrounding gas. As the drops roll or fall from the surface they merge with droplets in their path, effectively sweeping the surface clean of droplets. Droplets then begin to grow anew on the freshly-exposed solid surface. This sweeping and renewal of the droplet growth process is responsible for the high heat transfer coefficients associated with dropwise condensation. Theoretical aspects of dropwise condensation are described in two publications by Tanaka (1975, 1979). A discussion of correlations for the heat transfer coefficient associated with dropwise condensation is provided in the review article by Merte (1973).

External Film Condensation. If the liquid phase fully wets a cold surface in contact with a vapor near saturation conditions, the conversion of vapor to liquid will take the form of film condensation. As the name implies, the condensation takes place at the interface of a liquid film covering the solid surface. Because the latent heat of vaporization must be removed at the interface to sustain the process, the rate

of condensation is directly linked to the rate at which heat is transported across the liquid film from the interface to the surface.

The classic integral analysis of Nusselt (1916) for laminar falling-film condensation on a vertical surface considers the physical circumstances shown in Figure 4.4.7. The surface exposed to a motionless ambient of saturated vapor is taken to be isothermal with a temperature below the saturation temperature. Note that although a vertical surface is considered here, the analysis is identical for an inclined surface, except that the gravitational acceleration g is replaced by $g \sin \Omega$, with Ω being the angle between the surface and the horizontal. Because the liquid film flows down the surface because of gravity, this situation is sometimes referred to as *falling-film condensation.*

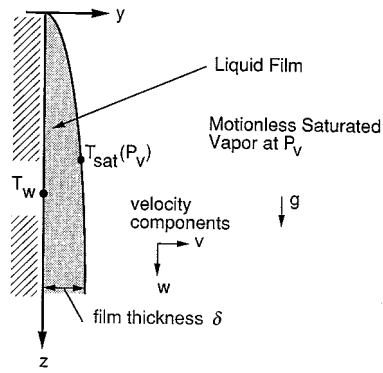


FIGURE 4.4.7 System model for the Nusselt analysis of falling-film condensation.

In its simplest form, the classic Nusselt analysis incorporates the following idealizations: (1) laminar flow, (2) constant properties, (3) that subcooling of liquid is negligible in the energy balance, (4) that inertia effects are negligible in the momentum balance, (5) that the vapor is stationary and exerts no drag, (6) that the liquid-vapor interface is smooth, and (7) that heat transfer across film is only by conduction (convection is neglected). With these idealizations, the following relation for the local heat transfer coefficient h can be obtained

$$\frac{hz}{k_l} = \left[\frac{\rho_l (\rho_l - \rho_v) g h_{fg} z^3}{4 k_l \mu_\ell (T_{sat} - T_w)} \right]^{1/4} \tag{4.4.18}$$

Modified versions of this analysis have been subsequently developed which relax many of these assumptions. Laminar film condensation on a vertical surface can also be analyzed with a full boundary layer formulation. An example of this type of approach is the analysis presented by Sparrow and Gregg (1959).

The analyses described above do not include two physical mechanisms which can significantly affect the transport: (1) the effects of waves on the liquid-vapor interface and (2) interfacial vapor shear drag on the interface. The effects of interfacial shear have been studied analytically by numerous investigators. The effects of surface waves on laminar film condensation are more difficult to incorporate into theoretical analyses. In general, interfacial waves are expected to enhance convective heat transport in the film since it intermittently thins the film, increases the interfacial area, and induces mixing. Because of these effects, laminar film condensation heat transfer data are often significantly higher than the values predicted by simple boundary layer models.

As for any boundary layer flow, when the film Reynolds number becomes large enough, it is expected that a transition to turbulent flow will occur. Eddy diffusivity models of the resulting turbulent transport have been developed by Seban (1954), Dukler (1960), and others. This methodology was later extended to evaporation of a falling liquid film (see, for example, Mills and Chung, 1973).

Subsequent studies (see, for example, Mills and Chung, 1973) have suggested that the presence of the interface tends to damp larger turbulent eddies near the interface in the liquid film. This implies that a viscous sublayer exists at the interface as well as at the wall. Recent efforts to model turbulent falling-film evaporation and condensation processes have therefore included a variation of the eddy viscosity in which it goes to zero at both the wall and the interface. The analysis tools and correlations described above work reasonably well for values of liquid Prandtl number above 1. However, deviation of the predictions using these methods from heat transfer data for liquid metals can be quite significant.

Because of its importance to the design of tube-and-shell condensers, condensation on the outside of horizontal tubes has been the subject of numerous studies. The length of the tube perimeter over which the condensate flows is usually small for commonly used tubes. Consequently, the film Reynolds number is usually low and the flow in the liquid film is laminar.

With slight modification, the Nusselt (1916) analysis of laminar falling-film condensation over a flat plate can be adapted to film condensation on an isothermal horizontal cylinder. Doing so yields the following relation for the mean heat transfer coefficient:

$$\frac{\bar{h}D}{k_l} = 0.728\left[\frac{(\rho_l - \rho_v)gh_{fg}D^3\mathrm{Pr}_l}{\rho_l \nu_l^2 c_{pl}(T_{\mathrm{sat}} - T_w)}\right]^{1/4} \tag{4.4.19}$$

Selin (1961) found that better agreement with film condensation data for horizontal tubes was obtained by replacing the constant factor in Equation (4.4.19) by 0.61. Correlations similar to the single-tube relation above have also been developed for the average condensation heat transfer coefficient for banks of round tubes.

Analytical treatment of laminar film condensation on a sphere is virtually the same as that for a horizontal cylinder. The only differences result from the angular variation of the body perimeter because of the spherical geometry. A general analytical prediction of the local heat transfer coefficient for laminar film condensation on arbitrary axisymmetric bodies has been developed by Dhir and Lienhard (1971).

Condensation in the Presence of a Noncondensable Gas. In nature and in a number of technological applications, condensation of one component vapor in a mixture may occur in the presence of other noncondensable components. The most common example is the condensation of water vapor in the air on a cold solid surface. If the component gases are considered to be a mixture of independent substances, condensation of one component vapor will occur if the temperature of the surface is below the saturation temperature of the pure vapor at its partial pressure in the mixture. This temperature threshold is referred to as the *dew point* of the mixture.

Because only the vapor is condensed, the concentration of the noncondensable gas at the interface is higher than its value in the far ambient. This, in turn, decreases the partial pressure of the vapor at the interface below its ambient value. The corresponding saturation temperature at the interface is therefore lower than the bulk temperature. The resulting depression of the interface temperature generally reduces the condensation heat transfer rate below that which would result for pure vapor alone under the same conditions. Space limitations here preclude a detailed discussion of the effects of noncondensable gases. The interested reader may find more-extensive discussions of this topic in the references by Collier (1981) and Carey (1992).

Internal Convective Condensation. In most power and refrigeration systems, the flow in the condenser is either horizontal or vertically downward. Figure 4.4.8 schematically depicts a typical condensation process in a horizontal round tube. Superheated vapor enters the tube and at the exit end the liquid is subcooled. At a point some distance downstream of the entrance, vapor begins to condense on the walls of the tube. The location at which this occurs is at or slightly before the bulk flow reaches the equilibrium saturation condition. In most condensers, the liquid readily wets the interior of the tube and at high vapor volume fractions the liquid forms a thin liquid film on the interior wall of the tube.

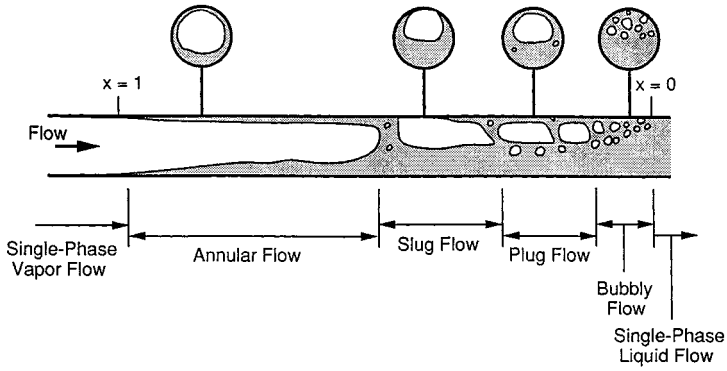


FIGURE 4.4.8 Flow regimes during horizontal cocurrent flow with condensation.

The vapor velocity is generally high at the inlet end of the condenser tube, and the liquid film is driven along the tube by strong vapor shear on the film. At low vapor flow rates, some stratification may occur and the film may be thicker on the bottom of the horizontal tube. At high vapor flow rates, turbulent stresses acting on the liquid may tend to keep the thickness of the liquid film nominally uniform over the perimeter of the tube.

In most condenser applications, shear-dominated annular flow persists to very low qualities and the overwhelming majority of the heat transfer occurs in this regime. The very last stage of the condensation process, corresponding to qualities less than a few percent, may occur in slug, plug, or bubbly two-phase flow. Generally these regimes represent such a small portion of the overall heat transfer in the condenser that some inaccuracy in estimating the heat transfer coefficient for them is tolerated. As a first estimate, the heat transfer coefficient may be predicted using a correlation for pure single-phase liquid flow in the tube at the same total flow rate, or a correlation for annular flow condensation may simply be extrapolated to zero quality.

Because most of the heat duty occurs in the annular flow regime, accurate prediction of the overall heat transfer performance of the condenser requires a predictive methodology that accurately treats the transport in this regime. For this reason, the form of most correlation methods for predicting local convective condensation heat transfer coefficients are optimized to match data in the annular flow regime. One example of such a correlation is the following relation for the local heat transfer coefficient for annular flow condensation proposed by Traviss et al. (1973):

$$\frac{hD}{k_l} = \frac{0.15\mathrm{Pr}_l\mathrm{Re}_l^{0.9}}{F_T}\left[\frac{1}{X_{tt}} + \frac{2.85}{X_{tt}^{0.476}}\right] \tag{4.4.20}$$

where

$$\mathrm{Re}_l = \frac{G(1-x)D}{\mu_l}, \qquad X_{tt} = \left(\frac{1-x}{x}\right)^{0.9}\left(\frac{\rho_v}{\rho_l}\right)^{0.5}\left(\frac{\mu_l}{\mu_v}\right)^{0.1}$$

and F_T is given by

$$F_T = 5\mathrm{Pr}_l + 5\ln\{1 + 5\mathrm{Pr}_l\} + 2.5\ln\{0.0031\mathrm{Re}_l^{0.812}\} \quad \text{for} \quad \mathrm{Re}_l > 1125$$

$$= 5\mathrm{Pr}_l + 5\ln\{1 + \mathrm{Pr}_l(0.0964\mathrm{Re}_l^{0.585} - 1)\} \quad \text{for} \quad 50 < \mathrm{Re}_l < 1125$$

$$= 0.707\mathrm{Pr}_l\mathrm{Re}_l^{0.5} \quad \text{for} \quad \mathrm{Re}_l < 50$$

Carey (1992) has shown that the generic form of this correlation can be derived from a theoretical model of annular flow condensation in a round tube. Several correlations of this general type have been developed as fits to experimental data; see Carey (1992) for a summary. The predictions of these correlations may vary significantly for a given set of conditions. When possible, a correlation should be selected which has been tested against data for conditions close to those for the application of interest.

A correlation methodology that can be used to predict internal convective condensation heat transfer for slug, plug, or wavy stratified flow has also been proposed by Rossen and Meyers (1965). To predict the overall heat transfer performance of a condenser, methods to predict the local heat transfer coefficient must be combined with a scheme to numerically integrate finite-difference forms of the energy, mass, and momentum balances in the tube. For further information on such schemes see the references by Collier (1981) and Carey (1992) (Figure 4.4.9).

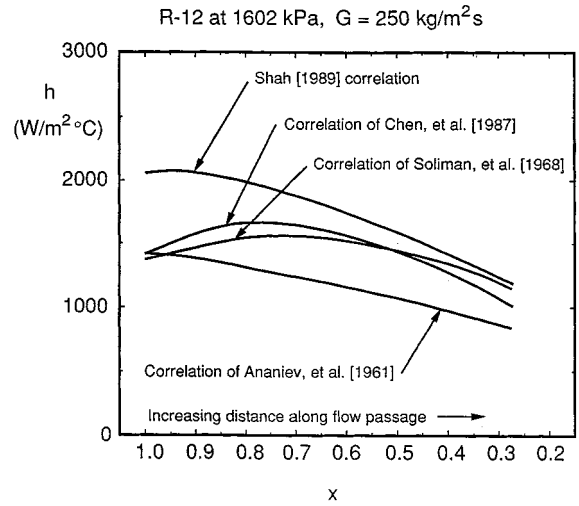


FIGURE 4.4.9 Comparison of the variation of h with x predicted by four correlation methods for internal convective condensation. References cited in this figure are listed in chapter 11 of Carey (1992).

Defining Terms

Critical heat flux (CHF): A maximum heat flux condition that characterizes the transition between nucleate boiling and transition boiling or film boiling.

Dropwise condensation: Condensation of vapor into liquid in discrete droplets, usually attained when a cold surface is poorly wetted by the liquid phase.

Film boiling: Generation of vapor at the interface of a vapor film which entirely covers the hot surface.

Film condensation: Condensation of vapor onto the interface of a liquid film that completely covers a cold surface.

Minimum heat flux: A minimum heat flux condition on the classic boiling curve that characterizes the transition between film boiling and transition boiling. Also, sometimes referred to as the Leidenfrost point, it is a lower bound for heat flux values at which stable film boiling may occur.

Nucleate boiling: Generation of vapor at a hot surface by formation of bubbles at discrete nucleation sites with full liquid wetting of the surface.

Polar molecules: Molecules which have a permanent electric dipole moment. Examples include water and ammonia.

Pool boiling: Generation of vapor at the surface of a hot body immersed in an extensive liquid pool.

Transition boiling: Generation of vapor at a hot surface with intermittent or partial liquid wetting of the surface.

References

Berenson, P.J. 1961. Film boiling heat transfer from a horizontal surface. *J. Heat Transfer*, 83, 351–356.

Carey, V.P. 1992. *Liquid-Vapor Phase Change Phenomena.* Taylor and Francis, Washington, D.C.

Chen, J.C. 1966. Correlation for boiling heat transfer to saturated fluids in convective flow. *Ind. Eng. Chem. Proc. Design and Dev.* 5(3), 322–339.

Collier, J.G. 1981. *Convective Boiling and Condensation,* 2nd ed. McGraw-Hill, New York.

Dhir, V.K. and Lienhard, J. 1971. Laminar film condensation on plane and axisymmetric bodies in nonuniform gravity. *J. Heat Transfer* 93, 97–100.

Dougall, R.S. and Rohsenow, W.M. 1963. Film boiling on the inside of vertical tubes with upward flow of the fluid at low qualities. MIT Report No. 9079-26. MIT, Cambridge, MA.

Dukler, A.E. 1960. Fluid mechanics and heat transfer in vertical falling film systems. *Chem. Eng. Prog. Symp. Ser.* 56(30), 1–10.

Gunnerson, F.S. and Cronenberg, A.W. 1980. On the minimum film boiling conditions for spherical geometries. *J. Heat Transfer* 102,335–341.

Haramura, Y. and Katto, Y. 1983. A new hydrodynamic model of the critical heat flux, applicable widely to both pool and forced convective boiling on submerged bodies in saturated liquids. *Int. J. Heat Mass Transfer* 26, 389–399.

Kandlikar, S.G. 1989. A general correlation for saturated two-phase flow boiling heat transfer inside horizontal and vertical tubes. *J. Heat Transfer* 112, 219–228.

Katto, Y. and Ohno, H. 1984. An improved version of the generalized correlation of critical heat flux for the forced convective boiling in uniformly heated vertical tubes. *Int. Heat Mass Transfer* 21, 1527–1542.

Levitan, L.L. and Lantsman, F.P. 1975. Investigating burnout with flow of a steam-water mixture in a round tube, *Therm. Eng.* (USSR). English trans., 22, 102–105.

Lienhard, J.H. and Dhir, V.K. 1973. Extended hydrodynamic theory of the peak and minimum pool boiling heat fluxes. NASA CR-2270.

Lienhard, J.H. and Witte, L.C. 1985. A historical review of the hydrodynamic theory of boiling. *Rev. Chem. Eng.* 3, 187–277.

Lienhard, J.H. and Wong, P.T.Y. 1964. The dominant unstable wavelength and minimum heat flux during film boiling on a horizontal cylinder. *J. Heat Transfer* 86, 220–226.

Merte, H. 1973. Condensation heat transfer. *Adv. Heat Transfer* 9, 181–272.

Mills, A.F. and Chung, D.K. 1973. Heat transfer across turbulent falling films. *Int. J. Heat Mass Transfer* 16, 694–696.

Nukiyama, S. 1934. The maximum and minimum values of Q transmitted from metal to boiling water under atmospheric pressure. *J. Jpn. Soc. Mech. Eng.* 37, 367–374.

Nusselt, W. 1916. Die Oberflachenkondensation des Wasser dampfes. *Z. Ver. Dtsch. Ininuere* 60, 541–575.

Ramilison, J.M. and Lienhard, J.H. 1987. Transition boiling heat transfer and the film transition regime. *J. Heat Transfer* 109, 746–752.

Rohsenow, W.M. 1962. A method of correlating heat transfer data for surface boiling of liquids. *Trans. ASME* 84, 969–975.

Rossen, H.F. and Meyers, J.A. 1965. Point values of condensing film coefficients inside a horizontal tube. *Chem. Eng. Prog. Symp. Ser.* 61(59), 190–199.

Seban, R. 1954. Remarks on film condensation with turbulent flow. *Trans. ASME* 76, 299–303.
Selin, G. 1961. Heat transfer by condensing pure vapors outside inclined tubes, in *Proc. First Int. Heat Transfer Conf.*, University of Colorado, Boulder, Part II, 279–289.
Sparrow, E.M. and Gregg, J.L. 1959. A boundary-layer treatment of laminar film condensation. *J. Heat Transfer* 81, 13–23.
Stephen, K. 1992. *Heat Transfer in Condensation and Boiling*. Springer-Verlag, New York.
Tanaka, H. 1975. A theoretical study of dropwise condensation. *J. Heat Transfer* 97, 72–78.
Tanaka, H. 1979. Further developments of dropwise condensation theory. *J. Heat Transfer* 101, 603–611.
Traviss, D.P., Rohsenow, W.M., and Baron, A.B. 1973. Forced convection condensation in tubes: a heat transfer correlation for condenser design. *ASHRAE Trans.* 79(I), 157–165.
Witte, L.C. and Lienhard, J.H. 1982. On the existence of two "transition" boiling curves. *Int. J. Heat Mass Transfer* 25, 771–779.
Zuber, N. 1959. Hydrodynamic aspects of boiling heat transfer. AEC Rep. AECU-4439.

Further Information

The texts *Heat Transfer in Condensation and Boiling* by K. Stephan (Springer-Verlag, New York, 1992) and *Liquid-Vapor Phase Change Phenomena* by V.P. Carey (Taylor and Francis, Washington, D.C., 1992) provide an introduction to the physics of boiling and condensation processes. The text by J.G. Collier, *Convective Boiling and Condensation* (2nd ed., McGraw-Hill, New York, 1981), summarizes more-advanced elements of convective boiling and condensation processes. The *ASHRAE Handbook of Fundamentals* (American Society of Heating, Refrigerating, and Air-Conditioning Engineers, Atlanta, GA, 1993) provides some information on boiling and condensation heat transfer and is a good source of thermophysical property data needed to analyze boiling and condensation processes.

Particle Gas Convection

John C. Chen

Introduction

Heat transfer in two-phase systems involving gas and solid particles are encountered in several types of operations important in chemical, power, and environmental technologies. Chief among these are gas fluidized beds which are widely used to achieve either physical processing or chemical reactions that require interfacial contact between gas and solid particles. Currently, fluidized beds operate in either the *bubbling regime* or the *fast-circulating regime*. In the first case, particles are retained in the fluidized bed while the gas passes upward past the particles, partially as rising bubbles. In the second case, gas velocities exceed terminal velocity for the individual particles and the two phases flow through the fluidized bed in cocurrent upward flow. For those applications which require thermal control, convective heat transfer between the fluidized medium and heat transfer surfaces (either immersed tubes or the vessel walls) is an essential element of the process design.

Bubbling Fluidized Beds

Bubbling fluidization occurs when the superficial gas velocity exceeds a critical value wherein the gravitational body force on the solid particles is balanced by the shear force between particles and flowing gas. The superficial gas velocity at this condition, commonly called the minimum fluidization velocity (U_{mf}), marks the boundary between gas flow through packed beds and gas flow in fluidized beds. Wen and Yu (1966) derived the following general equation to estimate U_{mf} for spherical particles:

$$\mathrm{Re}_{mf} = \left[(33.7)^2 + 0.0408\mathrm{Ar}\right]^{1/2} - 33.7 \tag{4.4.21}$$

where

$$\text{Re}_{mf} = \text{particle Reynolds number at } U_{\text{mf}} = \frac{U_{\text{mf}} d_p \rho_g}{\mu_g}$$

$$\text{Ar} = \text{Archimedes number} = \frac{d_p^3 \rho_g \left(\rho_s - \rho_g\right) g}{\mu_g^2}$$

Increasing gas velocity beyond minimum fluidization causes the excess gas to collect into discrete bubbles that grow and rise through the fluidized matrix of solid particles. In this bubbling fluidization regime, the total pressure drop over the height of the fluidized bed, H, is equal to the hydrostatic pressure of the solid mass,

$$\Delta P = g\rho_s(1-\varepsilon)H \tag{4.4.22}$$

where ε = volume fraction of gas (void fraction).

Tubes carrying cooling or heating fluids are often immersed in bubbling fluidized beds to extract or add thermal energy. The effective heat transfer coefficient at the surface of such tubes has been the objective of numerous experimental and analytical investigations. Data for the circumferentially averaged heat transfer coefficient for horizontal tubes are shown in Figure 4.4.10 for various types of solid particles. Characteristics representative of such systems are

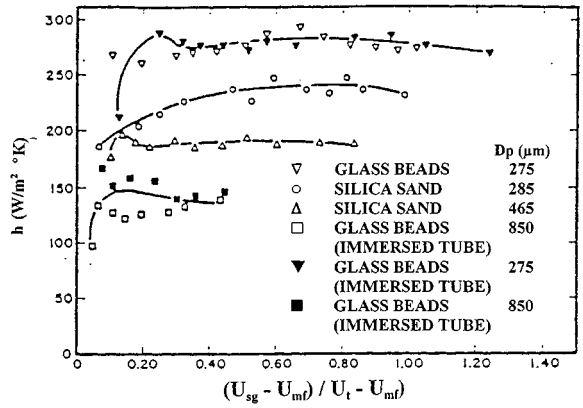


FIGURE 4.4.10 Average heat transfer coefficients for horizontal tubes immersed in bubbling fluidized beds. (From Biyikli, Tuzla and Chen, 1983.)

- The heat transfer coefficient increases sharply as the gas velocity exceeds minimum fluidization velocity,
- After the initial increase, the heat transfer coefficient remains fairly constant over a significant range of the gas velocity beyond minimum fluidization velocity,
- The absolute magnitude of the heat transfer coefficient is severalfold greater than single-phase gas convection at the same superficial velocity,
- The heat transfer coefficient increases as particle size decreases.

Kunii and Levenspiel (1991) have shown that increasing gas pressure and density significantly increases the magnitude of the heat transfer coefficient as well as promoting the occurrence of minimum fluidization at a lower value of superficial gas velocity. The effect of bundle spacing is insignificant at 1-atm pressure but becomes increasingly more important as gas pressure and density increase. The data

of Jacob and Osberg (1957) indicate that the convective heat transfer coefficient in fluidized beds increases with increasing thermal conductivity of the gas phase, for any given particle size.

Several different types of correlations have been suggested for predicting convective heat transfer coefficients at submerged surfaces in bubbling fluidized beds. The first type attributes the enhancement of heat transfer to the scouring action of solid particles on the gas boundary layer, thus decreasing the effective film thickness. These models generally correlate a heat transfer Nusselt number in terms of the fluid Prandtl number and a modified Reynolds number with either the particle diameter or the tube diameter as the characteristic length scale. Examples are

Leva's correlation for vertical surfaces and larger particles (Leva and Gummer, 1952);

$$\mathrm{Nu}_{dp} = \frac{h_c d_p}{k_g} = 0.525\left(\mathrm{Re}_p\right)^{0.75} \tag{4.4.23}$$

where

$$\mathrm{Re}_p = \frac{d_p \rho_g U}{\mu_g}$$

Vreedenberg's (1958) correlation for horizontal tubes refers to the particle of diameter D_t.

$$\mathrm{Nu}_{Dt} = \frac{h_c D_t}{k_g} = 420\left(\frac{\rho_s}{\rho_g}\mathrm{Re}_t\right)^{0.3}\left(\frac{\mu_g^2}{g\rho_s^2 d_p^3}\right)^{0.3}\left(\mathrm{Pr}_g\right)^{0.3} \tag{4.4.24}$$

for

$$\left(\frac{\rho_s}{\rho_g}\mathrm{Re}_t\right) > 2250$$

where

$$\mathrm{Re}_t = \frac{D_t \rho_g U}{\mu_g}$$

Molerus and Schweinzer (1989) developed an alternative type of correlation based on the supposition that the heat transfer is dominated by gas convection through the matrix of particles in the vicinity of the heat transfer surface. Their correlation takes the form:

$$\mathrm{Nu} = \frac{h_c d_p}{k_g} = 0.0247(\mathrm{Ar})^{0.4304}(\mathrm{Pr})^{0.33} \tag{4.4.25}$$

Figure 4.4.11 shows comparison of this model with experimental data obtained at three different pressures. The solid curve represents the relationship for fixed beds, while the dashed lines represent the behavior for fluidized beds (i.e., Equation 4.4.25) upon exceeding minimum fluidization.

A third type of model considers the heat transfer surface to be contacted alternately by gas bubbles and packets of packed particles, leading to a surface renewal process for heat transfer. Mickley and Fairbanks (1955) provided the first analysis of this renewal mechanism. Ozkaynak and Chen (1980)

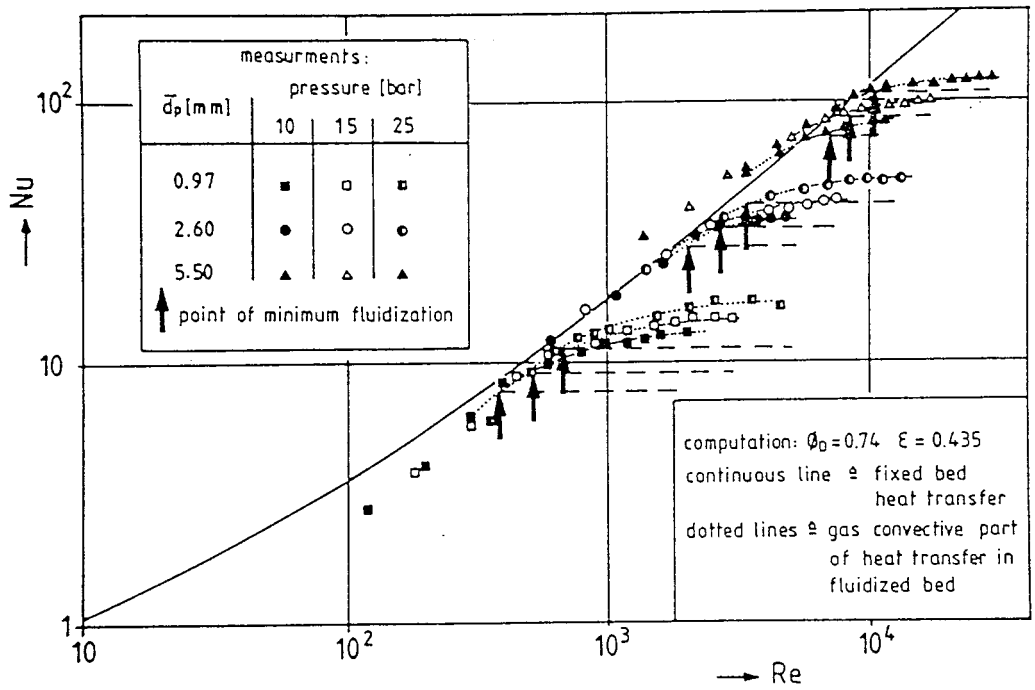


FIGURE 4.4.11 Correlation of Molerus and Schweinzer compared with experimental data (1989).

showed that if experimentally measured values of the packet contact time and residence times are used in the packet model analysis, excellent agreement is obtained.

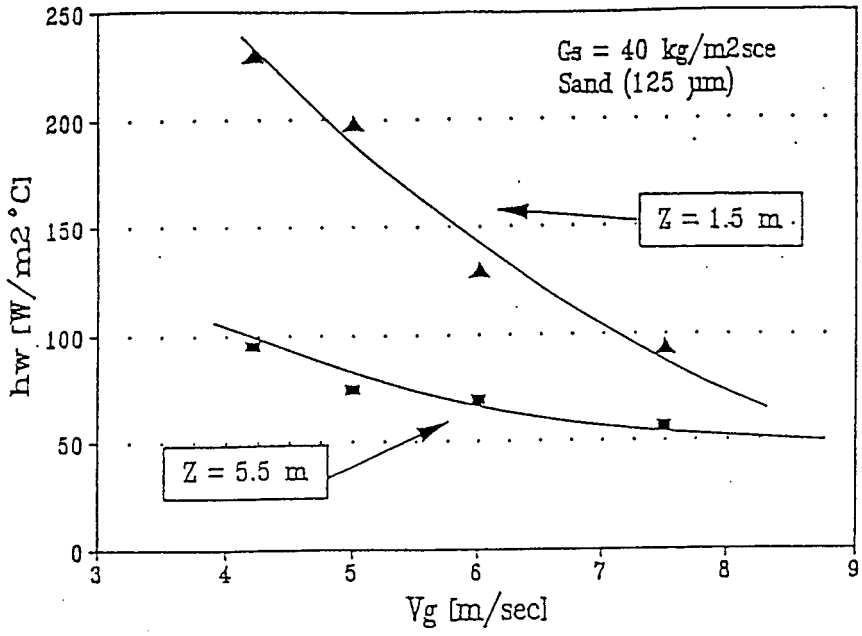
Fast-Circulating Fluidized Beds

Fast fluidization occurs when the superficial gas velocity exceeds the terminal velocity of the solid particles, causing the particles to be suspended in cocurrent upward flow with the gas. This upward flow occurs in "rise reactors" wherein desired physical or chemical reactions occur. In most applications, the two-phase flow exits the top of the riser into a cyclone where the gas phase is separated and exhausted while the solid particles are captured and returned for reinjection at the bottom of the riser. The volumetric concentration of solid particles in these fast fluidized beds (FFBs) tend to be fairly dilute, often with average concentrations of less than 2%. Heat exchange with the particle/gas suspension is usually accomplished through the vertical wall surfaces or through vertical tubes immersed in the duct.

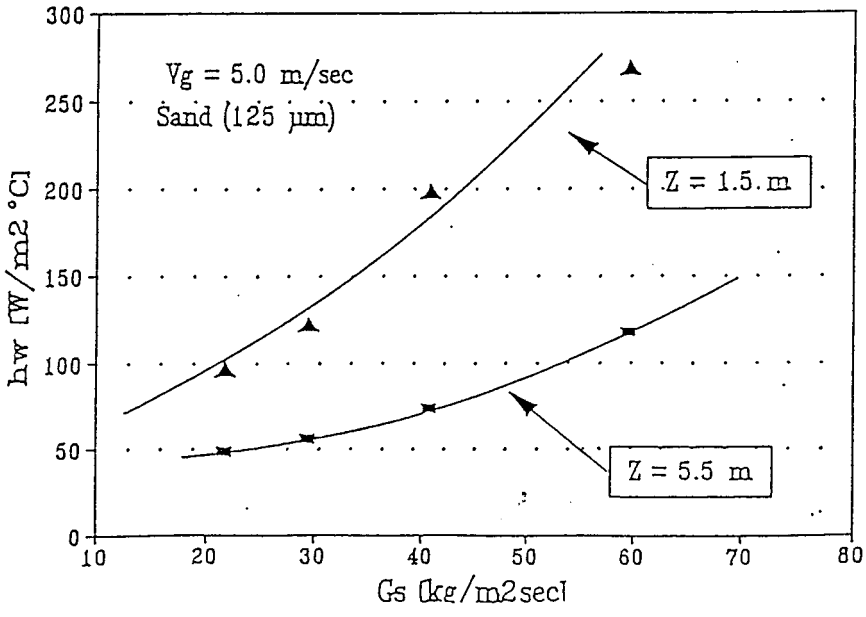
The heat transfer coefficient at vertical surfaces FFBs has been found to increase with increasing solid concentration, aside from other second-order parametric effects. Figure 4.4.12 shows heat transfer coefficients experimentally measured by Dou et al. (1994) for an FFB operating with sand particles of 124 μm mean diameter. Figure 4.4.12b shows that the heat transfer coefficient increased with solid mass flux, for a constant superficial gas velocity. Figure 4.4.12a shows that the heat transfer coefficient decreased parametrically with superficial gas velocity for a constant solid mass flux. Both figures indicate that heat transfer coefficients decrease with increasing elevation in the riser duct. These three parametric trends are all consistent with the hypothesis that heat transfer in FFBs increases with increasing concentration of the solid phase.

It is generally accepted that the effective heat transfer coefficient for surfaces in FFBs have contributions for gas-phase convection, particle-induced convection, and radiation:

$$h = h_g + h_p + h_r \tag{4.4.26}$$



(a)



(b)

FIGURE 4.4.12 Heat transfer coefficients in fast fluidized beds; Vg is superficial gas velocity, Gs is mass flux of particles, and Z is elevation in FFB. (From Dou, Tuzla and Chen, 1992.)

In contrast to the situation in dense-bubbling fluidized beds, the relatively dilute concentration of solid particles in FFBs often results in significant contributions from all three heat transfer mechanisms. The radiation coefficient can be obtained by a gray body model suggested by Grace (1985). The contribution of the gas phase convection (h_g) is commonly estimated based on correlations for gas flow alone at the same superficial gas velocity. Although the presence of particles may alter the turbulence characteristic of this gas flow, any errors caused by this procedure are usually small since h_g is generally smaller than the particle-phase convective coefficient h_p.

For most FFBs, the particle convective contribution to heat transfer is most important and the prediction of h_p is the major concern in thermal design. Unfortunately, mechanistically based models are still lacking and most design methods rely on empirical correlations which often combine the contributions of gas and particle phases into a single convective heat transfer coefficient (h_c). One such correlation proposed by Wen and Miller (1961) is

$$\mathrm{Nu}_{dp} = \frac{h_c d_p}{k_g} = \left(\frac{C_{pp}}{C_{pg}}\right)\left(\frac{\rho_{susp}}{\rho_p}\right)^{0.3}\left(\frac{V_t}{g d_p}\right)^{0.21} \mathrm{Pr}_g \tag{4.4.27}$$

where V_t = terminal velocity of particle.

Other correlations have been proposed by Fraley (1992) and Martin (1984). These correlations are useful as a starting point but have not yet been verified over wide parametric ranges. Large deviations can occur when compared with measurements obtained outside of the experimental parametric ranges.

References

Biyikli, K., Tuzla, K., and Chen, J.C. 1983. Heat transfer around a horizontal tube in freeboard region of fluidized beds, *AIChE J.*, 29(5), 712–716.

Dou, S., Herb, B., Tuzla, K., and Chen, J.C. 1992. Dynamic variation of solid concentration and heat transfer coefficient at wall of circulating fluidized bed, in *Fluidization VII*, Eds. Potter and Nicklin, Engineering Foundation, 793–802.

Fraley, L.D., Lin, Y.Y., Hsiao, K.H., and Solbakken, A. 1983. ASME Paper 83-HT-92, National Heat Transfer Conference, Seattle.

Grace, J.R. 1985. Heat transfer in circulating fluidized beds, *Circulating Fluidized Bed Technology I*, Peramon Press, New York, 63–81.

Jacob, A. and Osberg, G.L. 1957. Effect of gas thermal conductivity on local heat transfer in a fluidized bed, *Can. J. Chem. Eng.*, 35(6), 5–9.

Kunii, D. and Levenspiel, O. 1991. *Fluidization Engineering*, 2nd ed., Butterworth-Heinemann, Boston.

Leva, M. and Grummer, M. 1952. A correlation of solids turnovers in fluidized systems, *Chem. Eng. Prog.*, 48(6), 307–313.

Martin, H. 1984. *Chem. Eng. Process*, 18, 157–223.

Mickley, H.S. and Fairbanks, D.F. 1955. Mechanism of heat transfer to fluidized beds, *AIChE J.*, 1(3), 374–384.

Molerus, O. and Scheinzer, J. 1989. Prediction of gas convective part of the heat transfer to fluidized beds, in *Fluidization IV*, Engineering Foundation, New York, 685–693.

Ozkaynak, T.F. and Chen, J.C. 1980. Emulsion phase residence time and its use in heat transfer models in fluidized bed, *AIChE J.*, 26(4), 544–550.

Vreedenberg, H.A. 1958. Heat transfer between a fluidized bed and a horizontal tube, *Chem. Eng. Sci.*, 9(1), 52–60.

Wen, C.Y. and Yu, Y.H. 1966. A generalized method for predicting the minimum fluidization velocity, *AIChE J.*, 12(2), 610–612

Wen, C.Y. and MIller, E.N. 1961. *Ind. Eng. Chem.*, 53, 51–53.

Melting and Freezing

Noam Lior

Introduction and Overview

Melting and freezing occur naturally (Lunardini, 1981) as with environmental ice in the atmosphere (hail, icing on aircraft), on water bodies and ground regions at the Earth surface, and in the molten Earth core (Figure 4.4.13). They are also a part of many technological processes, such as preservation of foodstuffs (ASHRAE, 1990, 1993), refrigeration and air-conditioning (ASHRAE, 1990, 1993), snow and ice making for skiing and skating (ASHRAE, 1990), organ preservation and cryosurgery (Rubinsky and Eto, 1990), manufacturing (such as casting, molding of plastics, coating, welding, high-energy beam cutting and forming, crystal growth, electrodischarge machining, electrodeposition) (Flemings, 1974; Cheng and Seki, 1991; Tanasawa and Lior, 1992), and thermal energy storage using solid–liquid phase-changing materials (deWinter, 1990).

(a)

(b)

FIGURE 4.4.13 Melting and freezing in nature. (a) A melting icicle. (b) Frozen lava in Hawaii.

In simple thermodynamic systems (i.e., without external fields, surface tension, etc.) of a pure material, melting or freezing occurs at certain combinations of temperature and pressure. Since pressure typically has a relatively smaller influence, only the fusion (freezing or melting) temperature is often used to identify this phase transition. Fusion becomes strongly dependent on the concentration when the material contains more than a single species. Furthermore, melting and freezing are also sensitive to external effects, such as electric and magnetic fields, in more-complex thermodynamic systems.

The equilibrium thermodynamic system parameters during phase transition can be calculated from the knowledge that the partial molar Gibbs free energies or chemical potentials of each component in the two phases must be equal. One important result of using this principle for simple single-component systems is the Clapeyron equation relating the temperature (T) and pressure (P) during the phase transition, such that

$$\frac{dP}{dT} = \frac{h_{s\ell}}{T\Delta v} \tag{4.4.28}$$

where $h_{s\ell}$ is the enthalpy change from phase A to phase B ($=h_B - h_A$, the latent heat of fusion with appropriate sign) and Δv is the specific volume difference between phases A and B ($= v_B - v_A$). Considering for example that phase A is a solid and B a liquid ($h_{s\ell}$ is then positive), examination of Equation (4.4.28) shows that increasing the pressure will result in an increase in the melting temperature if $\Delta v > 0$ (i.e., when the specific volume of the liquid is higher than that of the solid, which is a property of tin, for example), but will result in a decrease of the melting temperature when $\Delta v < 0$ (for water, for example).

In some materials, called glassy, the phase change between the liquid and solid occurs with a gradual transition of the physical properties, from those of one phase to those of the other. When the liquid phase flows during the process, the flow is strongly affected because the viscosity increases greatly as the liquid changes to solid. Other materials, such as pure metals and ice, and eutectic alloys, have a definite line of demarcation between the liquid and the solid, the transition being abrupt. This situation is easier to analyze and is therefore more thoroughly addressed in the literature.

Gradual transition is most distinctly observed in mixtures. Consider the equilibrium phase diagram for a binary mixture (or alloy) composed of species *a* and *b*, shown in Figure 4.4.14. χ is the concentration of species *b* in the mixture, ℓ denotes the liquid, *s* the solid, s_a a solid with a lattice structure of species a in its solid phase but containing some molecules of species *b* in that lattice, and s_b a solid with a lattice structure of species *b* in its solid phase but containing some molecules of species *a* in that lattice. "Liquidus" denotes the boundary above which the mixture is just liquid, and "solidus" is the boundary separating the final solid mixture of species *a* and *b* from the solid–liquid mixture zones and from the other zones of solid s_a and solid s_b.

For illustration, assume that a liquid mixture is at point 1, characterized by concentration χ_1 and temperature T_1 (Figure 4.4.14), and is cooled (descending along the dashed line) while maintaining the concentration constant. When the temperature drops below the liquidus line, solidification starts, creating a mixture of liquid and of solid s_a. Such a two-phase mixture is called the **mushy zone**. At point 2 in that zone, the solid phase (s_a) portion contains a concentration χ_{2,s_a} of component *b*, and the liquid phase portion contains a concentration $\chi_{2,\ell}$ of component *b*. The ratio of the mass of the solid s_a to that of the liquid is determined by the lever rule, and is $(\chi_{2,\ell} - \chi_2)/(\chi_2 - \chi_{2,s_a})$ at point 2. Further cooling to below the solidus line, say to point 3, results in a solid mixture (or alloy) of s_a and s_b, containing concentrations χ_{3,s_A} and χ_{3,s_B} of species *b*, respectively. The ratio of the mass of the solid s_a to that of s_b is again determined by the lever rule, and is $(\chi_{3,s_b} - \chi_3)/(\chi_3 - \chi_{3,s_a})$ at point 3.

A unique situation occurs if the initial concentration of the liquid is χ_e: upon constant-concentration cooling, the liquid forms the solid mixture $s_a + s_b$ having the same concentration and without the formation of a two-phase zone. χ_e is called the **eutectic concentration**, and the resulting solid mixture (or alloy) is called a *eutectic*.

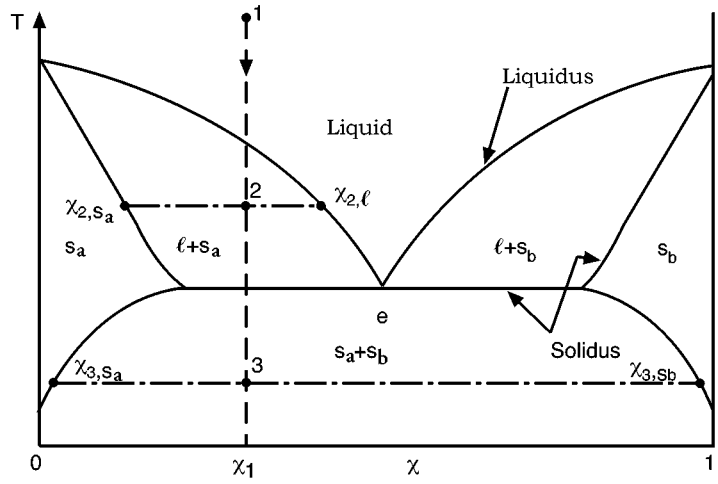


FIGURE 4.4.14 A liquid–solid phase diagram of a binary mixture.

The presence of a two-phase mixture zone with temperature-dependent concentration and phase proportion obviously complicates heat transfer analysis, and requires the simultaneous solution of both the heat and mass transfer equations. Furthermore, the liquid usually does not solidify on a simple planar surface. Crystals of the solid phase are formed at some preferred locations in the liquid, or on colder solid surfaces immersed in the liquid, and as freezing progesses the crystals grow in the form of intricately shaped fingers, called dendrites. This complicates the geometry significantly and makes mathematical modeling of the process very difficult. An introduction to such problems and further references are available in Hayashi and Kunimine (1992) and Poulikakos (1994).

Flow of the liquid phase often has an important role in the inception of, and during, melting and freezing (see Incropera and Viskanta, 1992). The flow may be forced, such as in the freezing of a liquid flowing through or across a cooled pipe, and/or may be due to natural convection that arises whenever there are density gradients in the liquid, here generated by temperature and possibly concentration gradients. It is noteworthy that the change in phase usually affects the original flow, such as when the liquid flowing in a cooled pipe gradually freezes and the frozen solid thus reduces the flow passage, or when the evolving dendritic structure gradually changes the geometry of the solid surfaces that are in contact with the liquid. Under such circumstances, strong coupling may exist between the heat transfer and fluid mechanics, and also with mass transfer when more than a single species is present. The process must then be modeled by an appropriate set of continuity, momentum, energy, mass conservation, and state equations, which need to be solved simultaneously.

More-detailed information about melting and freezing can be found in the monograph by Alexiades and Solomon (1993) and in the comprehensive reviews by Fukusako and Seki (1987) and Yao and Prusa (1989).

Melting and Freezing of Pure Materials

Thorough mathematical treatment of melting and freezing is beyond the scope of this section, but examination of the simplified one-dimensional case for a pure material and without flow effects provides important insights into the phenomena, identifies the key parameters, and allows analytical solutions and thus qualitative predictive capability for at least this class of problems.

In the freezing model, described in Figure 4.4.15, a liquid of infinite extent is to the right ($x > 0$) of the infinite surface at $x = 0$, initially at a temperature T_i higher than the fusion temperature T_f. At time $t = 0$ the liquid surface temperature at $x = 0$ is suddenly lowered to a temperature $T_0 < T_f$, and maintained at that temperature for $t > 0$. Consequently, the liquid starts to freeze at $x = 0$, and the freezing interface (separating in Figure 4.4.15 the solid to its left from the liquid on its right) located

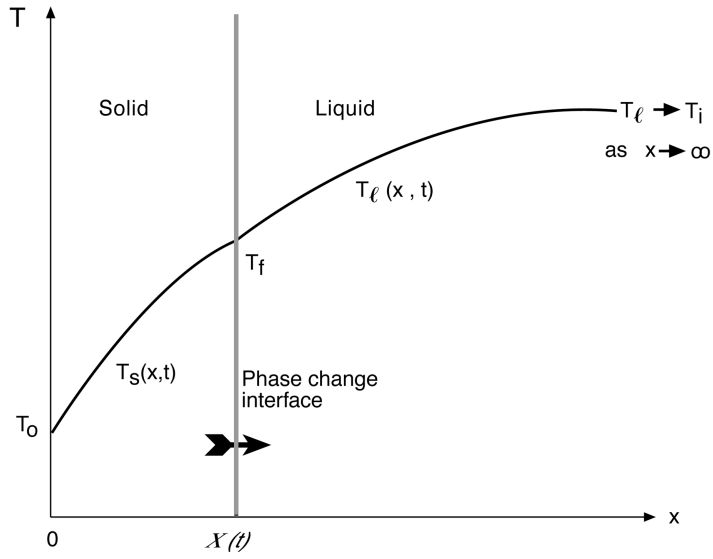


FIGURE 4.4.15 Freezing of semi-inifinite liquid with heat conduction in both phases.

at the position $x = X(t)$ moves gradually to the right (in the positive x direction). We note that in this problem heat is conducted in both phases.

Assuming for simplification that heat transfer is by conduction only — although at least natural convection (Incropera and Viskanta, 1992) and sometimes forced convection and radiation also take place — the governing equations are

In the liquid: The transient heat conduction equation is

$$\frac{\partial T_\ell(x,t)}{\partial t} = \alpha_\ell \frac{\partial^2 T_\ell(x,t)}{\partial x^2} \quad \text{in} \quad X(t) < x < \infty \quad \text{for} \quad t > 0 \tag{4.4.29}$$

$$T_\ell(x,t) = T_i \quad \text{in} \quad x > 0, \quad \text{at} \quad t = 0 \tag{4.4.30}$$

where α_ℓ is the thermal diffusivity of the liquid, with the initial condition and the boundary condition

$$T_\ell(x \rightarrow \infty, t) \rightarrow T_i \quad \text{for} \quad t > 0 \tag{4.4.31}$$

In the solid: The transient heat conduction equation is

$$\frac{\partial T_s(x,t)}{\partial t} = \alpha_s \frac{\partial^2 T_s(x,t)}{\partial x^2} \quad \text{in} \quad 0 < x < X(t) \quad \text{for} \quad t > 0 \tag{4.4.32}$$

where α_s is the thermal diffusivity of the solid, with the boundary condition

$$T_s(0,t) = T_0 \quad \text{for} \quad t > 0 \tag{4.4.33}$$

The remaining boundary conditions are those of temperature continuity and heat balance at the solid–liquid phase-change interface $X(t)$,

$$T_\ell\big[X(t)\big] = T_s\big[X(t)\big] = T_f \quad \text{for} \quad t > 0 \tag{4.4.34}$$

$$k_s\left(\frac{\partial T_s}{\partial x}\right)_{[X(t)]} - k_\ell\left(\frac{\partial T_\ell}{\partial x}\right)_{[X(t)]} = \rho h_{s\ell}\frac{dX(t)}{dt} \quad \text{for} \quad t > 0 \tag{4.4.35}$$

where k_s and k_ℓ are the thermal conductivities of the solid and liquid, respectively, ρ is the density (here it is assumed for simplicity to be the same for the liquid and solid), and $h_{s\ell}$ is the latent heat of fusion. The two terms on the left-hand side of Equation (4.4.35) thus represent the conductive heat flux away from the phase-change interface, into the solid at left and the liquid at right, respectively. Energy conservation at the interface requires that the sum of these fluxes leaving the interface be equal to the amount of heat generated due to the latent heat released there, represented by the term on the right-hand side of the equation.

The analytical solution of Equations (4.4.29) to (4.4.35) yields the temperature distributions in the liquid and solid phases,

$$T_\ell(x,t) = T_i - \left(T_i - T_f\right)\frac{\operatorname{erfc}\left(\dfrac{x}{2\sqrt{\alpha_\ell t}}\right)}{\operatorname{erfc}\left(\lambda\sqrt{\alpha_s/\alpha_\ell}\right)} \tag{4.4.36}$$

$$T_s(x,t) = T_0 + \left(T_f - T_0\right)\frac{\operatorname{erfc}\left(\dfrac{x}{2\sqrt{\alpha_s t}}\right)}{\operatorname{erfc}\lambda} \tag{4.4.37}$$

where erf and erfc are the *error function* and the *complementary error function*, respectively, and λ is a constant, obtained from the solution of the equation

$$\frac{e^{\lambda^2}}{\operatorname{erf}\lambda} - \frac{k_\ell}{k_s}\sqrt{\frac{\alpha_s}{\alpha_\ell}}\,\frac{T_i - T_f}{T_f - T_0}\,\frac{e^{(\alpha_s/\alpha_\ell)\lambda^2}}{\operatorname{erfc}\left(\lambda\sqrt{\alpha_s/\alpha_\ell}\right)} = \frac{\lambda\sqrt{\pi}}{\mathrm{Ste}_s} \tag{4.4.38}$$

where Ste is the Stefan number (dimensionless), here defined for the solid as

$$\mathrm{Ste}_s \equiv \frac{c_s\left(T_f - T_0\right)}{h_{s\ell}} \tag{4.4.39}$$

and c_s is the specific heat of the solid. Solutions of Equation (4.4.38) are available for some specific cases in several of the references, and can be obtained relatively easily by a variety of commonly used software packages.

The solution of Equations (4.4.29) to (4.4.35) also gives an expression for the transient position of the freezing interface,

$$X(t) = 2\lambda\left(\alpha_s t\right)^{1/2} \tag{4.4.40}$$

where λ is the solution of Equation 4.4.38, and thus the expression for the rate of freezing, i.e., the velocity of the motion of the solid liquid interface, is

$$\frac{dX(t)}{dt} = \lambda \alpha_s^{1/2} t^{1/2} \tag{4.4.41}$$

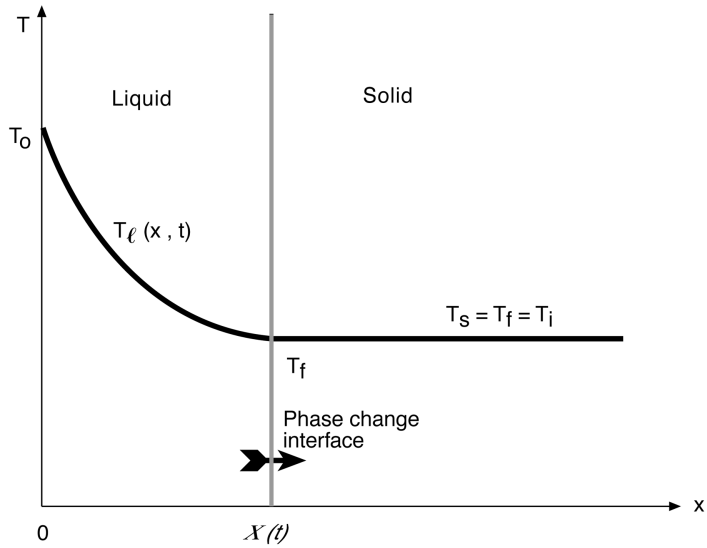


FIGURE 4.4.16 Melting of semi-infinite solid with conduction in the liquid phase only.

For a simple one-dimensional melting example of an analytical solution for melting, consider the semi-infinite solid described in Figure 4.4.16, initially at the fusion temperature T_f. For time $t > 0$ the temperature of the surface (at $x = 0$) is raised to $T_0 > T_f$, and the solid consequently starts to melt there. In this case the temperature in the solid remains constant, $T_s = T_f$, so the temperature distribution needs to be calculated only in the liquid phase. It is assumed that the liquid formed by melting remains motionless and in place. Very similarly to the above-described freezing case, the equations describing this problem are the heat conduction equation

$$\frac{\partial T_\ell(x,t)}{\partial t} = \alpha_\ell \frac{\partial^2 T_\ell(x,t)}{\partial x^2} \quad \text{in} \quad 0 < x < X(t) \quad \text{for} \quad t > 0 \tag{4.4.42}$$

with the initial condition

$$T_\ell(x,t) = T_f \quad \text{in} \quad x > 0, \quad \text{at} \quad t = 0 \tag{4.4.43}$$

the boundary condition

$$T_\ell(0,t) = T_0 \quad \text{for} \quad t > 0 \tag{4.4.44}$$

and the liquid–solid interfacial temperature and heat flux continuity conditions

$$T_\ell\left[X(t)\right] = T_f \quad \text{for} \quad t > 0 \tag{4.4.45}$$

$$-k_\ell \left(\frac{\partial T_\ell}{\partial x}\right)_{[X(t)]} = \rho h_{\ell s} \frac{dX(t)}{dt} \quad \text{for} \quad t > 0 \tag{4.4.46}$$

The analytical solution of this problem yields the temperature distribution in the liquid,

$$T_\ell(x,t) = T_0 - \left(T_0 - T_f\right)\frac{\operatorname{erf}\left(\dfrac{x}{2\sqrt{\alpha_\ell t}}\right)}{\operatorname{erf}\lambda'} \quad \text{for} \quad t > 0 \tag{4.4.47}$$

where λ' is the solution of the equation

$$\lambda' e^{\lambda'^2} \operatorname{erf}(\lambda') = \frac{\mathrm{Ste}_\ell}{\sqrt{\pi}} \tag{4.4.48}$$

with Ste_ℓ here defined for the liquid as

$$\mathrm{Ste}_\ell \equiv \frac{c_\ell\left(T_0 - T_f\right)}{h_{s\ell}} \tag{4.4.49}$$

λ' as a function of Ste, for $0 \le \mathrm{Ste} \le 5$, is given in Figure 4.4.17. The interface position is

$$X(t) = 2\lambda'\left(\alpha_\ell t\right)^{1/2} \tag{4.4.50}$$

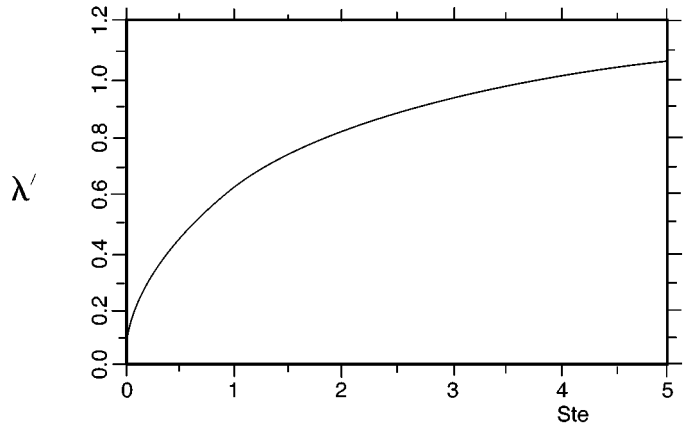


FIGURE 4.4.17 The root λ' of Equation 4.4.48.

The solution of the *freezing* problem under similar conditions, i.e., of a semi-infinite liquid initially at temperature T_f where $T(0,t)$ is abruptly reduced to $T_0 < T_f$ for $t > 0$, is identical to the above if every subscript ℓ is replaced by s and the latent heat $h_{s\ell}$ is replaced by $-h_{s\ell}$.

Example: The temperature of the vertical surface of a large volume of solid paraffin wax used for heat storage, initially at the fusion temperature, $T_i = T_f = 28°C$, is suddenly raised to 58°C. Any motion in the melt may be neglected. How long would it take for the paraffin to solidify to a depth of 0.1 m? Given properties: $\alpha_\ell = (1.09)\ 10^{-7}$ m^2/sec, $\rho_s = \rho_\ell = 814$ kg/m^3, $h_{s\ell} = 241$ kJ/kg, $c_\ell = 2.14$ kJ/kg°C. To find the required time we use Equation (4.4.50), in which the value of λ' needs to be determined. λ' is calculated from Equation (4.4.48), which requires the knowledge of Ste_ℓ. From Equation (4.4.49)

$$\mathrm{Ste}_\ell = \frac{(2.14\ \mathrm{kJ/kg°C})(58°\mathrm{C} - 28°\mathrm{C})}{241.2\ \mathrm{kJ/kg}} = 0.266$$

The solution of Equation (4.4.48) as a function of Ste_ℓ is given in Figure 4.4.17, yielding $\lambda \approx 0.4$. By using Equation (4.4.50), the time of interest is calculated by

$$t = \frac{[X(t)]^2}{4\lambda^2\alpha_\ell} = \frac{(0.1\ \mathrm{m})^2}{4(0.4)^2\left[(1.09)10^7\ \mathrm{m^2/sec}\right]} = (1.43)10^5\ \mathrm{sec} = 39.8\ \mathrm{hr}$$

The axisymmetric energy equation in *cylindrical coordinates*, applicable to both the solid phase and immobile liquid phase (with appropriate assignment of the properties) is

$$\frac{\partial T(r,t)}{\partial t} = \frac{1}{\rho c}\frac{\partial}{\partial r}\left(\frac{k}{r}\frac{\partial T(r,t)}{\partial r}\right) \quad \text{for} \quad t > 0 \tag{4.4.51}$$

and the temperature and heat balance conditions at the solid–liquid phase-change interface $r = R(t)$ are

$$T_\ell[R(t)] = T_s[R(t)] \quad \text{for} \quad t > 0 \tag{4.4.52}$$

$$k_s\left(\frac{\partial T_s}{\partial r}\right)_{R(t)} - k_\ell\left(\frac{\partial T_\ell}{\partial r}\right)_{R(t)} = h_{s\ell}\frac{dR(t)}{dt} \tag{4.4.53}$$

Because of the nature of the differential equations describing nonplanar and multidimensional geometries, analytical solutions are available for only a few cases, such as line heat sources in cylindrical coordinate systems or point heat sources in spherical ones, which have very limited practical application. Other phase-change problems in nonplanar geometries, and in cases when the melt flows during phase change, are solved by approximate and numerical methods (Yao and Prusa, 1989; Alexiades and Solomon, 1993).

Some Approximate Solutions

Two prominent approximate methods used for the solution of melting and freezing problems are the integral method and the *quasi-static* approximation. The integral method is described in Goodman (1964), and only the quasi-static approximation is described here.

To obtain rough estimates of melting and freezing processes quickly, in cases where heat transfer takes place in only one phase, it is assumed in this approximation that effects of sensible heat are negligible relative to those of latent heat (Ste $\rightarrow$ 0), thus eliminating the sensible-heat left-hand side of the energy equations (such as (4.4.29), (4.4.32), and (4.4.51)). This is a significant simplification, since the energy equation then becomes independent of time, and solutions to the steady-state heat conduction problem are much easier to obtain. At the same time, the transient phase-change interface condition (such as Equations (4.4.35) and (4.4.53)) 5 is retained, allowing the estimation of the transient interface position and velocity. This is hence a quasi-static approximation, and its use is shown below.

We emphasize that these are just approximations, without full information on the effect of specific problem conditions on the magnitude of the error incurred when using them. In fact, in some cases, especially with a convective boundary condition, they may produce incorrect results. It is thus necessary to examine the physical viability of the results, such as overall energy balances, when using these approximations.

All of the examples are for melting, but freezing problems have the same solutions when the properties are taken to be those of the solid and $h_{s\ell}$ is replaced everywhere by $-h_{s\ell}$. It is assumed here that the problems are one-dimensional, and that the material is initially at the fusion temperature T_f.

Examples of the Quasi-Static Approximation for Cartesian Coordinate Geometries. Given a semi-infinite solid (Figure 4.4.16), on which a *time-dependent temperature* $T_0(t) > T_f$ *is imposed at* $x = 0$, the above-described quasi-static approximation of Equations (4.4.42) to (4.4.46) easily yields the solution

$$X(t) = \left[2\frac{k_\ell}{\rho h_{s\ell}} \int_0^t \left[T_0(t) - T_f \right] dt \right]^{1/2} \quad \text{for} \quad t \geq 0 \tag{4.4.54}$$

$$T_\ell(x,t) = T_0(t) - \left[T_0(t) - T_f \right] \frac{x}{X(t)} \quad \text{in} \quad 0 \leq x \leq X(t) \quad \text{for} \quad t \geq 0 \tag{4.4.55}$$

The heat flux needed for melting, $q(x,t)$, can easily be determined from the temperature distribution in the liquid (Equation 4.4.55), which is linear because of the steady-state form of the heat conduction equation in this quasi-static approximation, so that

$$q(x,t) = -k_\ell \frac{dT_\ell(x,t)}{dx} = k_\ell \frac{T_0(t) - T_f}{X(t)} \tag{4.4.56}$$

For comparison of this approximate solution to the exact one (Equations (4.4.47) and (4.4.50)), consider the case where $T_0(t) = T_0 =$ constant. Rearranging to use the Stefan number, Equations (4.4.54) and (4.4.55) become

$$X(t) = 2\left(\mathrm{Ste}_\ell/2\right)^{1/2} \left(\alpha_\ell t\right)^{1/2} \quad \text{for} \quad t > 0 \tag{4.4.57}$$

$$T(x,t) = T_0 - \left[T_0 - T_f \right] \frac{x \Big/ \left[2\left(\alpha_\ell t\right)^{1/2} \right]}{\left(\mathrm{Ste}_\ell/2\right)^{1/2}} \quad \text{in} \quad 0 \leq x \leq X(t) \quad \text{for} \quad t \geq 0 \tag{4.4.58}$$

It is easy to show that λ' in the exact solution (Equation 4.4.48) approaches the value $(\mathrm{Ste}_\ell/2)^{1/2}$ when $\mathrm{Ste}_\ell \to 0$, and that otherwise $\lambda' < (\mathrm{Ste}_\ell/2)^{1/2}$. The approximate solution is therefore indeed equal to the exact one when $\mathrm{Ste}_\ell \to 0$, and it otherwise overestimates the values of both $X(t)$ and $T(x,t)$. While the errors depend on the specific problem, they are confined to about 10% in the above-described case (Alexiades and Solomon, 1993).

For the same melting problem but with the *boundary condition of an imposed time-dependent heat flux* $q_0(t)$,

$$-k_\ell \left(\frac{dT_\ell}{dx} \right)_{0,t} = q_0(t) \quad \text{for} \quad t > 0 \tag{4.4.59}$$

the quasi-static approximate solution is

$$X(t) \equiv \frac{1}{\rho h_{s\ell}} \int_0^t q_0(t)\, dt \quad \text{for} \quad t > 0 \tag{4.4.60}$$

$$T_\ell(x,t) = T_f + \frac{q_0}{k_\ell}\left[\frac{q_0}{\rho h_{s\ell}}t - x\right] \quad \text{in} \quad 0 \le x \le X(t) \quad \text{for} \quad t > 0 \tag{4.4.61}$$

For the same case if the *boundary condition is a convective heat flux* from an ambient fluid at the transient temperature $T_a(t)$, characterized by a heat transfer coefficient $\bar{h}$,

$$-k_\ell\left(\frac{dT_\ell}{dx}\right)_{0,t} = \bar{h}\left[T_a(t) - T_\ell(0,t)\right] \quad \text{for} \quad t \ge 0 \tag{4.4.62}$$

the quasi-static approximate solution is

$$X(t) = -\frac{k_\ell}{\bar{h}} + \left\{\left(\frac{k_\ell}{\bar{h}}\right)^2 + 2\frac{k_\ell}{\rho h_{s\ell}}\int_0^t \left[T_a(t) - T_f\right] dt\right\}^{1/2} \quad \text{for} \quad t \ge 0 \tag{4.4.63}$$

$$T_\ell(x,t) = T_f(t)\left[T_a(t) - T_f\right]\frac{\bar{h}\left[X(t) - x\right]}{\bar{h}X(t) + k_\ell} \quad \text{in} \quad 0 \le x \le X(t) \quad \text{for} \quad t > 0 \tag{4.4.64}$$

Examples of the Quasi-Static Approximation for Cylindrical Coordinate Geometries. It is assumed in these examples that the cylinders are very long and that the problems are axisymmetric. Just as in the Cartesian coordinate case, the energy equation (4.4.51) is reduced by the approximation to its steady-state form. Here

$$T_\ell(r_i,t) = T_0(t) > T_f \quad \text{for} \quad t > 0 \tag{4.4.65}$$

Consider the *outward-directed melting* of a hollow cylinder due to a temperature imposed at the internal radius r_i. The solution is

$$T_\ell(r,t) = T_f + \left[T_0(t) - T_f\right]\frac{\ln\left[r/R(t)\right]}{\ln\left[r_i/R(t)\right]} \quad \text{in} \quad r_i \le r \le R(t) \quad \text{for} \quad t > 0 \tag{4.4.66}$$

and the transient position of the phase front, $R(t)$, can be calculated from the transcendental equation

$$2R(t)^2 \ln\frac{R(t)}{r_i} = R(t)^2 - r_i^2 + \frac{4k_\ell}{\rho h_{s\ell}}\int_0^t \left[T_0(t) - T_f\right] dt \tag{4.4.67}$$

If the melting for the same case occurs due to the imposition of a *heat flux* q_0 *at* r_p

$$-k_\ell\left(\frac{dT_\ell}{dx}\right)_{r_i,t} = q_0(t) > 0 \quad \text{for} \quad t > 0 \tag{4.4.68}$$

the solution is

$$T_\ell(r,t) = T_f - \frac{q_0(t)r_i}{k_\ell}\ln\frac{r}{R(t)} \quad \text{in} \quad r_i \le r \le R(t) \quad \text{for} \quad t > 0 \tag{4.4.69}$$

$$R(t) = \left(r_i^2 + 2\frac{r_i}{\rho h_{s\ell}} \int_0^t q_0(t)\, dt \right)^{1/2} \quad \text{for} \quad t > 0 \tag{4.4.70}$$

If the melting for the same case occurs due to the imposition of a *convective heat flux from a fluid at the transient temperature* $T_a(t)$, *with a heat transfer coefficient* $\bar{h}$, *at* r_i

$$-k_\ell \left(\frac{dT_\ell}{dr} \right)_{r_i,t} = \bar{h}\left[T_a(t) - T_f(r_i,t) \right] > 0 \quad \text{for} \quad t > 0 \tag{4.4.71}$$

The solution is

$$T_\ell(r,t) = T_f + \left[T_a(t) - T_f \right] \frac{\ln\left[r/R(t) \right]}{\ln\left[r_i/R(t) \right] - k_\ell/\bar{h} r_i} \quad \text{in} \quad r_i \le r \le R(t) \quad \text{at} \quad t > 0 \tag{4.4.72}$$

with $R(t)$ calculated from the transcendental equation

$$2R(t)^2 \ln\frac{R(t)}{r_i} = \left(1 - \frac{2k_\ell}{\bar{h} r_i} \right) \left[R(t)^2 - r_i^2 \right] + \frac{4k_\ell}{\rho h_{s\ell}} \int_0^t \left[T_a(t) - T_f \right] dt \tag{4.4.73}$$

The solutions for *inward melting* of a cylinder, where heating is applied at the outer radius r_o, are the same as the above-described ones for the outward-melting cylinder, if the replacements $r_i \to r_o$, $q_0 \to -q_0$, and $\bar{h} \to -\bar{h}$ are made. If such a cylinder is not hollow, then $r_i = 0$ is used.

Estimation of Freezing and Melting Time

There are a number of approximate formulas for estimating the freezing and melting times of different materials having a variety of shapes. The American Society of Heating, Refrigerating, and Air-Conditioning Engineers (ASHRAE) provides a number of such approximations for estimating the freezing and thawing times of foods (ASHRAE, 1993). For example, if it can be assumed that the freezing or thawing occurs at a single temperature, the time to freeze or thaw, t_f, for a body that has shape parameters P and R (described below) and thermal conductivity k, initially at the fusion temperature T_f, and which is exchanging heat via heat transfer coefficient $\bar{h}$ with an ambient at the constant T_a, can be approximated by Plank's equation

$$t_f = \frac{h_{s\ell}\rho}{\left| T_f - T_a \right|} \left(\frac{Pd}{\bar{h}} + \frac{Rd^2}{k} \right) \tag{4.4.74}$$

where d is the diameter of the body if it is a cylinder or a sphere, or the thickness when it is an infinite slab, and where the shape coefficients P and R for a number of body forms are given in Table 4.4.2. Shape coefficients for other body forms are also available. To use Equation 4.4.74 for freezing, k and ρ should be the values for the food in its frozen state. In thawing, they should be for the unfrozen food. Other simple approximations for melting and thawing times can be found in Cleland et al. (1987).

Example of Using Plank's Equation (4.4.74) for Estimating Freezing Time. Estimate the time needed to freeze a fish, the shape of which can be approximated by a cylinder 0.5 m long having a diameter of 0.1 m. The fish is initially at its freezing temperature, and during the freezing process it is surrounded by air at $T_a = -25°C$, with the cooling performed with a convective heat transfer coefficient $\bar{h}$ = 68 W/m² K. For the fish, $T_f = -1°C$, $h_{s\ell}$ = 200 kJ/kg, ρ_s = 992 kg/m³, and ${}_sk_s$ = 1.35 W/m K.

TABLE 4.4.2 Shape Factors for Equation (4.4.74)

Forms	*P*	*R*
Slab	1/2	1/8
Cylinder	1/4	1/16
Sphere	1/6	1/24

From ASHRAE, in *Fundamentals*, ASHRAE, Atlanta, 1993, chap. 29. With permission.

By using Table 4.4.2, the geometric coefficients for the cylindrical shape of the fish are $P = {}^1/_2$ and $R = {}^1/_{16}$, while d is the cylinder diameter, = 0.1 m. Substituting these values into Equation (4.4.74) gives

$$t_f = \frac{200{,}000 \cdot 992}{-1-(-25)}\left(\frac{1/4(0.1)}{68} + \frac{1/16(0.1)^2}{1.35}\right) = 6866 \text{ sec} = 1.9 \text{ hr}$$

In fact, freezing or melting of food typically takes place over a range of temperatures, and approximate Plank-type formulas have been developed for various specific foodstuffs and shapes to represent reality more closely than Equation (4.4.74) (ASHRAE, 1993).

Alexiades and Solomon (1993) provide several easily computable approximate equations for estimating the time needed to melt a simple solid body initially at the fusion temperature T_f. It is assumed that conduction occurs in one phase (the liquid) only, that the problems are axi- and spherically symmetric for cylindrical and spherical bodies, respectively, and that the melting process for differently shaped bodies can be characterized by a single geometric parameter, r, in the body domain $0 \le r \le L$, using a shape factor, ω, defined by

$$\omega = \frac{LA}{V} - 1 \tag{4.4.75}$$

where A is the surface area across which the heat is transferred into the body and V is the body volume, to account for the specific body shape:

$$\omega = \begin{cases} 0 & \text{for a slab insulated at one end} \\ 1 & \text{for a cylinder} \\ 2 & \text{for a sphere} \end{cases} \tag{4.4.76}$$

$0 \le \omega \le 2$ always, and ω may be assigned appropriate values for shapes intermediate between the slab, cylinder, and sphere. For example, a football-shaped body, somewhere between a cylinder and sphere, may be assigned $\omega = 1.5$, and a short cylinder with a large diameter-to-height ratio may have $\omega = 0.5$.

For the case where the temperature $T_0 > T_f$ is imposed on the boundary at $t = 0$, the melt time, t_m can be estimated by

$$t_m = \frac{L^2}{2\alpha_\ell(1+\omega)\text{Ste}_\ell}\left[1 + \left(0.25 + 0.17\omega^{0.7}\right)\text{Ste}_\ell\right] \tag{4.4.77}$$

valid for $0 \le \text{Ste}_\ell \le 4$.

If the *heat input is convective*, with a heat transfer coefficient $\bar{h}$ from a fluid at temperature T_a, the approximate melt time is

$$t_m = \frac{L^2}{2\alpha_\ell(1+\omega)\mathrm{Ste}_\ell}\left[1+\frac{2k_\ell}{\bar{h}L}+\left(0.25+0.17\omega^{0.7}\right)\mathrm{Ste}_\ell\right] \tag{4.4.78}$$

valid for $0 \le \mathrm{Ste}_\ell \le 4$ and $\bar{h}L/k_\ell \ge 0.1$, and the temperature, $T(0,t)$, of the surface across which the heat is supplied can be estimated from the implicit time-temperature relationship:

$$t = \frac{\rho c_\ell k_\ell}{2\bar{h}^2\mathrm{Ste}_\ell}\left[1.18\mathrm{Ste}_\ell\left(\frac{T(0,t)-T_f}{T_a-T(0,t)}\right)^{1.83}+\left(\frac{T_a-T_f}{T_a-T(0,t)}\right)^2-1\right] \tag{4.4.79}$$

Both equations (4.4.78) and (4.4.79) are claimed to be accurate within 10%.

The suitability of using several simplified analytical solutions for the estimation of freezing and melting times for more-realistic problems was assessed by Dilley and Lior (1986).

Defining Terms

Eutectic concentration: A concentration of a component of a multicomponent liquid at which the liquid would upon freezing form a solid containing the same concentration, and at which the freezing process is completed at a single temperature.

Mushy zone: The zone composed of both liquid and solid, bounded by the liquidus and solidus curves, in a freezing or melting process.

References

Alexiades, V. and Solomon, A.D. 1993. *Mathematical Modeling of Melting and Freezing Processes*, Hemisphere Publishing, Washington, D.C.

ASHRAE (American Society of Heating, Refrigerating, and Air-Conditioning Engineers). 1993. Cooling and freezing times of foods, in *Fundamentals*, ASHRAE, Atlanta, GA, chap. 29.

ASHRAE (American Society of Heating, Refrigerating, and Air-Conditioning Engineers). 1990. *Refrigeration*, ASHRAE, Atlanta, GA.

Cheng, K.C. and Seki, N., Eds. 1991. *Freezing and Melting Heat Transfer in Engineering*, Hemisphere Publishing, Washington, D.C.

Cleland, D.J., Cleland, A.C., and Earle, R.L. 1987. Prediction of freezing and thawing times for multidimensional shapes by simple formulae: Part 1, regular shapes; Part 2, irregular shapes. *Int. J. Refrig.*, 10, 156–166; 234–240.

DeWinter, F. 1990. Energy storage of solar systems; in *Solar Collectors, Energy Storage, and Materials*, MIT Press, Cambridge, MA, Section II.

Dilley, J.F. and Lior, N. 1986. The evaluation of simple analytical solutions for the prediction of freeze-up time, freezing, and melting. *Proc. 8th International Heat Transfer Conf.*, 4, 1727–1732, San Francisco.

Flemings, M.C. 1974. *Solidification Processes*, McGraw-Hill, New York.

Fukusako, S. and Seki, N. 1987. Fundamental aspects of analytical and numerical methods on freezing and melting heat-transfer problems, in *Annual Review of Numerical Fluid Mechanics and Heat Transfer*, Vol. 1, T.C. Chawla, Ed., Hemisphere, Publishing, Washington, D.C., chap. 7, 351–402.

Goodman, T.R. 1964. Application of integral methods to transient nonlinear heat transfer, in *Advances in Heat Transfer*, Vol. 1, T.F. Irvine and J.P. Hartnett, Eds., Academic Press, San Diego, 51–122.

Hayashi, Y. and Kunimine, K. 1992. Solidification of mixtures with supercooling, in *Heat and Mass Transfer in Materials Processing*, I. Tanasawa and N. Lior, Eds., Hemisphere Publishing, New York, 265–277.

Incropera, F.P. and Viskanta, R. 1992. Effects of convection on the solidification of binary mixtures, in *Heat and Mass Transfer in Materials Processing*, I. Tanasawa and N Lior, Eds., Hemisphere Publishing, New York, 295–312.

Lunardini, V.J. 1981. *Heat Transfer in Cold Climate*, Van Nostrand-Reinhold, Princeton, NJ.

Poulikakos, D. 1994. *Conduction Heat Transfer*, Prentice-Hall, Englewood Cliffs, NJ.

Rubinsky, B. and Eto, T.K. 1990. Heat transfer during freezing of biological materials, in *Annual Review of Heat Transfer*, Vol. 3, C.L. Tien, Eds., Hemisphere Publishing, Washington, D.C., chap. 1, 1–38.

Tanasawa, I. and Lior, N., Ed. 1992. *Heat and Mass Transfer in Materials Processing*, Hemisphere Publishing, New York.

Yao, L.S. and Prusa, J. 1989. Melting and freezing, in *Advances in Heat Transfer*, Vol. 19, J.P. Hartnett and T.F. Irvine, Eds., Academic Press, San Diego, 1–95.

Further Information

Many textbooks on heat transfer (some listed in the References section above) contain material about melting and freezing, and many technical journals contain articles about this subject. Some of the major journals, classified by orientation, are

General: *ASME Journal of Heat Transfer, International Journal of Heat & Mass Transfer, Numerical Heat Transfer, Canadian Journal of Chemical Engineering, AIChE Journal*

Refrigeration: *Transactions of the ASHRAE, International Journal of Refrigeration, Referigeration, Journal of Food Science, Bulletin of the International Institute of Refrigeration*

Manufacturing: *ASME Journal of Engineering for Industry, Journal of Crystal Growth, Materials Science and Engineering A*

Geophysical, climate, cold regions engineering: *Limnology and Oceanography, Journal of Geophysical Research, ASCE Journal of Cold Regions Engineering, Cold Regions Science and Technology*

Medical: *Cryobiology, ASME Journal of Biomechanical Engineering, Journal of General Physiology*

4.5 Heat Exchangers

Ramesh K. Shah and Kenneth J. Bell

The two major categories of heat exchangers are shell-and-tube exchangers and compact exchangers. Basic constructions of gas-to-gas compact heat exchangers are plate-fin, tube-fin and all prime surface recuperators (including polymer film and laminar flow exchangers), and compact regenerators. Basic constructions of liquid-to-liquid and liquid-to-phase-change compact heat exchangers are gasketed and welded plate-and-frame, welded stacked plate (without frames), spiral plate, printed circuit, and dimple plate heat exchangers.

Shell-and-tube exchangers are custom designed for virtually any capacity and operating condition, from high vacuums to ultrahigh pressures, from cryogenics to high temperatures, and for any temperature and pressure differences between the fluids, limited only by the materials of construction. They can be designed for special operating conditions: vibration, heavy fouling, highly viscous fluids, erosion, corrosion, toxicity, radioactivity, multicomponent mixtures, etc. They are made from a variety of metal and nonmetal materials, and in surface areas from less than 0.1 to 100,000 m^2 (1 to over 1,000,000 ft^2). They have generally an order of magnitude less surface area per unit volume than the compact exchangers, and require considerable space, weight, support structure, and footprint.

Compact heat exchangers have a large heat transfer surface area per unit volume of the exchanger, resulting in reduced space, weight, support structure and footprint, energy requirement and cost, as well as improved process design, plant layout and processing conditions, together with low fluid inventory compared with shell-and-tube exchangers. From the operating condition and maintenance point of view, compact heat exchangers of different constructions are used for specific applications, such as for high-temperature applications (up to about 850°C or 1550°F), high pressure applications (over 200 bars), and moderate fouling applications. However, applications do not involve both high temperature and pressure simultaneously. Plate-fin exchangers are generally brazed, and the largest size currently manufactured is $1.2 \times 1.2 \times 6$ m ($4 \times 4 \times 20$ ft). Fouling is one of the major potential problems in many compact exchangers except for the plate heat exchangers. With a large frontal area exchanger, flow maldistribution could be another problem. Because of short transient times, a careful design of controls is required for startup of some compact heat exchangers compared with shell-and-tube exchangers. No industry standards or recognized practice for compact heat exchangers is yet available.

This section is divided into two parts: Compact Heat Exchangers and Shell-and-Tube Exchangers, written by R. K. Shah and K. J. Bell, respectively. In the compact heat exchangers section, the following topics are covered: definition and description of exchangers, heat transfer and pressure drop analyses, heat transfer and flow friction correlations, exchanger design (rating and sizing) methodology, flow maldistribution, and fouling. In the shell-and-tube heat exchangers section, the following topics are covered: construction features, principles of design, and an approximate design method with an example.

Compact Heat Exchangers

Ramesh K. Shah

Introduction

A heat exchanger is a device to provide for transfer of internal thermal energy (enthalpy) between two or more fluids, between a solid surface and a fluid, or between solid particulates and a fluid, in thermal contact without external heat and work interactions. The fluids may be single compounds or mixtures. Typical applications involve heating or cooling of a fluid stream of concern, evaporation or condensation of single or multicomponent fluid stream, and heat recovery or heat rejection from a system. In other applications, the objective may be to sterilize, pasteurize, fractionate, distill, concentrate, crystallize, or control process fluid. In some heat exchangers, the fluids transferring heat are in direct contact. In other heat exchangers, heat transfer between fluids takes place through a separating wall or into and out of a

wall in a transient manner. In most heat exchangers, the fluids are separated by a heat transfer surface and do not mix. Such exchangers are referred to as *direct transfer type*, or simply *recuperators*. Exchangers in which there is an intermittent flow of heat from the hot to cold fluid (via heat storage and heat rejection through the exchanger surface or matrix) are referred to as *indirect transfer type* or simply *regenerators.*

The heat transfer surface is a surface of the exchanger core which is in direct contact with fluids and through which heat is transferred by conduction in a recuperator. The portion of the surface which also separates the fluids is referred to as a *primary* or *direct surface*. To increase heat transfer area, appendages known as fins may be intimately connected to the primary surface to provide an *extended*, *secondary*, or *indirect surface*. Thus, the addition of fins reduces the thermal resistance on that side and thereby increases the net heat transfer from the surface for the same temperature difference.

Heat exchangers may be classified according to transfer process, construction, flow arrangement, surface compactness, number of fluids, and heat transfer mechanisms as shown in Figure 4.5.1.

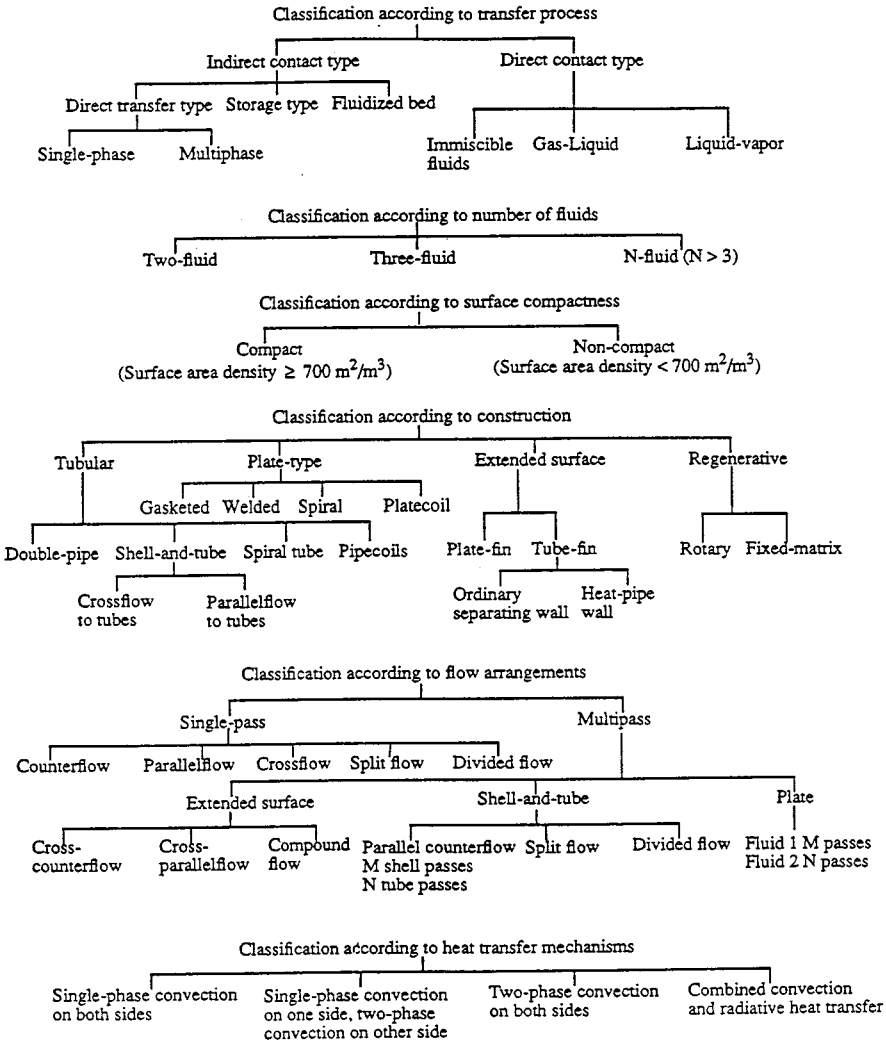


FIGURE 4.5.1 Classification of heat exchangers.

A gas-to-fluid heat exchanger is referred to as a compact heat exchanger if it incorporates a heat transfer surface having a surface area density above about 700 m^2/m^3 (213 ft^2/ft^3) on at least one of the fluid sides which usually has gas flow. It is referred to as a laminar flow heat exchanger if the surface area density is above about 3000 m^2/m^3 (914 ft^2/ft^3), and as a micro heat exchanger if the surface area density is above about 10,000 m^2/m^3 (3050 ft^2/ft^3). A liquid/two-phase heat exchanger is referred to as a compact heat exchanger is the surface area density on any one fluid side is above about 400 m^2/m^3 (122 ft^2/ft^3). A typical process industry shell-and-tube exchanger has a surface area density of less than 100 m^2/m^3 on one fluid side with plain tubes, and two to three times that with the high-fin-density low-finned tubing. Plate-fin, tube-fin, and rotary regenerators are examples of compact heat exchangers for gas flows on one or both fluid sides, and gasketed and welded plate heat exchangers are examples of compact heat exchangers for liquid flows.

Types and Description

Gas-to-Fluid Exchangers.

The important design and operating considerations for compact extended surface exchangers are (1) usually at least one of the fluids is a gas or specific liquid that has low h; (2) fluids must be clean and relatively noncorrosive because of small hydraulic diameter (D_h) flow passages and no easy techniques for mechanically cleaning them; (3) the fluid pumping power (i.e., pressure drop) design constraint is often equally as important as the heat transfer rate; (4) operating pressures and temperatures are somewhat limited compared with shell-and-tube exchangers as a result of the joining of the fins to plates or tubes such as brazing, mechanical expansion, etc.; (5) with the use of highly compact surfaces, the resultant shape of a gas-to-fluid exchanger is one having a large frontal area and a short flow length (the header design of a compact heat exchanger is thus important for a uniform flow distribution among the very large number of small flow passages); (6) the market potential must be large enough to warrant the sizable manufacturing research and tooling costs for new forms to be developed.

Some advantages of plate-fin exchangers over conventional shell-and-tube exchangers are as follows. Compact heat exchangers, generally fabricated from thin metallic plates, yield large heat transfer surface area per unit volume (β), typically up to ten times greater than the 50 to 100 m^2/m^3 provided by a shell-and-tube exchanger for general process application and from 1000 to 6000 m^2/m^3 for highly compact gas side surfaces. Compact liquid or two-phase side surfaces have a β ratio ranging from 500 to 600 m^2/m^3. A compact exchanger provides a tighter temperature control; thus it is useful for heat-sensitive materials, improves the product (e.g., refining fats from edible oil) and its quality (such as a catalyst bed). Also, a compact exchanger could provide rapid heating or cooling of a process stream, thus improving the product quality. The plate-fin exchangers can accommodate multiple (up to 12 or more) fluid streams in one exchanger unit with proper manifolding, thus allowing process integration and cost-effective compact solutions.

Fouling is one of the potential major problems in compact heat exchangers (except for plate-and-frame heat exchangers), particularly those having a variety of fin geometries or very fine circular or noncircular flow passages that cannot be cleaned mechanically. Chemical cleaning may be possible; thermal baking and subsequent rinsing is possible for small-size units. Hence, extended surface compact heat exchangers may not be used in heavy fouling applications.

Liquid-to-Liquid Exchangers.

Liquid-to-liquid and phase-change exchangers are plate-and-frame and welded plate heat exchangers (PHE), spiral plate, and printed circuit exchangers; some of them are described next in some detail along with other compact heat exchangers and their applications.

Plate-Fin Heat Exchangers.

This type of exchanger has “corrugated” fins or spacers sandwiched between parallel plates (referred to as plates or parting sheets) as shown in Figure 4.5.2. Sometimes fins are incorporated in a flat tube with rounded corners (referred to as a formed tube), thus eliminating a need for the side bars. If liquid or

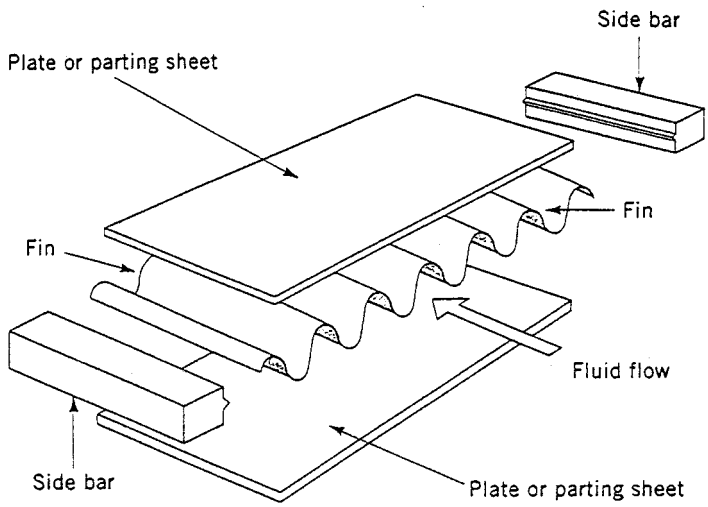


FIGURE 4.5.2 Typical components of a plate-fin exchanger.

phase-change fluid flows on the other side, the parting sheet is usually replaced by a flat tube with or without inserts/webs. Other plate-fin constructions include drawn-cup (see Figure 4.5.3) or tube-and-

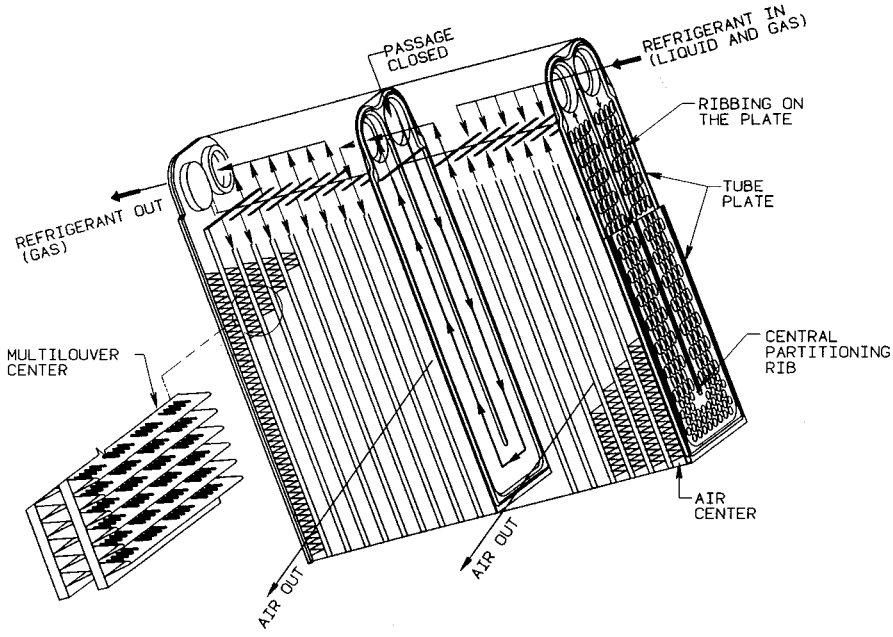


FIGURE 4.5.3 U-channel ribbed plates and multilouver fin automotive evaporator. (Courtesy of Delphi Harrison Thermal Systems, Lockport, NY.)

center configurations. Fins are die- or roll-formed and are attached to the plates by brazing, soldering, adhesive bonding, welding, mechanical fit, or extrusion. Fins may be used on both sides in gas-to-gas heat exchangers. In gas-to-liquid applications, fins are usually used only on the gas side; if employed on the liquid side, they are used primarily for structural strength and flow-mixing purposes. Fins are also sometimes used for pressure containment and rigidity.

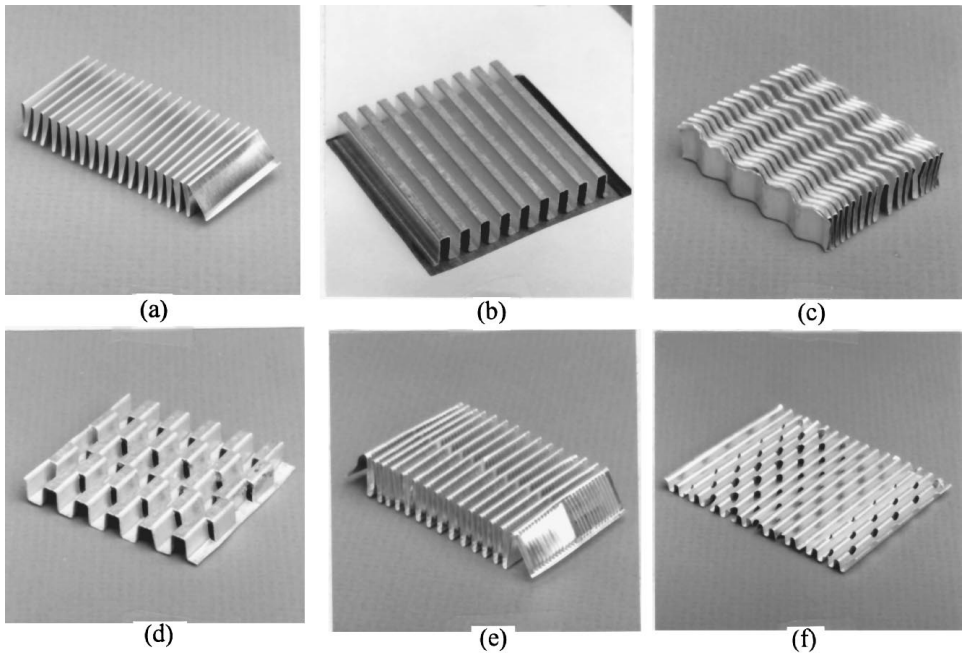


FIGURE 4.5.4 Fin geometries for plate-fin heat exchangers: (a) plain triangular fin, (b) plain rectangular fin, (c) wavy fin, (d) offset strip fin, (e) multilouver fin, and (f) perforated fin.

Plate fins are categorized as (1) plain (i.e., uncut) and straight fins, such as plain triangular and rectangular fins; (2) plain but wavy fins (wavy in the main fluid flow direction); and (3) interrupted fins such as offset strip, louver, and perforated. Examples of commonly used fins are shown in Figure 4.5.4.

Plate-fin exchangers have been built with a surface area density of up to about 5900 m^2/m^3 (1800 ft^2/ft^3). There is a total freedom of selecting fin surface area on each fluid side, as required by the design, by varying fin height and fin density. Although typical fin densities are 120 to 700 fins/m (3 to 18 fins/in.), applications exist for as many as 2100 fins/m (53 fins/in.). Common fin thicknesses range from 0.05 to 0.25 mm (0.002 to 0.010 in.). Fin heights range from 2 to 25 mm (0.08 to 1.0 in.). A plate-fin exchanger with 600 fins/m (15.2 fins/in.) provides about 1300 m^2 (400 ft^2/ft^3) of heat transfer surface area per cubic meter volume occupied by the fins. Plate-fin exchangers are manufactured in virtually all shapes and sizes, and made from a variety of materials.

Tube-Fin Heat Exchangers.

In this type of exchanger, round and rectangular tubes are the most common, although elliptical tubes are also used. Fins are generally used on the outside, but they may be used on the inside of the tubes in some applications. They are attached to the tubes by a tight mechanical fit, tension winding, adhesive bonding, soldering, brazing, welding, or extrusion. Fins on the outside of the tubes may be categorized as follows: (1) normal fins on individual tubes, referred to as individually finned tubes or simply as *finned tubes*, as shown in Figures 4.5.6 and 4.5.5a; (2) flat or continuous (plain, wavy, or interrupted) external fins on an array of tubes, as shown in Figures 4.5.7 and 4.5.5b; (3) longitudinal fins on individual tubes. The exchanger having flat (continuous) fins on tubes has also been referred to as a *plate-fin and tube* exchanger in the literature. In order to avoid confusion with plate-fin surfaces, we will refer to it as a tube-fin exchanger having flat (plain, wavy, or interrupted) fins. Individually finned tubes are probably more rugged and practical in large tube-fin exchangers. Shell-and-tube exchangers sometimes employ low-finned tubes to increase the surface area on the shell side when the shell-side heat transfer coefficient is low compared with the tube side coefficient. The exchanger with flat fins is usually less expensive on a unit heat transfer surface area basis because of its simple and mass-production-type construction

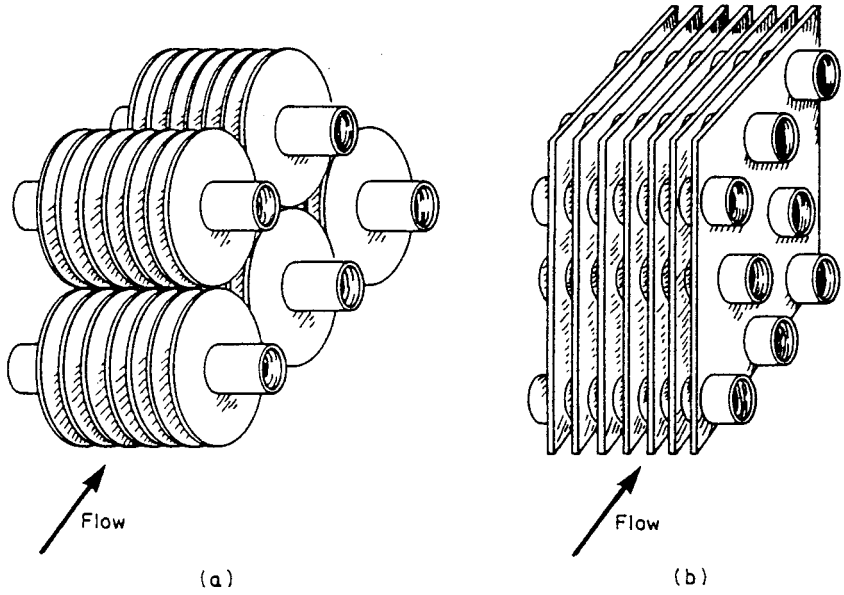


FIGURE 4.5.5 (a) Individually finned tubes, (b) flat or continuous fins on an array of tubes.

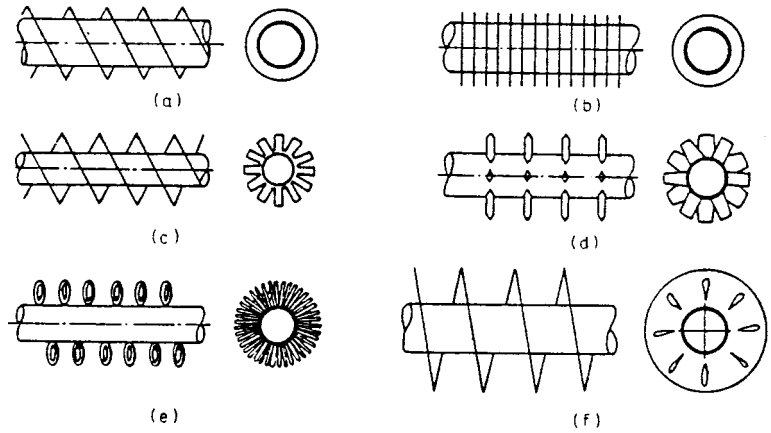


FIGURE 4.5.6 Individually finned tubes: (a) helical, (b) annular disk, (c) segmented, (d) studded, (e) wire loop, and (f) slotted helical.

features. Longitudinal fins are generally used in condensing applications and for viscous fluids in double-pipe heat exchangers.

Tube-fin exchangers can withstand high pressures on the tube side. The highest temperature is again limited by the type of bonding, the materials employed, and the material thickness. Tube-fin exchangers with an area density of about 3300 m^2/m^3 (1000 ft^2/ft^2) are commercially available. On the fin side, the desired surface area can be employed by using the proper fin density and fin geometry. The typical fin densities for flat fins vary from 250 to 800 fins/m (6 to 20 fins/in.), fin thicknesses vary from 0.08 to 0.25 mm (0.003 to 0.010 in.), and fin flow lengths from 25 to 250 mm (1 to 10 in.). A tube-fin exchanger having flat fins with 400 fins/m (10 fins/in.) has a surface area density of about 720 m^2/m^3(220 ft^2/ft^3). These exchangers are extensively used as condensers and evaporators in air-conditioning and refrigeration applications, as condensers in electric power plants, as oil coolers in propulsive power plants, and as air-cooled exchangers (also referred to as a fin-fan exchanger) in process and power industries.

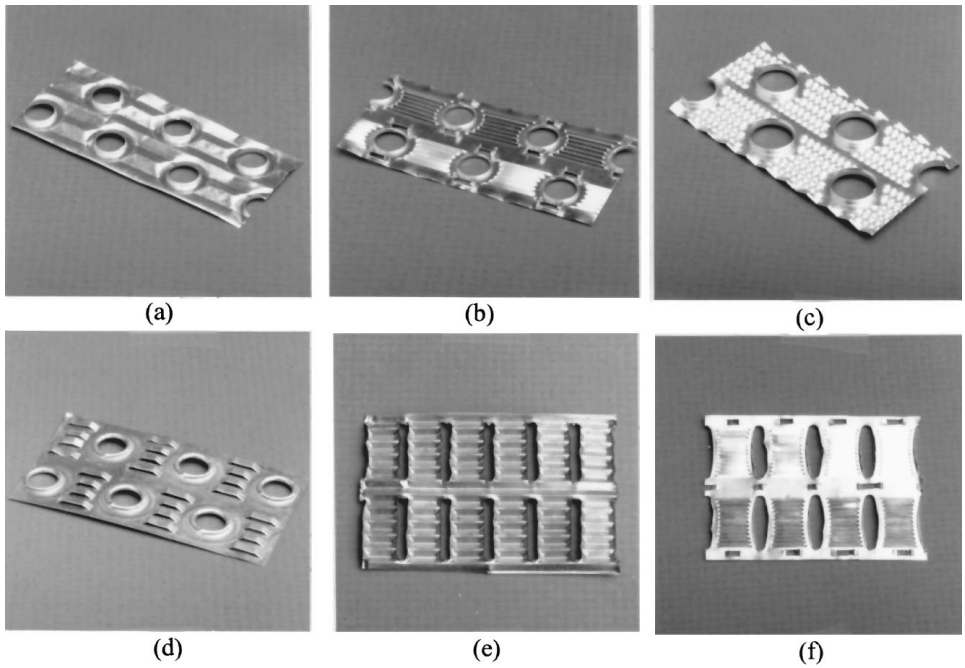


FIGURE 4.5.7 Flat or continuous fins on an array of tubes: On round tubes: (a) wavy fin, (b) multilouver fin, (c) fin with structured surface roughness (dimples), (d) parallel louver fin; (e) louver fin on flat tubes; (f) multilouver fin on elliptical tubes.

Regenerators.

The regenerator is a storage-type exchanger. The heat transfer surface or elements are usually referred to as a matrix in the regenerator. In order to have continuous operation, either the matrix must be moved periodically into and out of the fixed streams of gases, as in a *rotary* regenerator (Figure 4.5.8a), or the gas flows must be diverted through valves to and from the fixed matrices as in a *fixed-matrix* regenerator (Figure 4.5.8b). The latter is also sometimes referred to as a *periodic-flow regenerator* or a *reversible heat accumulator*. A third type of regenerator has a fixed matrix (in the disk form) and the fixed stream of gases, but the gases are ducted through rotating hoods (headers) to the matrix as shown in Figure 4.5.8c. This Rothemuhle regenerator is used as an air preheater in some power-generating plants. The thermodynamically superior counterflow arrangement is usually employed in regenerators.

The **rotary regenerator** is usually a disk type in which the matrix (heat transfer surface) is in a disk form and fluids flow axially. It is rotated by a hub shaft or a peripheral ring gear drive. For a rotary regenerator, the design of seals to prevent leakage of hot to cold fluids and vice versa becomes a difficult task, especially if the two fluids are at significantly differing pressures. Rotating drives also pose a challenging mechanical design problem.

Major advantages of rotary regenerators are the following. For a highly compact regenerator, the cost of the regenerator surface per unit of heat transfer area is usually substantially lower than that for the equivalent recuperator. A major disadvantage of a regenerator is an unavoidable carryover of a small fraction of the fluid trapped in the passage to the other fluid stream just after the periodic flow switching. Since fluid contamination (small mixing) is prohibited with liquids, the regenerators are used exclusively for gas-to-gas heat or energy recovery applications. Cross contamination can be minimized significantly by providing a purge section in the disk and using double-labyrinth seals.

Rotary regenerators have been designed for a surface area density of up to about 6600 m^2/m^3 (2000 ft^2/ft^3), and exchanger effectivenesses exceeding 85% for a number of applications. They can employ thinner stock material, resulting in the lowest amount of material for a given effectiveness and pressure

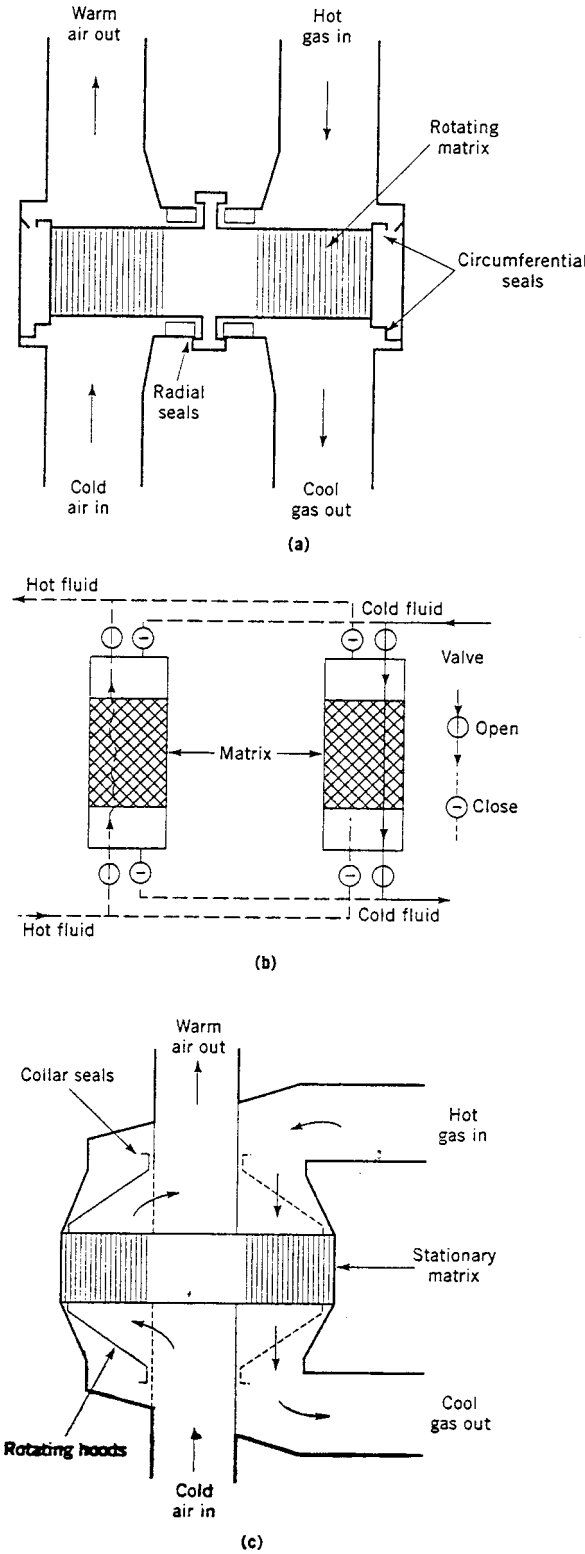


FIGURE 4.5.8 Regenerators: (a) rotary, (b) fixed-matrix, and (c) Rothemuhle.

drop of any heat exchanger known today. The metal rotary regenerators have been designed for continuous inlet temperatures up to about 790°C (1450°F) and ceramic matrices for higher-temperature applications; these regenerators are designed up to 400 kPa or 60 psi pressure differences between hot and cold gases. Plastic, paper, and wool are used for regenerators operating below 65°C (150°F) inlet temperature of the hot gas and 1 atm pressure. Typical regenerator rotor diameters and rotational speeds are as follows: up to 10 m (33 ft) and 0.5 to 3 rpm for power plant regenerators, 0.25 to 3 m (0.8 to 9.8 ft) and up to 10 rpm for air-ventilating regenerators, and up to 0.6 m (24 in.) and up to 18 rpm for vehicular regenerators. Refer to Shah (1994) for the description of **fixed-matrix regenerator**, also referred to as a *periodic-flow*, *fixed bed*, *valved*, or *stationary* regenerator.

Plate-Type Heat Exchangers.

These exchangers are usually built of thin plates (all prime surface). The plates are either smooth or have some form of corrugations, and they are either flat or wound in an exchanger. Generally, these exchangers cannot accommodate very high pressures, temperatures, and pressure and temperature differentials. These exchangers may be further classified as plate, spiral plate, lamella, and plate-coil exchangers as classified in Figure 4.5.1. The plate heat exchanger, being the most important of these, is described next.

The **plate-and-frame** or **gasketed PHE** consists of a number of thin rectangular corrugated or embossed metal plates sealed around the edges by gaskets and held together in a frame as shown in Figure 4.5.9. The plate pack with fixed and movable end covers is clamped together by long bolts, thus compressing the gaskets and forming a seal. Sealing between the two fluids is accomplished by elastomeric molded gaskets (typically 5 mm or 0.2 in. thick) that are fitted in peripheral grooves mentioned earlier. The most conventional flow arrangement is one pass to one pass counterflow with all inlet and outlet connections on the fixed end cover. By blocking flow through some ports with proper gasketing, either one or both fluids could have more than one pass. Also more than one exchanger can be accommodated in a single frame with the use of intermediate connector plates such as up to five "exchangers" or sections to heat, cool, and regenerate heat between raw milk and pasteurized milk in a milk pasteurization application.

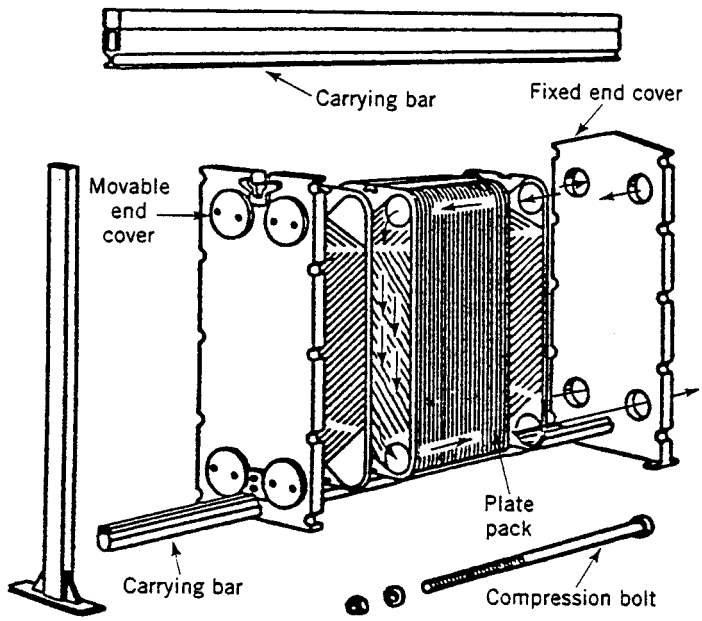


FIGURE 4.5.9 A plate-and-frame or gasketed PHE.

TABLE 4.5.1 Some Geometric and Operating Condition Characteristics of Plate-and-Frame Heat Exchangers

Unit		**Operation**	
Maximum surface area	2500 m^2	Pressure	0.1 to 2.5 MPa
Number of plates	3–700	Temperature	–40 to 260°C
Port size	Up to 400 mm	Maximum port velocity	6 m/sec
Plates		Channel flow rates	0.05 to 12.5 m^3/hr
Thickness	0.5–1.2 mm	Max unit flow rate	2500 m^3/hr
Size	0.03–3.6 m^2	**Performance**	
Spacing	1.5–5 mm	Temperature approach	As low as 1°C
Width	70–1200 mm	Heat exchanger efficiency	Up to 93%
Length	0.6–5 m	Heat transfer coefficients for water-water duties	3000 to 7000 W/m^2 K

Source: From Shah, R.K., in *Encyclopedia of Energy Technology and the Environment*, A. Bision and S.G. Boots, Eds., John Wiley & Sons, New York, 1994, 1651–1670. With permission.

Typical PHE dimensions and performance parameters are given in Table 4.5.1 (Shah, 1994). Any metal which can be cold-worked is suitable for PHE applications. The most common plate materials are stainless steel (AISI 304 or 316) and titanium. Plates made from Incoloy 825, Inconel 625, Hastelloy C-276 are also available. Nickel, cupronickel, and monel are rarely used. Carbon steel is not used because of low corrosion resistance for thin plates. The heat transfer surface area per unit volume for plate exchangers ranges from 120 to 660 m^2/m^3 (37 to 200 ft^2/ft^3).

In PHEs, the high turbulence due to plates reduces fouling from about 10 to 25% of that of a shell-and-tube exchanger. High thermal performance can be achieved in plate exchangers because the high degree of counterflow in PHEs makes temperature approaches of up to 1°C (2°F) possible. The high thermal effectiveness (up to about 93%) makes low-grade heat recovery economical. PHEs are most suitable for liquid-liquid heat transfer duties.

Welded PHEs. One of the limitations of gasketed PHE is the presence of the gaskets which restricts the use to compatible fluids and which limits operating temperatures and pressures. In order to overcome this limitation, a number of welded PHE designs have surfaced with a welded pair of plates for one or both fluid sides. However, the disadvantage of such design is the loss of disassembling flexibility on the fluid side where the welding is done. Essentially, welding is done around the complete circumference where the gasket is normally placed. A *stacked plate heat exchanger* is another welded PHE design from Pacinox in which rectangular plates are stacked and welded at the edges. The physical size limitations of PHEs (1.2 m wide × 4 m long max, 4 × 13 ft) are considerably extended to 1.5 m wide × 20 m long (5 × 66 ft) in this exchanger. A maximum surface area of (10,000 m^2 or over 100,000 ft^2) can be accommodated in one unit. The potential maximum operating temperature is 815°C (1500°F) with an operating pressure of up to 20 MPa (3000 psig) when the stacked plate assembly is placed in a cylindrical pressure vessel. For operating pressures below 2 MPa (300 psig) and operating temperatures below 200°C (400°F), the plate bundle is not contained in a pressure vessel, but is bolted between two heavy plates. Some of the applications of this exchanger are catalytic reforming, hydrosulfurization, crude distillation, synthesis converter feed effluent exchanger for methanol, propane condenser, etc.

A number of other PHE constructions have been developed to address some of the limitations of the conventional PHEs. A double-wall PHE is used to avoid mixing of the two fluids. A wide-gap PHE is used for fluids having high fiber content or coarse particles. A graphite PHE is used for highly corrosive fluids. A flow-flex exchanger has plain fins on one side between plates and the other side has conventional plate channels, and is used to handle asymmetric duties (flow rate ratio of 2 to 1 and higher).

A vacuum **brazed PHE** is a compact PHE for high-temperature and high-pressure duties, and it does not have gaskets, tightening bolts, frame, or carrying and guide bars. It simply consists of stainless steel plates and two end plates. The brazed unit can be mounted directly on piping without brackets and foundations.

Printed Circuit Heat Exchangers. This exchanger, as shown in Figure 4.5.10, has only primary heat transfer surfaces as PHEs. Fine grooves are made in the plate by using the same techniques as those employed for making printed electrical circuits. High surface area densities (650 to 1350 m^2/m^3 or 200 to 400 ft^2/ft^3 for operating pressures of 500 to 100 bar respectively) are achievable. A variety of materials including stainless steel, nickel, and titanium alloys can be used. It has been successfully used with relatively clean gases, liquids and phase-change fluids in chemical processing, fuel processing, waste heat recovery, and refrigeration industries. Again, this exchanger is a new construction with limited special applications currently.

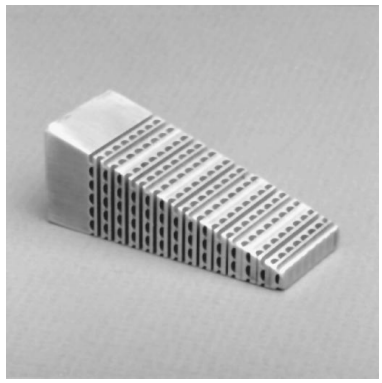


FIGURE 4.5.10 A section of a printed circuit heat exchanger. (Courtesy of Heatric Ltd., Dorset, U.K.)

Exchanger Heat Transfer and Pressure Drop Analysis

In this subsection, starting with the thermal circuit associated with a two-fluid exchanger, ε-NTU, P-NTU, and mean temperature difference (MTD) methods used for an exchanger analysis are presented, followed by the fin efficiency concept and various expressions. Finally, pressure drop expressions are outlined for various single-phase exchangers.

Two energy conservation differential equations for a two-fluid exchanger with any flow arrangement are (see Figure 4.5.11 for counterflow)

$$dq = q'' \, dA = -C_h \, dT_h = \pm C_c \, dT_c \tag{4.5.1}$$

where the ± sign depends upon whether dT_c is increasing or decreasing with increasing dA or dx. The local overall rate equation is

$$dq = q'' \, dA = U\left(T_h - T_c\right)_{\text{local}} \, dA = U \, \Delta T \, dA \tag{4.5.2}$$

Integration of Equations (4.5.1) and (4.5.2) across the exchanger surface area results in

$$q = C_h\left(T_{h,i} - T_{h,o}\right) = C_c\left(T_{c,o} - T_{c,i}\right) \tag{4.5.3}$$

and

$$q = UA\Delta T_m = \Delta T_m / R_o \tag{4.5.4}$$

where ΔT_m is the true mean temperature difference (or MTD) that depends upon the exchanger flow arrangement and degree of fluid mixing within each fluid stream. The inverse of the overall thermal conductance UA is referred to as the overall thermal resistance R_o as follows (see Figure 4.5.12).

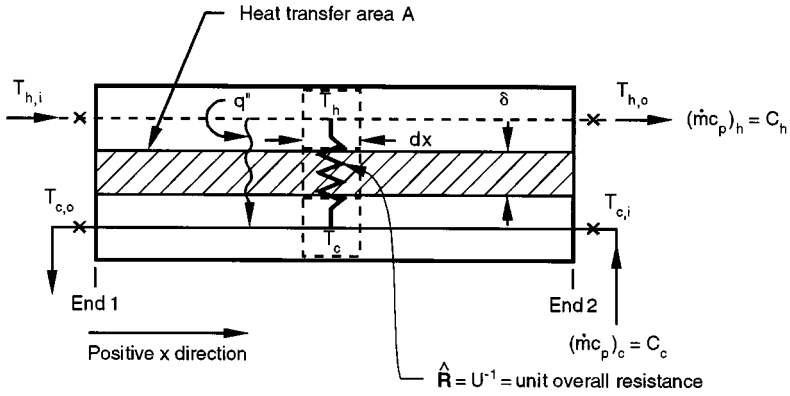


FIGURE 4.5.11 Nomenclature for heat exchanger variables.

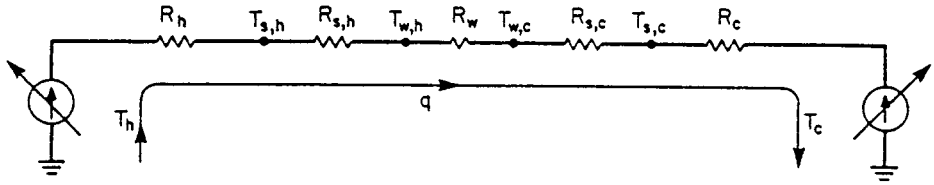


FIGURE 4.5.12 Thermal circuit for heat transfer in an exchanger.

$$R_o = R_h + R_{s,h} + R_w + R_{s,c} + R_c \tag{4.5.5}$$

where the subscripts *h*, *c*, *s*, and *w* denote hot, cold, fouling (or scale), and wall, respectively. In terms of the overall and individual heat transfer coefficients, Equation (4.5.5) is represented as

$$\frac{1}{UA} = \frac{1}{(\eta_o hA)_h} + \frac{1}{(\eta_o h_s A)_h} + R_w + \frac{1}{(\eta_o h_s A)_c} + \frac{1}{(\eta_o hA)_c} \tag{4.5.6}$$

where η_o = the overall surface efficiency of an extended (fin) surface and is related to the fin efficiency η_f, fin surface area A_f, and the total surface area A as follows:

$$\eta_o = 1 - \frac{A_f}{A}(1 - \eta_f) \tag{4.5.7}$$

The wall thermal resistance R_w of Equation (4.5.5) is given by

$$R_w = \begin{cases} \delta / A_w k_w & \text{for a flat wall} \\ \dfrac{\ln\left(d_o/d_j\right)}{2\pi k_w L N_t} & \text{for a circular tube with a single-layer wall} \\ \dfrac{1}{2\pi L N_t}\left[\displaystyle\sum_j \frac{\ln\left(d_{j+1}/d_j\right)}{k_{w,j}}\right] & \text{for a circular tube with a multiple-layer wall} \end{cases} \tag{4.5.8}$$

If one of the resistances on the right-hand side of Equation (4.5.5) or (4.5.6) is significantly higher than the other resistances, it is referred to as the *controlling thermal resistance*. A reduction in the controlling thermal resistance will have much more impact in reducing the exchanger surface area (A) requirement compared with the reduction in A as a result of the reduction in other thermal resistances.

UA of Equation (4.5.6) may be defined in terms of hot or cold fluid side surface area or wall conduction area as

$$UA = U_h A_h = U_c A_c = U_w A_w \tag{4.5.9}$$

When R_w is negligible, $T_{w,h} = T_{w,c} = T_w$ of Figure 4.5.12 is computed from

$$T_w = \frac{T_h + \left[\left(R_h + R_{s,h}\right)/\left(R_c + R_{s,c}\right)\right] T_c}{1 + \left[\left(R_h + R_{s,h}\right)/\left(R_c + R_{s,c}\right)\right]} \tag{4.5.10}$$

When $R_{s,h} = R_{s,c} = 0$, Equation (4.5.10) reduces to

$$T_w = \frac{T_h/R_h + T_c/R_c}{1/R_h + 1/R_c} = \frac{\left(\eta_o hA\right)_h T_h + \left(\eta_o hA\right)_c T_c}{\left(\eta_o hA\right)_h + \left(\eta_o hA\right)_c} \tag{4.5.11}$$

ε-NTU, P-NTU, and MTD Methods. If we consider the fluid outlet temperatures or heat transfer rate as dependent variables, they are related to independent variable/parameters of Figure 4.5.11 as follows.

$$T_{h,o},\ T_{c,o},\ \text{or}\ \ q = \phi\left\{T_{h,i},\ T_{c,i},\ C_c,\ C_h,\ U,\ A,\ \text{flow arrangement}\right\} \tag{4.5.12}$$

Six independent and three dependent variables of Equation (4.5.12) for a given flow arrangement can be transferred into two independent and one dependent dimensionless groups; three different methods are presented in Table 4.5.2 based on the choice of three dimensionless groups. The relationship among three dimensionless groups is derived by integrating Equations (4.5.1) and (4.5.2) across the surface area for a specified exchanger flow arrangement. Such expressions are presented later in Table 4.5.4 for the industrially most important flow arrangements. Now we briefly describe the three methods.

In *the ε-NTU method*, the heat transfer rate from the hot fluid to the cold fluid in the exchanger is expressed as

$$q = \varepsilon C_{\min}\left(T_{h,i} - T_{c,i}\right) \tag{4.5.13}$$

TABLE 4.5.2 General Functional Relationships and Dimensionless Groups for ε-NTU, P-NTU, and MTD Methods

ε-NTU Method	P-NTU Method[a]	MTD Method[a]
$q = \varepsilon C_{\min}(T_{h,i} - T_{c,i})$	$q = P_1 C_1 \lvert T_{1,i} - T_{2,i} \rvert$	$q = UAF\Delta T_{lm}$
$\varepsilon = \phi(\text{NTU}, C^*, \text{flow arrangement})$	$P_1 = \phi(\text{NTU}_1, R_1, \text{flow arrangement})$	$F = \phi(\text{P}, \text{R}, \text{flow arrangement})$[b]
$\varepsilon = \dfrac{C_h\left(T_{h,i} - T_{h,o}\right)}{C_{\min}\left(T_{h,i} - T_{c,i}\right)} = \dfrac{C_c\left(T_{c,o} - T_{c,i}\right)}{C_{\min}\left(T_{h,i} - T_{c,i}\right)}$	$P_1 = \dfrac{T_{1,o} - T_{1,i}}{T_{2,i} - T_{1,i}}$	$F = \dfrac{\Delta T_m}{\Delta T_{\text{lm}}}$
$\text{NTU} = \dfrac{UA}{C_{\min}} = \dfrac{1}{C_{\min}} \int_A U\, dA$	$\text{NTU}_1 = \dfrac{UA}{C_1} = \dfrac{\lvert T_{1,o} - T_{1,i} \rvert}{\Delta T_m}$	$\text{LMTD} = \Delta T_{\text{lm}} = \dfrac{\Delta T_1 - \Delta T_2}{\ln\left(\Delta T_1 / \Delta T_2\right)}$
$C^* = \dfrac{C_{\min}}{C_{\max}} = \dfrac{\left(\dot{m}c_p\right)_{\min}}{\left(\dot{m}c_p\right)_{\max}}$	$R_1 = \dfrac{C_1}{C_2} = \dfrac{T_{2,i} - T_{2,o}}{T_{1,o} - T_{1,i}}$	$\Delta T_1 = T_{h,i} - T_{c,o} \quad \Delta T_2 = T_{h,o} - T_{c,i}$

[a] Although P, R, and NTU_1 are defined on fluid side 1, it must be emphasized that all the results of the P-NTU and MTD methods are valid if the definitions of P, NTU, and R are consistently based on C_c, C_s, C_h, or C.

[b] P and R are defined in the P-NTU method.

Here the exchanger effectiveness ε is an efficiency factor. It is a ratio of the actual heat transfer rate from the hot fluid to the cold fluid in a given heat exchanger of any flow arrangement to the maximum possible heat transfer rate $q_{\max}$ thermodynamically permitted. The $q_{\max}$ is obtained in a *counterflow* heat exchanger (recuperator) of *infinite surface area* operating with the fluid flow rates (heat capacity rates) and fluid inlet temperatures equal to those of an actual exchanger (constant fluid properties are idealized). As noted in Table 4.5.1, the exchanger effectiveness ε is a function of NTU and C^* in this method. The number of transfer units NTU is a ratio of the overall conductance UA to the smaller heat capacity rate $C_{\min}$. NTU designates the dimensionless "heat transfer size" or "thermal size" of the exchanger. Other interpretations of NTU are given by Shah (1983). The heat capacity rate ratio C^* is simply a ratio of the smaller to the larger heat capacity rate for the two fluid streams. Note that $0 \le \varepsilon \le 1$, $0 \le \text{NTU} \le \infty$ and $0 \le C^* \le 1$.

The P-NTU method represents a variant of the ε-NTU method. The ε-NTU relationship is different depending upon whether the shell fluid is the $C_{\min}$ or $C_{\max}$ fluid in the (stream unsymmetric) flow arrangements commonly used for shell-and-tube exchangers. In order to avoid possible errors and to avoid keeping track of the $C_{\min}$ fluid side, an alternative is to present the temperature effectiveness P as a function of NTU and R, where P, NTU, and R are defined consistently either for Fluid 1 side or Fluid 2 side; in Table 4.5.2, they are defined for Fluid 1 side (regardless of whether that side is the hot or cold fluid side), and Fluid 1 side is clearly identified for each flow arrangement in Table 4.5.4; it is the shell side in a shell-and-tube exchanger. Note that

$$q = P_1 C_1 \left| T_{1,i} - T_{2,i} \right| = P_2 C_2 \left| T_{2,i} - T_{1,i} \right| \tag{4.5.14}$$

$$P_1 = P_2 R_2 \qquad P_2 = P_1 R_1 \tag{4.5.15}$$

$$\text{NTU}_1 = \text{NTU}_2 R_2 \qquad \text{NTU}_2 = \text{NTU}_1 R_1 \tag{4.5.16}$$

and

$$R_1 = 1/R_2 \tag{4.5.17}$$

In *the MTD method*, the heat transfer rate from the hot fluid to the cold fluid in the exchanger is given by

$$q = UA\Delta T_m = UAF\Delta T_{\text{lm}} \tag{4.5.18}$$

where ΔT_m the log-mean temperature difference (LMTD), and F the LMTD correction factor, a ratio of true (actual) MTD to the LMTD, where

$$\text{LMTD} = \Delta T_{\text{lm}} = \frac{\Delta T_1 - \Delta T_2}{\ln\left(\Delta T_1/\Delta T_2\right)} \tag{4.5.19}$$

Here ΔT_1 and ΔT_2 are defined as

$$\Delta T_1 = T_{h,i} - T_{c,o} \qquad \Delta T_2 = T_{h,o} - T_{c,i} \qquad \text{for all flow arrangements except for parallel flow} \tag{4.5.20}$$

$$\Delta T_1 = T_{h,i} - T_{c,i} \qquad \Delta T_2 = T_{h,o} - T_{c,o} \qquad \text{for parallel flow} \tag{4.5.21}$$

The LMTD represents a true MTD for a counterflow arrangement under the idealizations listed below. Thus, the LMTD correction factor F represents a degree of departure for the MTD from the counterflow LMTD; it does not represent the effectiveness of a heat exchanger. It depends on two dimensionless group P_1 and R_1 or P_2 and R_2 for a given flow arrangement.

TABLE 4.5.3 Relationships between Dimensionless Groups of the P-NTU and LMTD Methods and Those of the ε-NTU Method

$$\text{P}_1 = \frac{C_{\min}}{C_1}\varepsilon = \begin{cases} \varepsilon & \text{for} \quad C_1 = C_{\min} \\ \varepsilon C^* & \text{for} \quad C_1 = C_{\max} \end{cases}$$

$$\text{R}_1 = \frac{C_1}{C_2} = \begin{cases} C^* & \text{for} \quad C_1 = C_{\min} \\ 1/C^* & \text{for} \quad C_1 = C_{\max} \end{cases}$$

$$\text{NTU}_1 = \text{NTU}\frac{C_{\min}}{C_1} = \begin{cases} \text{NTU} & \text{for} \quad C_1 = C_{\min} \\ \text{NTU}\,C^* & \text{for} \quad C_1 = C_{\max} \end{cases}$$

$$F = \frac{\text{NTU}_{\text{cf}}}{\text{NTU}} = \frac{1}{\text{NTU}\left(1 - C^*\right)}\ln\left[\frac{1 - C^*\varepsilon}{1 - \varepsilon}\right] \xrightarrow{C^* = 1} \frac{\varepsilon}{\text{NTU}(1 - \varepsilon)}$$

$$F = \frac{1}{\text{NTU}_1\left(1 - \text{R}_1\right)}\ln\left[\frac{1 - \text{RP}_1}{1 - \text{P}_1}\right] \xrightarrow{\text{R}_1 = 1} \frac{\text{P}_1}{\text{NTU}_1\left(1 - \text{P}_1\right)}$$

The relationship among the dimensionless groups of the ε-NTU, P-NTU, and MTD methods are presented in Table 4.5.3. The closed-form formulas for industrially important exchangers are presented in terms of P_1, NTU_1, and R_1 in Table 4.5.4. These formulas are valid under idealizations which include: (1) steady-state conditions; (2) negligible heat losses to the surrounding; (3) no phase changes in the fluid streams flowing through the exchanger, or phase changes (condensation or boiling) occurring at constant temperature and constant effective specific heat; (4) uniform velocity and temperature at the entrance of the heat exchanger on each fluid side; (5) the overall extended surface efficiency η_o is uniform and constant; (6) constant individual and overall heat transfer coefficients; (7) uniformly distributed heat transfer area on each fluid side; (7) the number of baffles as large in shell-and-tube exchangers; (8) no flow maldistribution; and (9) negligible longitudinal heat conduction in the fluid and exchanger wall.

The overall heat transfer coefficient can vary as a result of variations in local heat transfer coefficients due to two effects: (1) change in heat transfer coefficients in the exchanger as a result of changes in the fluid properties or radiation due to rise or drop of fluid temperatures and (2) change in heat transfer coefficients in the exchanger due to developing thermal boundary layers; it is referred to as the *length effect*. The first effect due to fluid property variations (or radiation) consists of two components: (1) distortion of velocity and temperature profiles at a given flow cross section due to fluid property variations — this effect is usually taken into account by the so-called property ratio method, with the correction scheme of Equations (4.5.55) and (4.5.56) — and (2) variations in the fluid temperature along the axial and transverse directions in the exchanger depending upon the exchanger flow arrangement — this effect is referred to as the *temperature effect*. The resultant axial changes in the overall mean heat transfer coefficient can be significant; the variations in U_{local} could be nonlinear, dependent upon the type of the fluid. The effect of varying U_{local} can be taken into account by evaluating U_{local} at a few points in the exchanger and subsequently integrating U_{local} values by the Simpson or Gauss method (Shah, 1993). The temperature effect can increase or decrease mean U slightly or significantly, depending upon the fluids and applications. The length effect is important for developing laminar flows for which high heat transfer coefficients are obtained in the thermal entrance region. However, in general it will have less impact on the overall heat transfer coefficient because the other thermal resistances in series in an exchanger may be controlling. The length effect reduces the overall heat transfer coefficient compared with the mean value calculated conventionally (assuming uniform mean heat transfer coefficient on each fluid side). It is shown that this reduction is up to about 11% for the worst case (Shah, 1993).

Shah and Pignotti (1997) have shown that the following are the specific number of baffles beyond which the influence of the finite number of baffles on the exchanger effectiveness is not significantly larger than 2%: $N_b \geq 10$ for 1-1 TEMA E counterflow exchanger; $N_b \geq 6$ for 1-2 TEMA E exchanger for $NTU_s \leq 2$, $R_s \leq 5$; $N_b \geq 9$ for 1-2 TEMA J exchanger for $NTU_s \leq 2$, $R_s \leq 5$; $N_b \geq 5$ for 1-2 TEMA G exchanger for $NTU_s \leq 3$, all R_s; $N_b \geq 11$ for 1-2 TEMA H exchanger for $NTU_s \leq 3$, all R_s. Various shell-and-tube heat exchangers (such as TEMA E, G, H, J, etc.) are classified by the Tubular Exchanger Manufacturers' Association (TEMA, 1988).

If any of the basic idealizations are not valid for a particular exchanger application, the best solution is to work directly with either Equations 4.5.1 and 4.5.2 or their modified form by including a particular effect, and to integrate them over a small exchanger segment numerically in which all of the idealizations are valid.

Fin Efficiency and Extended Surface Efficiency.

Extended surfaces have fins attached to the primary surface on one or both sides of a two-fluid or a multifluid heat exchanger. Fins can be of a variety of geometries — plain, wavy, or interrupted — and can be attached to the inside, outside, or both sides of circular, flat, or oval tubes, or parting sheets. Fins are primarily used to increase the surface area (when the heat transfer coefficient on that fluid side is relatively low) and consequently to increase the total rate of heat transfer. In addition, enhanced fin geometries also increase the heat transfer coefficient compared to that for a plain fin. Fins may also be used on the high heat transfer coefficient fluid side in a heat exchanger primarily for structural strength purposes (for example, for high-pressure water flow through a flat tube) or to provide a thorough mixing of a highly viscous liquid (such as for laminar oil flow in a flat or a round tube). Fins are attached to the primary surface by brazing, soldering, welding, adhesive bonding, or mechanical expansion, or they are extruded or integrally connected to the tubes. Major categories of extended surface heat exchangers are plate-fin (Figures 4.5.2 to 4.5.4) and tube-fin (Figures 4.5.5 to 4.5.7) exchangers. Note that shell-and-tube exchangers sometimes employ individually finned tubes — low finned tubes (similar to Figure 4.5.5a but with low-height fins) (Shah, 1985).

The concept of fin efficiency accounts for the reduction in temperature potential between the fin and the ambient fluid due to conduction along the fin and convection from or to the fin surface depending upon the fin cooling or heating situation. The fin efficiency is defined as the ratio of the actual heat transfer rate through the fin base divided by the maximum possible heat transfer rate through the fin

TABLE 4.5.4 P_1 - NTU_1 Formulas and Limiting Values P_1 and $R_1 = 1$ and $NTU_1 \to \infty$ for Various Exchanger Flow Arrangements[a]

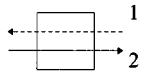
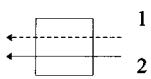
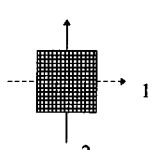
Flow Arrangement	Eq. no.	General formula	Value for $R_1 = 1$	Value for $NTU_1 \to \infty$
	I.1.1	$P_1 = \dfrac{1-\exp[-NTU_1(1-R_1)]}{1-R_1\exp[-NTU_1(1-R_1)]}$	$P_1 = \dfrac{NTU_1}{1+NTU_1}$	$P_1 \to 1$ for $R_1 \le 1$ $P_1 \to 1/R_1$ for $R_1 \ge 1$
	I.1.2	$NTU_1 = \dfrac{1}{(1-R_1)}\ln\left[\dfrac{1-R_1P_1}{1-P_1}\right]$	$NTU_1 = \dfrac{P_1}{1-P_1}$	$NTU_1 \to \infty$
Counterflow exchanger, stream symmetric.	I.1.3	$F = 1$	$F = 1$	$F = 1$
	I.2.1	$P_1 = \dfrac{1-\exp[-NTU_1(1+R_1)]}{1+R_1}$	$P_1 = \dfrac{1}{2}[1-\exp(-2NTU_1)]$	$P_1 \to \dfrac{1}{1+R_1}$
	I.2.2	$NTU_1 = \dfrac{1}{1+R_1}\ln\left[\dfrac{1}{1-P_1(1+R_1)}\right]$	$NTU_1 = \dfrac{1}{2}\ln\left[\dfrac{1}{1-2P_1}\right]$	$NTU_1 \to \infty$
Parallel flow exchanger, stream symmetric.	I.2.3	$F = \dfrac{(R_1+1)\ln\left[\dfrac{1-R_1P_1}{1-P_1}\right]}{(R_1-1)\ln[1-P_1(1+R_1)]}$	$F = \dfrac{2P_1}{(P_1-1)\ln(1-2P_1)}$	$F \to 0$
	II.1	$P_1 = 1-\exp(NTU_1) - \exp[-(1+R_1)NTU_1] \cdot \sum_{n=1}^{\infty} R_1^n P_n(NTU_1)$	same as Eq. (II.1) with $R_1 = 1$	$P_1 \to 1$ for $R_1 \le 1$ $P_1 \to \dfrac{1}{R_1}$ for $R_1 \ge 1$
Single-pass crossflow exchanger, both fluids unmixed, stream symmetric		$P_n(y) = \dfrac{1}{(n+1)!}\sum_{j=1}^{n}\dfrac{(n+1-j)}{j!}y^{n+j}$		

TABLE 4.5.4 (continued) P_1 - NTU_1 Formulas and Limiting Values P_1 and $R_1 = 1$ and $NTU_1 \to \infty$ for Various Exchanger Flow Arrangements[a]

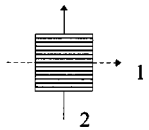
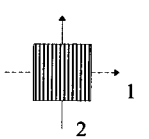
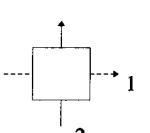
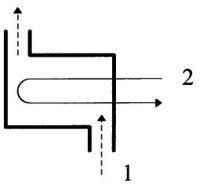
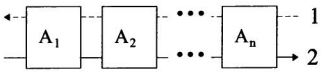
Flow Arrangement	Eq. no.	General formula	Value for $R_1 = 1$	Value for $NTU_1 \to \infty$
 Single-pass crossflow exchanger, fluid 1 unmixed, fluid 2 mixed.	II.2.1	$P_1 = [1-\exp(-KR_1)]/R_1$ $K = 1-\exp(-NTU_1)$	$P_1 = 1-\exp(-K)$	$P_1 \to \dfrac{1-\exp(-R_1)}{R_1}$
	II.2.2	$NTU = \ln\left[\dfrac{1}{1+\dfrac{1}{R_1}\ln(1-R_1P_1)}\right]$	$NTU_1 = \ln\left[\dfrac{1}{1+\ln(1-P_1)}\right]$	$NTU_1 \to \infty$
	II.2.3	$F = \dfrac{\ln[(1-R_1P_1)/(1-P_1)]}{(R_1-1)\ln\left[1+\dfrac{1}{R_1}\ln(1-R_1P_1)\right]}$	$F = \dfrac{P_1}{(P_1-1)\ln[1+\ln(1-P_1)]}$	$F \to 0$
 Single-pass crossflow exchanger, fluid 1 mixed, fluid 2 unmixed.	II.3.1	$P = 1-\exp(-K/R_1)$ $K = 1-\exp(-R_1NTU_1)$	$P = 1-\exp(-K)$ $K = 1-\exp(-NTU_1)$	$P_1 \to 1-\exp(-1/R_1)$
	II.3.2	$NTU_1 = \dfrac{1}{R_1}\ln\left[\dfrac{1}{1+R_1}\ln(1-P_1)\right]$	$NTU_1 = \ln\left[\dfrac{1}{1+\ln(1-P_1)}\right]$	$NTU_1 \to \infty$
	II.3.3	$F = \dfrac{\ln(1-R_1P_1)/(1-P_1)]}{(1-1/R_1)\ln[1+R_1\ln(1-P_1)]}$	$F = \dfrac{P_1}{(P_1-1)\ln[1+\ln(1-P_1)]}$	$P_1 \to \dfrac{1}{1+R_1}$
 Single-pass crossflow exchanger, both fluids mixed, stream symmetric.	II.4	$P_1 = \left[\dfrac{1}{K_1}+\dfrac{R_1}{K_2}-\dfrac{1}{NTU_1}\right]^{-1}$ $K_1 = 1-\exp(-NTU_1)$ $K_2 = 1-\exp(-R_1NTU_1)$	$P_1 = \left[\dfrac{2}{K_1}-\dfrac{1}{NTU_1}\right]^{-1}$	$P_1 \to \dfrac{1}{1+R_1}$

TABLE 4.5.4 (continued) P_1 - NTU_1 Formulas and Limiting Values P_1 and $R_1 = 1$ and $NTU_1 \to \infty$ for Various Exchanger Flow Arrangements[a]

Flow Arrangement	Eq. no.	General formula	Value for $R_1 = 1$	Value for $NTU_1 \to \infty$
 1-2 TEMA E shell-and-tube exchanger, shell fluid mixed, stream symmetric	III.1.1	$P_1 = \dfrac{2}{1+R_1+E\coth(E\,NTU_1/2)}$	$P_1 = \dfrac{1}{1+\coth(NTU_1/\sqrt{2})/\sqrt{2}}$	$P_1 \to \dfrac{2}{1+R_1+E}$
		$E = \left[1+R_1^2\right]^{1/2}$		
	III.1.2	$NTU_1 = \dfrac{1}{E}\ln\left[\dfrac{2-P_1(1+R_1-E)}{2-P_1(1+R_1+E)}\right]$	$NTU_1 = \ln\left[\dfrac{2-P_1}{2-3P_1}\right]$	$NTU_1 \to \infty$
	III.1.3	$F = \dfrac{E\ln[(1-R_1P_1)/(1-P_1)]}{(1-R_1)\ln\left[\dfrac{2-P_1(1+R_1-E)}{2-P_1(1+R_1+E)}\right]}$	$F = \dfrac{P_1/(1-P_1)}{\ln[(2-P_1)/(2-3P_1)]}$	$F \to 0$
 Series coupling of n exchangers, overall counterflow arrangement. Stream symmetric if all A_i are stream symmetric.	IV.1.1	$P_1 = \dfrac{\prod_{i=1}^{n}(1-R_1P_{1,A_i}) - \prod_{i=1}^{n}(1-P_{1,A_i})}{\prod_{i=1}^{n}(1-R_1P_{1,A_i}) - R_1\prod_{i=1}^{n}(1-P_{1,A_i})}$	$P_1 = \dfrac{\sum_{i=1}^{n}\frac{P_{1,A_i}}{1-P_{1,A_i}}}{1+\sum_{i=1}^{n}\frac{P_{1,A_i}}{1-P_{1,A_i}}}$	same as Eq. (I.1.1) counterflow
	IV.1.2	$R_1 = R_{1,A_i}, \quad i = 1,\ldots,n$	$1 = R_{1,A_i}, \quad i = 1,\ldots,n$	same as Eq. (IV.1.2)
	IV.1.3	$NTU_1 = \sum_{i=1}^{n} NTU_{1,A_i}$	same as for Eq. (IV.1.3)	same as Eq. (IV.1.3)
	IV.1.4	$F = \dfrac{1}{NTU_1}\sum_{i=1}^{n} NTU_{1,A_i}F_{A_i}$	same as Eq. (IV.1.4)	same as Eq. (IV.1.4)

[a] In this table, all variables, except P_1, R_1, NTU_1 and F, are local or dummy variables not necessarily related to similar ones defined in the nomenclature and the text. Source: Shah, R.K. and Mueller, A.C., 1988. With permission.

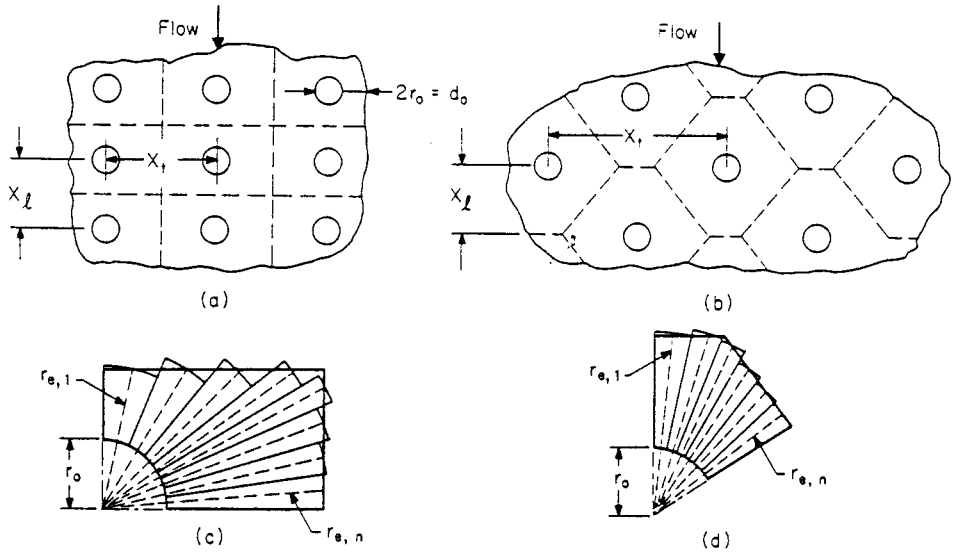


FIGURE 4.5.13 A flat fin over (a) an in-line and (b) staggered tube arrangement; the smallest representative segment of the fin for (c) an in-line and (d) a staggered tube arrangement.

base which would be obtained if the entire fin were at the base temperature (i.e., its material thermal conductivity were infinite). Since most of the real fins are "thin", they are treated as one-dimensional (1-D) with standard idealizations used for the analysis (Huang and Shah, 1992). This 1-D fin efficiency is a function of the fin geometry, fin material thermal conductivity, heat transfer coefficient at the fin surface, and the fin tip boundary condition; it is not a function of the fin base or fin tip temperature, ambient temperature, and heat flux at the fin base or fin tip in general. Fin efficiency formulas for some common fins are presented in Table 4.5.5 (Shah, 1985). Huang and Shah (1992) also discuss the influence on η_f if any of the basic idealizations used in the fin analysis are violated.

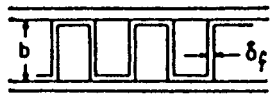
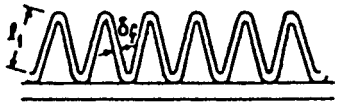
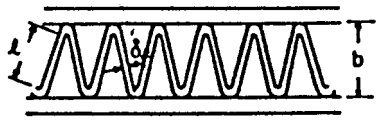
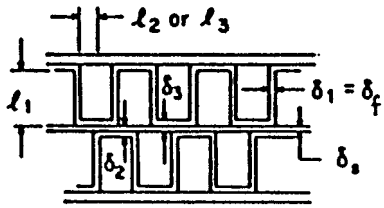
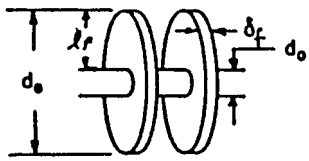
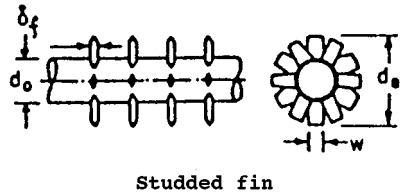
The fin efficiency for flat fins (Figure 4.5.5b) is obtained by a sector method (Shah, 1985). In this method, the rectangular or hexagonal fin around the tube (Figures 4.5.7a and b) or its smallest symmetrical section is divided into n sectors (Figure 4.5.13). Each sector is then considered as a circular fin with the radius $r_{e,i}$ equal to the length of the centerline of the sector. The fin efficiency of each sector is subsequently computed using the circular fin formula of Table 4.5.5. The fin efficiency η_f for the whole fin is then the surface area weighted average of $\eta_{f,i}$ of each sector.

$$\eta_f = \frac{\sum_{i=1}^{n} \eta_{f,i} A_{f,i}}{\sum_{i=1}^{n} A_{f,i}} \tag{4.5.22}$$

Since the heat flow seeks the path of least thermal resistance, actual η_f will be equal to or higher than that calculated by Equation (4.5.22); hence, Equation (4.5.22) yields a somewhat conservative value of η_f.

The η_f values of Table 4.5.5 or Equation (4.5.22) are not valid in general when the fin is thick, when it is subject to variable heat transfer coefficients or variable ambient fluid temperature, or when it has a temperature depression at the fin base. See Huang and Shah (1992) for details. For a thin rectangular fin of constant cross section, the fin efficiency as presented in Table 4.5.5 is given by

TABLE 4.5.5 Fin Efficiency Expressions for Plate-Fin and Tube-Fin Geometries of Uniform Fin Thickness

Geometry	Fin Efficiency Formula
	$m_i = \left[\frac{2h}{k_f \delta_i}\left(1+\frac{\delta_i}{l_f}\right)\right]^{1/2} \qquad E_1 = \frac{\tanh(m_i l_1)}{m_i l_1} \qquad i = 1, 2$
 Plain, wavy, or offset strip fin of rectangular cross section	$\eta_f = E_1$ $l_1 = \frac{b}{2} - \delta_1 \quad \delta_1 = \delta_f$
 Triangular fin heated from one side	$\eta_f = \frac{hA_1(T_0 - T_a)\frac{\sinh(m_1 l_1)}{m_1 l_1} + q_e}{\cosh(m_1 l_1)\left[hA_1(T_0 - T_a) + q_e \frac{T_0 - T_a}{T_1 - T_a}\right]}$
 Plain, wavy, or louver fin of triangular cross section	$\eta_f = E_1$ $l_1 = \frac{l}{2} \quad \delta_1 = \delta_f$
 Double sandwich fin	$\eta_f = \frac{E_1 l_1 + E_2 l_2}{l_1 + l_2}\frac{1}{1 + m_1^2 E_1 E_2 l_1 l_2}$ $\delta_1 = \delta_f \quad \delta_2 = \delta_3 = \delta_f + \delta_s$ $l_1 = b - \delta_f + \frac{\delta_s}{2} \quad l_2 = l_3 = \frac{p_f}{2}$
 Circular fin	$\eta_j = \begin{cases} a(ml_e)^{-b} & \text{for } \Phi > 0.6 + 2.257(r^*)^{-0.445} \\ \frac{\tanh\Phi}{\Phi} & \text{for } \Phi \le 0.6 + 2.257(r^*)^{-0.445} \end{cases}$ $a = (r^*)^{-0.246} \qquad \Phi = ml_e(r^*)^{\exp(0.13ml_e - 1.3863)}$ $b = \begin{cases} 0.9107 + 0.0893 r^* & \text{for } r^* \le 2 \\ 0.9706 + 0.17125 \ln r^* & \text{for } r^* > 2 \end{cases}$ $m = \left(\frac{2h}{k_f \delta_f}\right)^{1/2} \quad l_e = l_j + \frac{\delta_f}{2} \quad r^* = \frac{d_e}{d_o}$
 Studded fin	$\eta_j = \frac{\tanh(ml_e)}{ml_e}$ $m = \left[\frac{2h}{k_f \delta_f}\left(1 + \frac{\delta_f}{w}\right)\right]^{1/2} \quad l_e = l_j + \frac{\delta_f}{2} \quad l_f = \frac{(d_e - d_o)}{2}$

$$\eta_f = \frac{\tanh(m\ell)}{m\ell} \tag{4.5.23}$$

where $m = [2h(1 + \delta_f \ell_f)/k_f\delta_f]^{1/2}$. For a thick rectangular fin of constant cross section, the fin efficiency (a counterpart of Equation (4.5.23) is given by (Huang and Shah, 1992)

$$\eta_f = \frac{\left(\mathrm{Bi}^+\right)^{1/2}}{K\mathrm{Bi}} \tanh\left[K\left(\mathrm{Bi}^+\right)^{1/2}\right] \tag{4.5.24}$$

where $\mathrm{Bi}^+ = \mathrm{Bi}/(1 + \mathrm{Bi}/4)$, $\mathrm{Bi} = (h\delta_f/2k_f)^{1/2}$, $K = 2\ell/\delta_f$. Equation (4.5.23) is accurate (within 0.3%) for a "thick" rectangular fin having $\eta_f > 80\%$; otherwise, use Equation (4.5.24) for a thick fin.

In an extended-surface heat exchanger, heat transfer takes place from both the fins ($\eta_f < 100\%$) and the primary surface ($\eta_f = 100\%$). In that case, the total heat transfer rate is evaluated through a concept of extended surface efficiency η_o defined as

$$\eta_o = \frac{A_p}{A} + \eta_f \frac{A_f}{A} = 1 - \frac{A_f}{A}\left(1 - \eta_f\right) \tag{4.5.25}$$

where A_f is the fin surface area, A_p is the primary surface area, and $A = A_f + A_p$. In Equation 4.5.25, heat transfer coefficients over the finned and unfinned surfaces are idealized to be equal. Note that $\eta_o \geq \eta_f$ and η_o is always required for the determination of thermal resistances of Equation (4.5.5) in heat exchanger analysis.

Pressure Drop Analysis.

Usually a fan, blower, or pump is used to flow fluid through individual sides of a heat exchanger. Due to potential initial and operating high cost, low fluid pumping power requirement is highly desired for gases and viscous liquids. The fluid pumping power $\wp$ is approximately related to the core pressure drop in the exchanger as (Shah, 1985).

$$\wp = \frac{\dot{m}\Delta p}{\rho} \approx \begin{cases} \dfrac{1}{2g_c}\dfrac{\mu}{\rho^2}\dfrac{4L}{D_h} f\,\mathrm{Re} & \text{for laminar flow} \quad (4.5.26) \\ \dfrac{0.046}{2g_c}\dfrac{\mu^{0.2}}{\rho^2}\dfrac{4L}{D_h}\dfrac{\dot{m}^{2.8}}{A_0^{1.8}D_h^{0.2}} & \text{for turbulent flow} \quad (4.5.27) \end{cases}$$

It is clear from Equations (4.2.26) and (4.2.27) that the fluid pumping power is strongly dependent upon the fluid density ($\wp \propto 1/\rho^2$) particularly for low-density fluids in laminar and turbulent flows, and upon the viscosity in laminar flow. In addition, the pressure drop itself can be an important consideration when blowers and pumps are used for the fluid flow since they are head limited. Also for condensing and evaporating fluids, the pressure drop affects the heat transfer rate. Hence, the pressure drop determination in the exchanger is important.

The pressure drop associated with a heat exchanger consists of (1) core pressure drop, and (2) the pressure drop associated with the fluid distribution devices such as inlet and outlet manifolds, headers, tanks, nozzles, ducting, and so on, which may include bends, valves, and fittings. This second Δp component is determined from Idelchik (1994) and Miller (1990). The core pressure drop may consist of one or more of the following components depending upon the exchanger construction: (1) friction losses associated with fluid flow over heat transfer surface (this usually consists of skin friction, form (profile) drag, and internal contractions and expansions, if any); (2) the momentum effect (pressure drop or rise due to fluid density changes) in the core; (3) pressure drop associated with sudden contraction and expansion at the core inlet and outlet, and (4) the gravity effect due to the change in elevation

between the inlet and outlet of the exchanger. The gravity effect is generally negligible for gases. For vertical flow through the exchanger, the pressure drop or rise ("static head") due to the elevation change is given by

$$\Delta p = \pm \frac{\rho_m g L}{g_c} \tag{4.5.28}$$

Here the "+" sign denotes vertical upflow (i.e., pressure drop), the "–" sign denotes vertical downflow (i.e., pressure rise). The first three components of the core pressure drop are now presented for plate-fin, tube-fin, plate, and regenerative heat exchangers.

Plate-fin heat exchangers. For the plate-fin exchanger (Figure 4.5.2), all three components are considered in the core pressure drop evaluation as follows:

$$\frac{\Delta p}{p_i} = \frac{G^2}{2g_c}\frac{1}{p_i\rho_i}\left[\left(1-\sigma^2+K_c\right)+f\frac{L}{r_h}\rho_i\left(\frac{1}{\rho}\right)_m+2\left(\frac{\rho_i}{\rho_o}-1\right)-\left(1-\sigma^2-K_e\right)\frac{\rho_i}{\rho_o}\right] \tag{4.5.29}$$

where f is the Fanning friction factor, K_c and K_e are flow contraction (entrance) and expansion (exit) pressure loss coefficients (see Figure 4.5.14), and σ is a ratio of minimum free flow area to frontal area. K_c and K_e for four different entrance flow passage geometries are presented by Kays and London (1984). The entrance and exit losses are important at low values of σ and L (short cores), high values of Re, and for gases; they are negligible for liquids. The values of K_c and K_e apply to long tubes for which flow is fully developed at the exit. For partially developed flows, K_c and K_e is higher than that for fully developed flows. For interrupted surfaces, flow is never a fully developed boundary layer type. For highly interrupted fin geometries, the entrance and exit losses are generally small compared to the core pressure drop and the flow is well mixed; hence, K_c and K_e for Re $\to \infty$ should represent a good approximation. The mean specific volume v_m or $(1/\rho)_m$ in Equation (4.5.29) is given as follows. For liquids with any flow arrangement, or for a perfect gas with $C^* = 1$ and any flow arrangement (except for parallel flow),

$$\left(\frac{1}{\rho}\right)_m = v_m = \frac{v_i + v_o}{2} = \frac{1}{2}\left(\frac{1}{\rho_i}+\frac{1}{\rho_o}\right) \tag{4.5.30}$$

where v is the specific volume in m³/kg. For a perfect gas with $C^* = 0$ and any flow arrangement,

$$\left(\frac{1}{\rho}\right)_m = \frac{\tilde{R}}{p_{\text{ave}}}T_{\text{lm}} \tag{4.5.31}$$

Here $\tilde{R}$ is the gas constant in J/(kg K), $p_{\text{ave}} = (p_i + p_o)/2$, and $T_{\text{lm}} = T_{\text{const}} + \Delta T_{\text{lm}}$ where T_{const} is the mean average temperature of the fluid on the other side of the exchanger; the LMTD ΔT_{lm} is defined in Table 4.5.2. The core frictional pressure drop in Equation 4.5.29 may be approximated as

$$\Delta p \approx \frac{4fLG^2}{2g_c D_h}\left(\frac{1}{\rho}\right)_m \tag{4.5.32}$$

Tube-fin heat exchangers. The pressure drop inside a circular tube is computed using Equation (4.5.29) with proper values of f factors, and K_c and K_e from Figure (4.5.18) for circular tubes.

For flat fins on an array of tubes (see Figure 4.5.5b), the components of the core pressure drop (such as those in Equation 4.5.29) are the same with the following exception: the core friction and momentum effect take place within the core with $G = \dot{m}/A_0$, where A_o is the minimum free-flow area within the

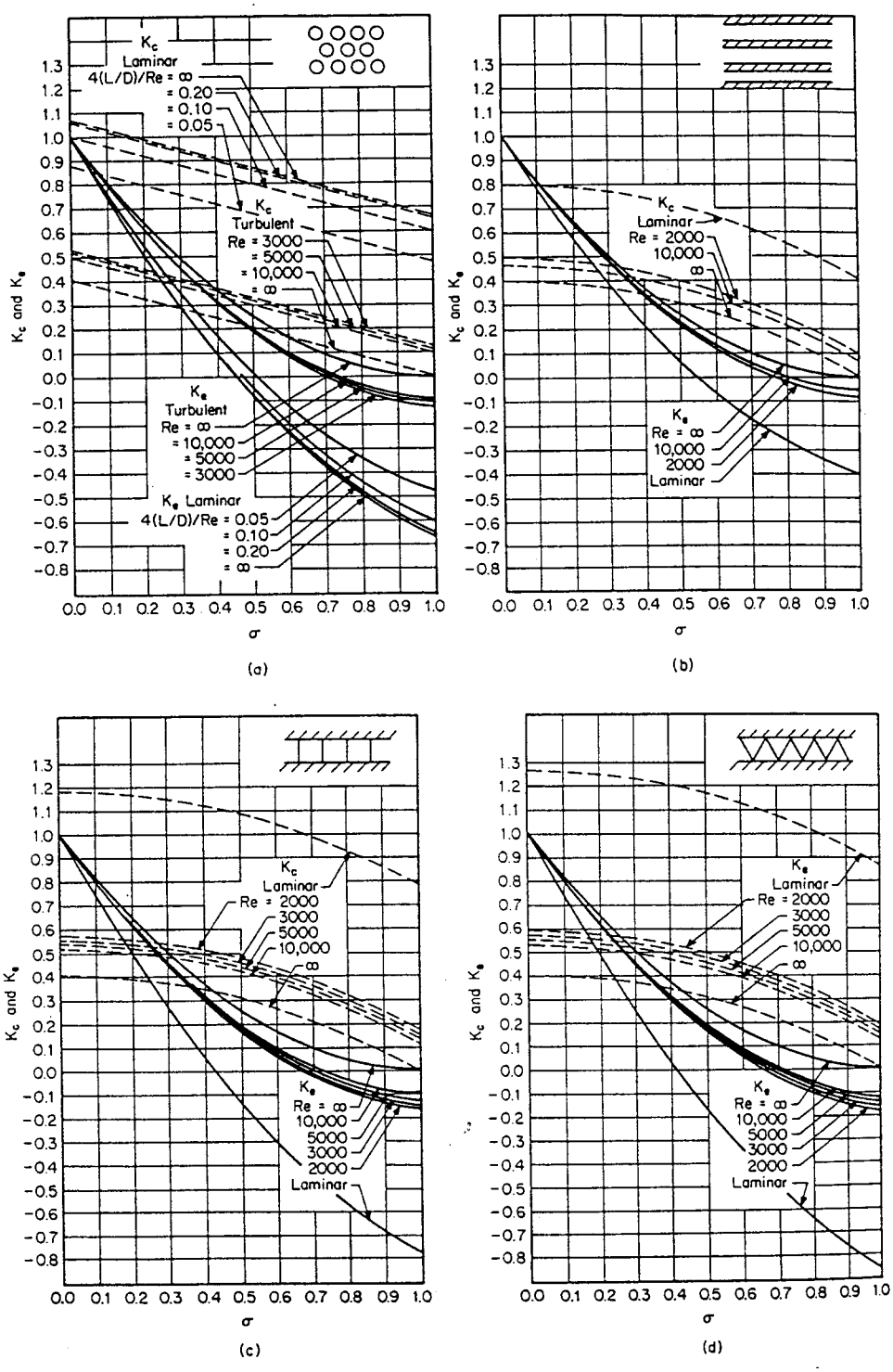


FIGURE 4.5.14 Entrance and exit pressure loss coefficients: (a) circular tubes, (b) parallel plates, (C) square passages, and (d) triangular passages. (From Kays, W.M. and London, A.L., *Compact Heat Exchangers*, 3rd ed., McGraw-Hill, New York, 1984. With permission.) For each of these flow passages, shown in the inset, the fluid flows perpendicular to the plane of the paper into the flow passages.

core, and the entrance and exit losses occur at the leading and trailing edges of the core with the associated flow area A_0' such that

$$\dot{m} = GA_o = G'A_o' \quad \text{or} \quad G'\sigma' = G\sigma \tag{4.5.33}$$

where σ' is the ratio of free-flow area to frontal area at the fin leading edges. The pressure drop for flow normal to a tube bank with flat fins is then given by

$$\frac{\Delta p}{p_i} = \frac{G^2}{2g_c}\frac{1}{p_i\rho_i}\left[f\frac{L}{r_h}\rho_i\left(\frac{1}{\rho}\right)_m + 2\left(\frac{\rho_i}{\rho_o}-1\right)\right] + \frac{G'^2}{2g_c}\frac{1}{p_i\rho_i}\left[\left(1-\sigma'^2-K_c\right)-\left(1-\sigma'^2-K_e\right)\frac{\rho_i}{\rho_o}\right] \tag{4.5.34}$$

For individually finned tubes as shown in Figure 4.5.5a, flow expansion and contraction take place along each tube row, and the magnitude is of the same order as that at the entrance and exit. Hence, the entrance and exit losses are generally lumped into the core friction factor. Equation (4.5.29) then reduces for individually finned tubes to

$$\frac{\Delta p}{p_i} = \frac{G^2}{2g_c}\frac{1}{p_i\rho_i}\left[f\frac{L}{r_h}\rho_i\left(\frac{1}{\rho}\right)_m + 2\left(\frac{\rho_i}{\rho_o}-1\right)\right] \tag{4.5.35}$$

Regenerators. For regenerator matrices having cylindrical passages, the pressure drop is computed using Equation (4.5.29) with appropriate values of f, K_c, and K_e. For regenerator matrices made up of any porous material (such as checkerwork, wire, mesh, spheres, copper wools, etc.), the pressure drop is calculated using Equation (4.5.35) in which the entrance and exit losses are included in the friction factor f.

Plate heat exchangers. Pressure drop in a PHE consists of three components: (1) pressure drop associated with the inlet and outlet manifolds and ports, (2) pressure drop within the core (plate passages), and (3) pressure drop due to the elevation change. The pressure drop in the manifolds and ports should be kept as low as possible (generally < 10%, but it is found as high as 25 to 30% of higher in some designs). Empirically, it is calculated as approximately 1.5 times the inlet velocity head per pass. Since the entrance and exit losses in the core (plate passages) cannot be determined experimentally, they are included in the friction factor for the given plate geometry. The pressure drop (rise) caused by the elevation change for liquids is given by Equation (4.5.28). Hence, the pressure drop on one fluid side in a PHE is given by

$$\Delta p = \frac{1.5G^2N_p}{2g_c\rho_i} + \frac{4fLG^2}{2g_cD_e}\left(\frac{1}{\rho}\right)_m + \left(\frac{1}{\rho_o}-\frac{1}{\rho_i}\right)\frac{G^2}{g_c} \pm \frac{\rho_m gL}{g_c} \tag{4.5.36}$$

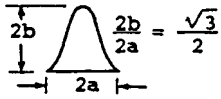
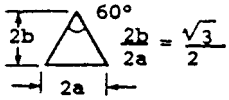
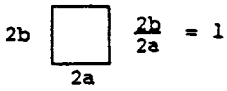
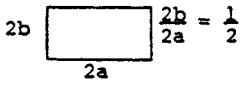
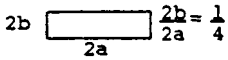
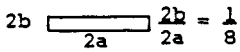
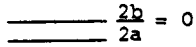
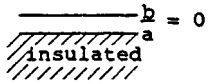
where N_p is the number of passes on the given fluid side and D_e is the equivalent diameter of flow passages (usually twice the plate spacing). Note that the third term on the right-hand side of the equality sign of Equation (4.5.36) is for the momentum effect which is generally negligible in liquids.

Heat Transfer and Flow Friction Correlations

Accurate and reliable surface heat transfer and flow friction characteristics are a key input to the exchanger heat transfer and pressure drop analyses or to the rating and sizing problems (Shah, 1985). Some important analytical solutions and empirical correlations are presented next for selected exchanger geometries.

The heat transfer rate in laminar duct flow is very sensitive to the thermal boundary condition. Hence, it is essential to identify carefully the thermal boundary condition in laminar flow. The heat transfer rate in turbulent duct flow is insensitive to the thermal boundary condition for most common fluids (Pr >

TABLE 4.5.6 Solutions for Heat Transfer and Friction for Fully Developed Flow-Through Specified Ducts

Geometry ($L/D_h > 100$)	Nu_{H1}	Nu_{H2}	Nu_T	fRe	j_{H1}/f [a]	Nu_{H1}/Nu_T
	3.014	1.474	2.39[b]	12.630	0.269	1.26
	3.111	1.892	2.47	13.333	0.263	1.26
	3.608	3.091	2.976	14.227	0.286	1.21
	4.002	3.862	3.34[b]	15.054	0.299	1.20
	4.123	3.017	3.391	15.548	0.299	1.22
	4.364	4.364	3.657	16.000	0.307	1.19
	5.331	2.94	4.439	18.233	0.329	1.20
	6.490	2.94	5.597	20.585	0.355	1.16
	8.235	8.235	7.541	24.000	0.386	1.09
	5.385	—	4.861	24.000	0.253	1.11

[a] This heading is the same as $Nu_{H1}\ Pr^{-1/3}/f\ Re$ with Pr = 0.7.

[b] Interpolated values.

0.7); the exception is liquid metals (Pr < 0.03). Hence, there is generally no need to identify the thermal boundary condition in turbulent flow for all fluids except liquid metals.

Fully developed laminar flow analytical solutions for some duct shapes of interest in compact heat exchangers are presented in Table 4.5.6 for three important thermal boundary conditions denoted by the subscripts *H*1, *H*2, and *T* (Shah and London, 1978; Shah and Bhatti, 1987). Here, *H*1 denotes constant axial wall heat flux with constant peripheral wall temperature, *H*2 denotes constant axial and peripheral wall heat flux, and *T* denotes constant wall temperature. The entrance effects, flow maldistribution, free convection, property variation, fouling, and surface roughness all affect fully developed analytical solutions. In order to account for these effects in real plate-fin plain fin geometries having fully developed flows, it is best to reduce the magnitude of the analytical Nu by at least 10% and to increase the value of the analytical *f*Re by 10% for design purposes.

The initiation of transition flow, the lower limit of the critical Reynolds number (Re_{crit}), depends upon the type of entrance (e.g., smooth vs. abrupt configuration at the exchanger flow passage entrance). For a sharp square inlet configuration, Re_{crit} is about 10 to 15% lower than that for a rounded inlet configuration. For most exchangers, the entrance configuration would be sharp. Some information on Re_{crit} is provided by Ghajar and Tam (1994).

Transition flow and fully developed turbulent flow Fanning friction factors (within ±2% accuracy) are given by Bhatti and Shah (1987) as

$$f = A + B\mathrm{Re}^{-1/m} \tag{4.5.37}$$

where

$$A = 0.0054, \quad B = 2.3\times10^{-8}, \quad m = -2/3, \quad \text{for} \quad 2100 \le \mathrm{Re} \le 4000$$

$$A = 0.00128, \quad B = 0.1143, \quad m = 3.2154, \quad \text{for} \quad 4000 \le \mathrm{Re} \le 10^7$$

The transition flow and fully developed turbulent flow Nusselt number correlation for a circular tube is given by Gnielinski as reported in Bhatti and Shah (1987) as

$$\mathrm{Nu} = \frac{(f/2)(\mathrm{Re}-1000)\mathrm{Pr}}{1+12.7(f/2)^{1/2}\left(\mathrm{Pr}^{2/3}-1\right)} \tag{4.5.38}$$

which is accurate within about ±10% with experimental data for $2300 \le \mathrm{Re} \le 5\times10^6$ and $0.5 \le \mathrm{Pr} \le 2000$.

A careful observation of accurate experimental friction factors for all noncircular smooth ducts reveals that ducts with laminar $f\mathrm{Re} < 16$ have turbulent f factors lower than those for the circular tube; whereas ducts with laminar $f\mathrm{Re} > 16$ have turbulent f factors higher than those for the circular tube (Shah and Bhatti, 1988). Similar trends are observed for the Nusselt numbers. Within ±15% accuracy, Equations (4.5.37) and (4.5.38) for f and Nu can be used for noncircular passages with the hydraulic diameter as the characteristic length in f, Nu, and Re; otherwise, refer to Bhatti and Shah (1987) for more accurate results for turbulent flow.

For hydrodynamically and thermally developing flows, the analytical solutions are boundary condition dependent (for laminar flow heat transfer only) and geometry dependent. The hydrodynamic entrance lengths for developing laminar and turbulent flows are given by Shah and Bhatti (1987) and Bhatti and Shah (1987) as

$$\frac{L_{hy}}{D_h} = \begin{cases} 0.0565\mathrm{Re} & \text{for laminar flow } (\mathrm{Re} \le 2100) \quad (4.5.39) \\ 1.359\mathrm{Re}^{1/4} & \text{for turbulent flow } \left(\mathrm{Re} \ge 10^4\right) \quad (4.5.40) \end{cases}$$

Analytical results are useful for well-defined constant-cross-sectional surfaces with essentially unidirectional flows. The flows encountered in heat exchangers are generally very complex having flow separation, reattachment, recirculation, and vortices. Such flows significantly affect Nu and f for the specific exchanger surfaces. Since no analytical or accurate numerical solutions are available, the information is derived experimentally. Kays and London (1984) and Webb (1994) present most of the experimental results reported in the open literature. In the following, empirical correlations for only some important surfaces are summarized due to space limitations. Refer to section 4.2, subsection External Flow Forced Convection for bare tube banks.

Plate-Fin Extended Surfaces.

Offset strip fins. This is one of the most widely used enhanced fin geometries (Figure 4.5.15) in aircraft, cryogenics, and many other industries that do not require mass production. This surface has one of the highest heat transfer performances relative to the friction factor. The most comprehensive correlations for j and f factors for the offset strip fin geometry is provided by Manglik and Bergles (1995) as follows.

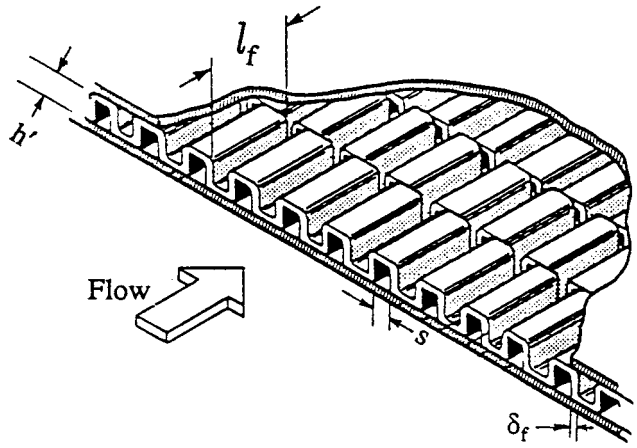


FIGURE 4.5.15 An offset strip fin geometry.

$$j = 0.6522\mathrm{Re}^{-0.5403}\left(\frac{s}{h'}\right)^{-0.1541}\left(\frac{\delta_f}{l_f}\right)^{0.1499}\left(\frac{\delta_f}{s}\right)^{-0.0678} \times\left[1+5.269\times10^{-5}\mathrm{Re}^{1.340}\left(\frac{s}{h'}\right)^{0.504}\left(\frac{\delta_f}{l_f}\right)^{0.456}\left(\frac{\delta_f}{s}\right)^{-1.055}\right]^{0.1} \tag{4.5.41}$$

$$f = 9.6243\mathrm{Re}^{-0.7422}\left(\frac{s}{h'}\right)^{-0.1856}\left(\frac{\delta_f}{l_f}\right)^{0.3053}\left(\frac{\delta_f}{s}\right)^{-0.2659} \times\left[1+7.669\times10^{-8}\mathrm{Re}^{4.429}\left(\frac{s}{h'}\right)^{0.920}\left(\frac{\delta_f}{l_f}\right)^{3.767}\left(\frac{\delta_f}{s}\right)^{0.236}\right]^{0.1} \tag{4.5.42}$$

where

$$D_h = 4A_o/\left(A/l_f\right) = 4sh'l_f/\left[2\left(sl_f + h'l_f + \delta_f h'\right) + \delta_f s\right] \tag{4.5.43}$$

Geometric symbols in Equation (4.5.43) are shown in Figure 4.5.15.

These correlations predict the experimental data of 18 test cores within ±20% for $120 \le \mathrm{Re} \le 10^4$. Although all the experimental data for these correlations are obtained for air, the *j* factor takes into consideration minor variations in the Prandtl number, and the above correlations should be valid for $0.5 < \mathrm{Pr} < 15$.

Louver fins. Louver or multilouver fins are extensively used in the auto industry because of their mass production manufacturability and hence lower cost. The louver fin has generally higher *j* and *f* factors than those for the offset strip fin geometry, and also the increase in the friction factors is in general higher than the increase in the *j* factors. However, the exchanger can be designed for higher heat transfer and the same pressure drop compared to that with the offset strip fins by a proper selection of exchanger frontal area, core depth, and fin density. Published literature and correlations on the louver fins are summarized by Webb (1994) and Cowell et al. (1995), and the understanding of flow and heat transfer phenomena is summarized by Cowell et al. (1995). Because of the lack of systematic studies reported

in the open literature on modern louver fin geometries, no correlation can be recommended for the design purpose.

Tube-Fin Extended Surfaces.

Two major types of tube-fin extended surfaces as shown in Figure 4.5.5 are (1) individually finned tubes and (2) flat fins (also sometimes referred to as plate fins) with or without enhancements/interruptions on an array of tubes. An extensive coverage of the published literature and correlations for these extended surfaces are provided by Webb (1994), Kays and London (1984), and Rozenman (1976). Empirical correlations for some important geometries are summarized below.

Individually finned tubes. This fin geometry, helically wrapped (or extruded) circular fins on a circular tube as shown in Figure 4.5.5a, is commonly used in process and waste heat recovery industries. The following correlation for j factors is recommended by Briggs and Young (see Webb, 1994) for individually finned tubes on staggered tube banks.

$$j = 0.134\mathrm{Re}_d^{-0.319}\left(s/l_f\right)^{0.2}\left(s/\delta_f\right)^{0.11} \tag{4.5.44}$$

where l_f is the radial height of the fin, δ_f the fin thickness, $s = p_f - \delta_f$ is the distance between adjacent fins and p_f is the fin pitch. Equation (4.5.44) is valid for the following ranges: $1100 \le \mathrm{Re}_d \le 18{,}000$, $0.13 \le s/l_f \le 0.63$, $1.01 \le s/\delta_f \le 6.62$, $0.09 \le l_f/d_o \le 0.69$, $0.011 \le \delta_f/d_o \le 0.15$, $1.54 \le X_t/d_o \le 8.23$, fin root diameter d_o between 11.1 and 40.9 mm, and fin density N_f (= $1/p_f$) between 246 and 768 fin/m. The standard deviation of Equation (4.5.44) with experimental results was 5.1%.

For friction factors, Robinson and Briggs (see Webb, 1994) recommended the following correlation:

$$f_{tb} = 9.465\mathrm{Re}_d^{-0.316}\left(X_t/d_o\right)^{-0.927}\left(X_t/X_d\right)^{0.515} \tag{4.5.45}$$

Here $X_d = (X_t^2 + X_l^2)^{1/2}$ is the diagonal pitch, and X_t and X_l are the transverse and longitudinal tube pitches, respectively. The correlation is valid for the following ranges: $2000 \le \mathrm{Re}_d \le 50{,}000$, $0.15 \le s/l_f \le 0.19$, $3.75 \le s/\delta_f \le 6.03$, $0.35 \le l_f/d_o \le 0.56$, $0.011 \le \delta_f/d_o \le 0.025$, $1.86 \le X_t/d_o \le 4.60$, $18.6 \le d_o \le 40.9$ mm, and $311 \le N_f \le 431$ fin/m. The standard deviation of Equation (4.5.45) with correlated data was 7.8%

For crossflow over low-height finned tubes, a simple but accurate correlation for heat transfer is given by Ganguli and Yilmaz (1987) as

$$j = 0.255\mathrm{Re}_d^{-0.3}\left(d_e/s\right)^{-0.3} \tag{4.5.46}$$

A more accurate correlation for heat transfer is given by Rabas and Taborek (1987). Chai (1988) provides the best correlation for friction factors:

$$f_{tb} = 1.748\mathrm{Re}_d^{-0.233}\left(\frac{l_f}{s}\right)^{0.552}\left(\frac{d_o}{X_t}\right)^{0.599}\left(\frac{d_o}{X_1}\right)^{0.1738} \tag{4.5.47}$$

This correlation is valid for $895 < \mathrm{Re}_d < 713{,}000$, $20 < \theta < 40°$, $X_t/d_o < 4$, $N \ge 4$, and θ is the tube layout angle. It predicts 89 literature data points within a mean absolute error of 6%; the range of actual error is from –16.7 to 19.9%.

Flat plain fins on a staggered tubebank. This geometry, as shown in Figure 4.5.5b, is used in air-conditioning/refrigeration industry as well as where the pressure drop on the fin side prohibits the use of enhanced/interrupted flat fins. An inline tubebank is generally not used unless very low fin side pressure drop is the essential requirement. Heat transfer correlation for Figure 4.5.5b flat plain fins on

staggered tubebanks is provided by Gray and Webb (see Webb, 1994) as follows for four or more tube rows.

$$j_4 = 0.14\mathrm{Re}_d^{-0.328}\left(X_t/X_l\right)^{-0.502}\left(s/d_o\right)^{0.031} \tag{4.5.48}$$

For the number of tube rows N from 1 to 3, the j factor is lower and is given by

$$\frac{j_N}{j_4} = 0.991\left[2.24\mathrm{Re}_d^{-0.092}\left(N/4\right)^{-0.031}\right]^{0.607(4-N)} \tag{4.5.49}$$

Gray and Webb (see Webb, 1994) hypothesized the friction factor consisting of two components: one associated with the fins and the other associated with the tubes as follows.

$$f = f_f\frac{A_f}{A} + f_t\left(1-\frac{A_f}{A}\right)\left(1-\frac{\delta_f}{p_f}\right) \tag{4.5.50}$$

where

$$f_f = 0.508\mathrm{Re}_d^{-0.521}\left(X_t/d_o\right)^{1.318} \tag{4.5.51}$$

and f_t (defined the same way as f) is the Fanning friction factor associated with the tube and can be determined from Eu of Figure 19 of Zukaukas (1987) as $f_t = \mathrm{EuN}(X_t - d_o)/\pi d_o$. Equation (4.5.50) correlated 90% of the data for 19 heat exchangers within ±20%. The range of dimensionless variables of Equations (4.5.50) and (4.5.51) are $500 \le \mathrm{Re} \le 24{,}700$, $1.97 \le X_t/d_o \le 2.55$, $1.7 \le X_l/d_o \le 2.58$, and $0.08 \le s/d_o \le 0.64$.

Exchanger Design Methodology

The problem of heat exchanger design is complex and multidisciplinary (Shah, 1991). The major design considerations for a new heat exchanger include process/design specifications, thermal and hydraulic design, mechanical design, manufacturing and cost considerations, and trade-offs and system-based optimization, as shown in Figure 4.5.16 with possible strong interactions among these considerations as indicated by double-sided arrows. The thermal and hydraulic design methods are mainly analytical, and the structural design is analytical to some extent. Most of the other major design considerations involve qualitative and experience-based judgments, trade-offs, and compromises. Therefore, there is no unique solution to designing a heat exchanger for given process specifications. Further details on this design methodology is given by Shah (1991).

Two important heat exchanger design problems are the rating and sizing problems. Determination of heat transfer and pressure drop performance of either an existing exchanger or an already sized exchanger is referred to as the rating problem. The objective here is to verify vendor's specifications or to determine the performance at off-design conditions. The rating problem is also sometimes referred to as the performance problem. In contrast, the design of a new or existing type of exchanger is referred to as the sizing problem. In a broad sense, it means the determination of the exchanger construction type, flow arrangement, heat transfer surface geometries and materials, and the physical size of an exchanger to meet the specified heat transfer and pressure drops. However, from the viewpoint of quantitative thermal-hydraulic analysis, we will consider that the selection of the exchanger construction type, flow arrangement, and materials has already been made. Thus, in the sizing problem, we will determine the physical size (length, width, height) and surface areas on each side of the exchanger. The sizing problem is also sometimes referred to as the design problem.

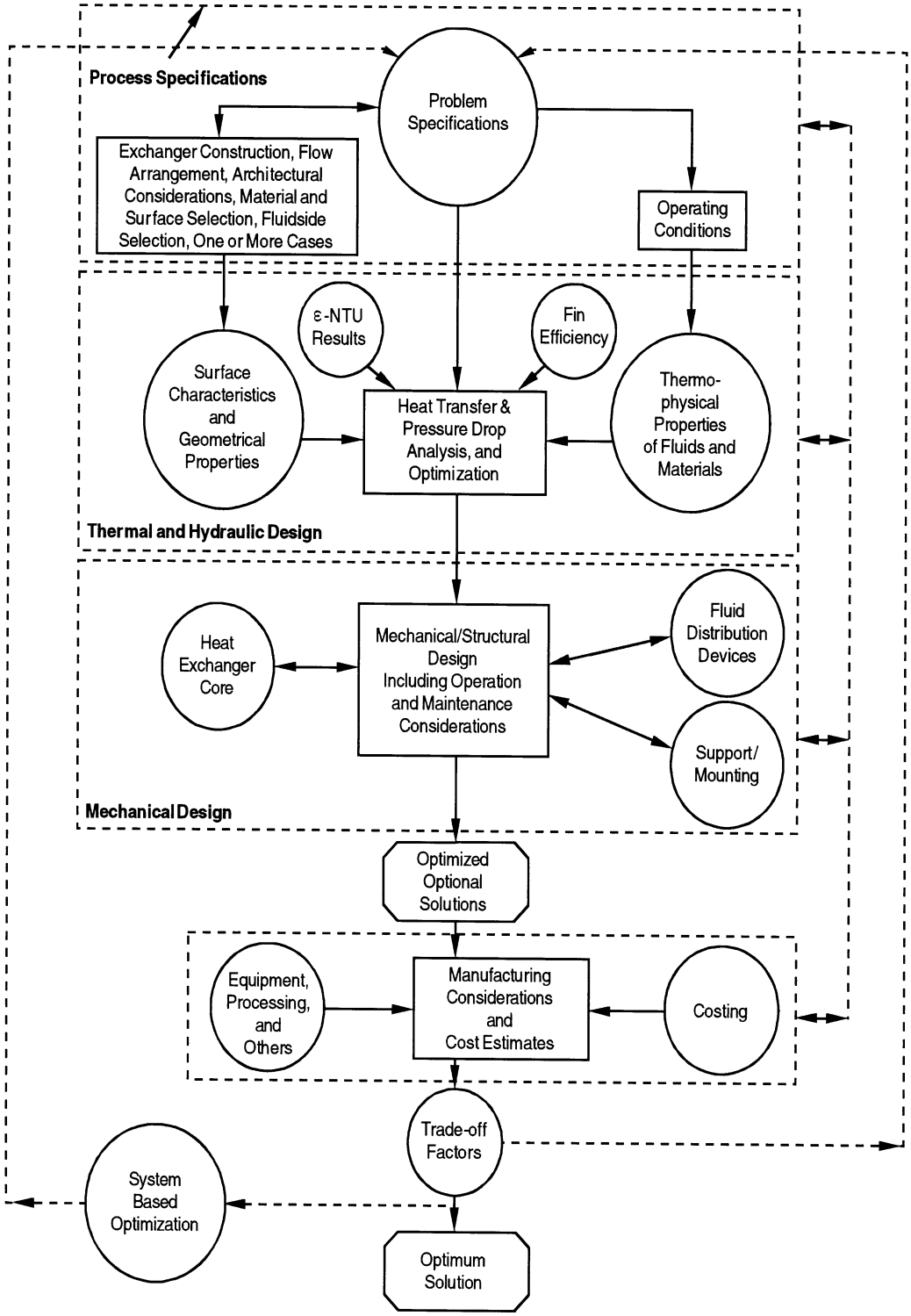


FIGURE 4.5.16 Heat exchanger design methodology.

The step-by-step solution procedures for the rating and sizing problems for counterflow and crossflow single-pass plate-fin heat exchangers have been presented with a detailed illustrative example by Shah (1981). Shah (1988a) presented further refinements in these procedures as well as step-by-step procedures for two-pass cross-counterflow plate-fin exchangers, and single-pass crossflow and two-pass cross-counterflow tube-fin exchangers. Also, step-by-step solution procedures for the rating and sizing problems for rotary regenerators (Shah, 1988b), heat pipe heat exchangers (Shah and Giovannelli, 1988) and PHEs (Shah and Wanniarachchi, 1991) are available. As an illustration, the step-by-step solution procedures will be covered here for a single-pass crossflow exchanger.

Rating Problem for a Crossflow Plate-Fin Exchanger.
Following is a step-by-step procedure for rating a crossflow plate-fin exchanger. Inputs to the rating problem for a two-fluid exchanger are the exchanger construction, flow arrangement and overall dimensions, complete details on the materials and surface geometries on both sides including their nondimensional heat transfer and pressure drop characteristics (j and f vs. Re), fluid flow rates, inlet temperatures, and fouling factors. The fluid outlet temperatures, total heat transfer rate, and pressure drops on each side of the exchanger are then determined as the rating problem solution.

1. Determine the surface geometric properties on each fluid side. This includes the minimum free-flow area A_o, heat transfer surface area A (both primary and secondary), flow lengths L, hydraulic diameter D_h, heat transfer surface area density β, the ratio of minimum free-flow area to frontal area σ, fin length l_f, and fin thickness δ for fin efficiency determination, and any specialized dimensions used for heat transfer and pressure drop correlations.
2. Compute the fluid bulk mean temperature and fluid thermophysical properties on each fluid side. Since the outlet temperatures are not known for the rating problem, they are estimated initially. Unless it is known from past experience, assume an exchanger effectiveness as 60 to 75% for most single-pass crossflow exchangers, or 80 to 85% for single-pass counterflow exchangers. For the assumed effectiveness, calculate the fluid outlet temperatures.

$$T_{h,o} = T_{h,i} - \varepsilon\left(C_{\min}/C_h\right)\left(T_{h,i} - T_{c,i}\right) \tag{4.5.52}$$

$$T_{c,o} = T_{c,i} - \varepsilon\left(C_{\min}/C_c\right)\left(T_{h,i} - T_{c,i}\right) \tag{4.5.53}$$

Initially, assume $C_c/C_h = \dot{m}_c/\dot{m}_h$ for a gas-to-gas exchanger, or $C_c/C_h = \dot{m}_c c_{p,c}/\dot{m}_h c_{p,h}$ for a gas-to-liquid exchanger with very approximate values of c_p for the fluids in question.

For exchangers with $C^* > 0.5$ (usually gas-to-gas exchangers), the bulk mean temperatures on each fluid side will be the arithmetic mean of the inlet and outlet temperatures on each fluid side (Shah, 1981). For exchangers with $C^* > 0.5$ (usually gas-to-gas exchangers), the bulk mean temperature on the $C_{\max}$ side will be the arithmetic mean of inlet and outlet temperatures; the bulk mean temperature on the $C_{\min}$ side will be the log-mean average temperature obtained as follows:

$$T_{m,C_{\min}} = T_{m,C_{\max}} \pm \Delta T_{\text{lm}} \tag{4.5.54}$$

where ΔT_{lm} is the LMTD based on the terminal temperatures (see Equation 4.5.19). Use the plus sign if the $C_{\min}$ side is hot; otherwise, use the negative sign.

Once the bulk mean temperature is obtained on each fluid side, obtain the fluid properties from thermophysical property books or from handbooks. The properties needed for the rating problem are μ, c_p, k, Pr, and ρ. With this c_p, one more iteration may be carried out to determine $T_{h,o}$ or $T_{c,o}$ from Equation 4.5.52 or 4.5.53 on the $C_{\max}$ side, and subsequently T_m on the $C_{\max}$ side, and refine fluid properties accordingly.

3. Calculate the Reynolds number Re = GD_h/μ and/or any other pertinent dimensionless groups (from the basic definitions) needed to determine the nondimensional heat transfer and flow friction characteristics (e.g., j or Nu and f) of heat transfer surfaces on each side of the exchanger. Subsequently, compute j or Nu and f factors. Correct Nu (or j) for variable fluid property effects (Shah, 1981) in the second and subsequent iterations from the following equations.

For gases:
$$\frac{\text{Nu}}{\text{Nu}_{\text{cp}}} = \left[\frac{T_w}{T_m}\right]^{n'} \qquad \frac{f}{f_{\text{cp}}} = \left[\frac{T_w}{T_m}\right]^{m'} \tag{4.5.55}$$

For liquids:
$$\frac{\text{Nu}}{\text{Nu}_{\text{cp}}} = \left[\frac{\mu_w}{\mu_m}\right]^{n'} \qquad \frac{f}{f_{\text{cp}}} = \left[\frac{\mu_w}{\mu_m}\right]^{m'} \tag{4.5.56}$$

where the subscript cp denotes constant properties, and m' and n' are empirical constants provided in Table 4.5.7. Note that T_w and T_m in Equations (4.5.55) and (4.5.56) and in Tables 4.5.7a and b and are absolute temperatures.

TABLE 4.5.7a Property Ratio Method Exponents of Equations (4.5.55) and (4.5.56) for Laminar Flow

Fluid	Heating	Cooling
Gases	$n' = 0.00$, $m' = 1.00$ for $1 < T_w/T_m < 3$	$n' = 0.0$, $m' = 0.81$ for $0.5 < T_w/T_m < 1$
Liquids	$n' = -0.14$, $m' = 0.58$ for $\mu_w/\mu_m < 1$	$n' = -0.14$, $m' = 0.54$ for $\mu_w/\mu_m > 1$

Source: Shah, R.K., in *Heat Exchangers: Thermal-Hydraulic Fundamentals and Design*, S. Kakaç et al., Eds., Hemisphere Publishing, Washington, D.C., 1981. With permission.

TABLE 4.5.7b Property Ratio Method Correlations of Exponents of Equations (4.5.55) and (4.5.56) for Turbulent Flow

Fluid	Heating	Cooling
Gases	$\text{Nu} = 5 + 0.012\,\text{Re}^{0.83}(\text{Pr} + 0.29)(T_w/T_m)^n$	$n' = 0$
	$n = -[\log_{10}(T_w/T_m)]^{1/4} + 0.3$ for $1 < T_w/T_m < 5$, $0.6 < \text{Pr} < 0.9$, $10^4 < \text{Re} < 10^6$, and $L/D_h > 40$	
	$m' = -0.1$ for $1 < T_w/T_m < 2.4$	$m' = -0.1$ (tentative)
Liquids	$n' = -0.11$[a] for $0.08 < \mu_w/\mu_m < 1$	$n' = -0.25$[a] for $1 < \mu_w/\mu_m < 40$
	$f/f_{\text{cp}} = (7 - \mu_m/\mu_w)/6$[b] or $m' = 0.25$ for $0.35 < \mu_w/\mu_m < 1$	$m' = 0.24$[b] for $1 < \mu_w/\mu_m < 2$

[a] Valid for $2 \le \text{Pr} \le 140$, $10^4 \le \text{Re} \le 1.25 \times 10^5$.

[b] Valid for $1.3 \le \text{Pr} \le 10$, $10^4 \le \text{Re} \le 2.3 \times 10^5$.

Source: Shah, R.K., in *Heat Exchangers: Thermal-Hydraulic Fundamentals and Design*, S. Kakaç et al., Eds., Hemisphere Publishing, Washington, D.C., 1981. With permission.

4. From Nu or j, compute the heat transfer coefficients for both fluid streams.

$$h = \text{Nu}\,k/D_h = jGc_p\text{Pr}^{-2/3} \tag{4.5.57}$$

Subsequently, determine the fin efficiency η_f and the extended surface efficiency η_o

$$\eta_f = \frac{\tanh m\ell}{m\ell} \quad \text{where} \quad m^2 = \frac{h\tilde{P}}{k_f A_k} \tag{4.5.58}$$

where $\tilde{P}$ is the wetted perimeter of the fin surface.

$$\eta_o = 1 - \frac{A_f}{A}\left(1 - \eta_f\right) \tag{4.5.59}$$

Also calculate the wall thermal resistance $R_w = \delta/A_w k_w$. Finally, compute the overall thermal conductance UA from Equation (4.5.6) knowing the individual convective film resistances, wall thermal resistances, and fouling resistances, if any.

5. From the known heat capacity rates on each fluid side, compute $C^* = C_{\min}/C_{\max}$. From the known UA, determine NTU = $UA/C_{\min}$. Also calculate the longitudinal conduction parameter λ. With the known NTU, C^*, λ, and the flow arrangement, determine the exchanger effectiveness ε from either closed-form equations of Table 4.5.4 or tabular/graphical results from Kays and London (1984).
6. With this ε, finally compute the outlet temperatures from Equations (4.5.52) and (4.5.53). If these outlet temperatures are significantly different from those assumed in Step 2, use these outlet temperatures in Step 2 and continue iterating Steps 2 to 6, until the assumed and computed outlet temperatures converge within the desired degree of accuracy. For a gas-to-gas exchanger, most probably one or two iterations will be sufficient.
7. Finally, compute the heat duty from

$$q = \varepsilon C_{\min}\left(T_{h,i} - T_{c,i}\right) \tag{4.5.60}$$

8. For the pressure drop calculations, first we need to determine the fluid densities at the exchanger inlet and outlet (ρ_i and ρ_o) for each fluid. The mean specific volume on each fluid side is then computed from Equation (4.5.30).

   Next, the entrance and exit loss coefficients, K_c and K_e, are obtained from Figure 4.5.14 for known σ, Re, and the flow passage entrance geometry.

   The friction factor on each fluid side is corrected for variable fluid properties using Equation (4.5.55) or (4.5.56). Here, the wall temperature T_w is computed from

$$T_{w,h} = T_{m,h} - \left(R_h + R_{s,h}\right)q \tag{4.5.61}$$

$$T_{w,c} = T_{m,c} + \left(R_c + R_{s,c}\right)q \tag{4.5.62}$$

where the various resistance terms are defined by Equation (4.5.6).

The core pressure drops on each fluid side are then calculated from Equation (4.5.29). This then completes the procedure for solving the rating problem.

Sizing Problem for a Crossflow Plate-Fin Exchanger.

As defined earlier, we will concentrate here to determine the physical size (length, width, and height) of a single-pass crossflow exchanger for specified heat duty and pressure drops. More specifically, inputs to the sizing problem are surface geometries (including their nondimensional heat transfer and pressure drop characteristics), fluid flow rates, inlet and outlet fluid temperatures, fouling factors, and pressure drops on each side.

For the solution to this problem, there are four unknowns — two flow rates or Reynolds numbers (to determine correct heat transfer coefficients and friction factors) and two surface areas — for the two-

fluid crossflow exchanger. The following four equations (Equations (4.5.63), (4.5.65), and (4.5.67) are used to solve iteratively the surface areas on each fluid side: *UA* in Equation (4.5.63) is determined from NTU computed from the known heat duty or ε and *C**; *G* in Equation (4.5.65) represents two equations, for Fluids 1 and 2 (Shah, 1988a); and the volume of the exchanger in Equation (4.5.67) is the same based on the surface area density of Fluid 1 or Fluid 2.

$$\frac{1}{UA} \approx \frac{1}{(\eta_o hA)_h} + \frac{1}{(\eta_o hA)_c} \tag{4.5.63}$$

Here we have neglected the wall and fouling thermal resistances. This equation in nondimensional form is given by

$$\frac{1}{\text{NTU}} = \frac{1}{\text{ntu}_h (C_h / C_{\min})} + \frac{1}{\text{ntu}_c (C_c / C_{\min})} \tag{4.5.64}$$

$$G_i = \left[\frac{2 g_c \Delta p}{\text{Deno}} \right]_i^{1/2} \qquad i = 1,2 \tag{4.5.65}$$

where

$$\text{Deno}_i = \left[\frac{f}{j} \frac{\text{ntu}}{\eta_o} \text{Pr}^{2/3} \left(\frac{1}{\rho} \right)_m + 2 \left(\frac{1}{\rho_o} - \frac{1}{\rho_i} \right) + (1 - \sigma^2 + K_c) \frac{1}{\rho_i} - (1 - \sigma^2 - K_e) \frac{1}{\rho_o} \right]_i \tag{4.5.66}$$

$$V = \frac{A_1}{\alpha_1} = \frac{A_2}{\alpha_2} \tag{4.5.67}$$

In the iterative solutions, the first time one needs ntu_h and ntu_c to start the iterations. These can be either determined from the past experience or by estimations. If both fluids are gases or both fluids are liquid, one could consider that the design is "balanced," i.e., that the thermal resistances are distributed approximately equally on the hot and cold sides. In that case, $C_h = C_c$, and

$$\text{ntu}_h \approx \text{ntu}_c \approx 2\text{NTU} \tag{4.5.68}$$

Alternatively, if we have liquid on one side and gas on the other side, consider 10% thermal resistance on the liquid side, i.e.,

$$0.10 \left(\frac{1}{UA} \right) = \frac{1}{(\eta_o hA)_{\text{liq}}} \tag{4.5.69}$$

Then, from Equations (4.5.63) and (4.5.64) with $C_{\text{gas}} = C_{\min}$, we can determine the ntu values on each side as follows:

$$\text{ntu}_{\text{gas}} = 1.11\text{NTU}, \qquad \text{ntu}_{\text{liq}} = 10 C^* \text{NTU} \tag{4.5.70}$$

Also note that initial guesses of η_o and *j/f* are needed for the first iteration to solve Equation (4.5.66). For a good design, consider $\eta_o = 0.80$ and determine an approximate value of *j/f* from the plot of *j/f* vs.

Re curve for the known j and f vs. Re characteristics of each fluid side surface. The specific step-by-step design procedure is as follows:

1. In order to compute the fluid bulk mean temperature and the fluid thermophysical properties on each fluid side, determine the fluid outlet temperatures from the specified heat duty

$$q = \left(\dot{m}c_p\right)_h \left(T_{h,i} - T_{h,o}\right) = \left(\dot{m}c_p\right)_c \left(T_{c,o} - T_{c,i}\right) \tag{4.5.71}$$

   or from the specified exchanger effectiveness using Equation (4.5.52) and (4.5.53). For the first time, estimate the values of c_p.

   For exchangers with $C^* \geq 0.5$, the bulk mean temperature on each fluid side will be the arithmetic mean of inlet and outlet temperatures on each side. For exchangers with $C^* < 0.5$, the bulk mean temperature on the C_{max} side will be the arithmetic mean of the inlet and outlet temperatures on that side and the bulk mean temperature on the C_{min} side will be the log-mean average as given by Equation (4.5.54). With these bulk mean temperatures, determine c_p and iterate one more time for the outlet temperatures if warranted. Subsequently, determine μ, c_p, k, Pr, and ρ on each fluid side.
2. Calculate C^* and ε (if q is given), and determine NTU from the ε-NTU expression, tables, or graphical results for the selected flow arrangement (in this case, it is unmixed–unmixed cross-flow, Table 4.5.4). The influence of longitudinal heat conduction, if any, is ignored in the first iteration since we don't know the exchanger size yet.
3. Determine ntu on each side by the approximations discussed with Equations (4.5.68) and (4.5.70) unless it can be estimated from past experience.
4. For the selected surfaces on each fluid side, plot j/f vs. Re curve from the given surface characteristics and obtain an approximate value of j/f. If fins are employed, assume $\eta_o = 0.80$ unless a better value can be estimated.
5. Evaluate G from Equation (4.5.65) on each fluid side using the information from Steps 1 to 4 and the input value of Δp.
6. Calculate Reynolds number Re, and determine j and f on each fluid side from the given design data for each surface.
7. Compute h, η_f, and η_o using Equations (4.5.57) to (4.5.59). For the first iteration, determine U_1 on Fluid 1 side from the following equation derived from Equations (4.5.6) and (4.5.67).

$$\frac{1}{U_1} = \frac{1}{\left(\eta_o h\right)_1} + \frac{1}{\left(\eta_o h_s\right)_1} + \frac{\alpha_1/\alpha_2}{\left(\eta_o h_s\right)_2} + \frac{\alpha_1/\alpha_2}{\left(\eta_o h\right)_2} \tag{4.5.72}$$

   where $\alpha_1/\alpha_2 = A_1/A_2$, $\alpha = A/V$, V is the exchanger total volume, and subscripts 1 and 2 denote Fluid 1 and 2 sides. For a plate-fin exchanger, α terms are given by Shah (1981) and Kays and London (1984):

$$\alpha_1 = \frac{b_1\beta_1}{b_1 + b_2 + 2a} \qquad \alpha_2 = \frac{b_2\beta_2}{b_1 + b_2 + 2a} \tag{4.5.73}$$

   Note that the wall thermal resistance in Equation (4.5.72) is ignored in the first iteration. In second and subsequent iterations, compute U_1 from

$$\frac{1}{U_1} = \frac{1}{\left(\eta_o h\right)_1} + \frac{1}{\left(\eta_o h_s\right)_1} + \frac{\delta A_1}{k_w A_w} + \frac{A_1/A_2}{\left(\eta_o h_s\right)_2} + \frac{A_1/A_2}{\left(\eta_o h\right)_2} \tag{4.5.74}$$

where the necessary geometry information A_1/A_2 and A_1/A_w is determined from the geometry calculated in the previous iteration.

8. Now calculate the core dimensions. In the first iteration, use NTU computed in Step 2. For subsequent iterations, calculate longitudinal conduction parameter λ (and other dimensionless groups for a crossflow exchanger). With known ε, C^*, and λ, determine the correct value of NTU using either a closed-form equation or tabular/graphical results (Kays and London, 1984). Determine A_1 from NTU using U_1 from the previous step and known $C_{\min}$.

$$A_1 = \text{NTU}\, C_{\min}/U_1 \tag{4.5.75}$$

and hence

$$A_2 = \left(A_2/A_1\right)A_1 = \left(\alpha_2/\alpha_1\right)A_1 \tag{4.5.76}$$

A_o from known $\dot{m}$ and G is given by

$$A_{o,1} = \left(\dot{m}/G\right)_1 \qquad A_{o,2} = \left(\dot{m}/G\right)_2 \tag{4.5.77}$$

so that

$$A_{fr,1} = A_{o,1}/\sigma_1 \qquad A_{fr,2} = A_{o,2}/\sigma_2 \tag{4.5.78}$$

where σ_1 and σ_2 are generally specified for the surface or can be computed for plate-fin surfaces from Shah (1981) and Kays and London (1984):

$$\sigma_1 = \frac{b_1\beta_1 D_{h,1}/4}{b_1 + b_2 + 2\delta} \qquad \sigma_2 = \frac{b_2\beta_2 D_{h,2}/4}{b_1 + b_2 + 2\delta} \tag{4.5.79}$$

Now compute the fluid flow lengths on each side (see Figure 4.5.17) from the definition of the hydraulic diameter of the surface employed on each side.

$$L_1 = \left(\frac{D_h A}{4A_o}\right)_1 \qquad L_2 = \left(\frac{D_h A}{4A_o}\right)_2 \tag{4.5.80}$$

Since $A_{fr,1} = L_2L_3$ and $A_{fr,2} = L_1L_3$, we can obtain

$$L_3 = \frac{A_{fr,1}}{L_2} \quad \text{or} \quad L_3 = \frac{A_{fr,2}}{L_1} \tag{4.5.81}$$

Theoretically, L_3 calculated from both expressions of Equation (4.5.81) should be identical. In reality, they may differ slightly because of the round-off error. In that case, consider an average value for L_3.

9. Finally, compute the pressure drop on each fluid side, after correcting f factors for variable property effects, in a manner similar to Step 8 of the rating problem for a Crossflow Plate Fin Exchanger.
10. If the calculated values of Δp are within and close to input specifications, the solution to the sizing problem is completed. Finer refinements in the core dimensions, such as integer numbers of flow passages, etc., may be carried out at this time. Otherwise, compute the new value of G on each fluid side using Equation (4.5.29) in which Δp is the input-specified value, and f, K_c, K_e, and geometric dimensions are from the previous iteration.

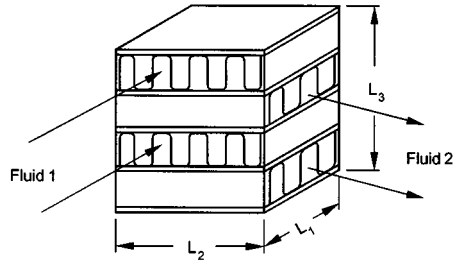


FIGURE 4.5.17 A single-pass crossflow exchanger.

11. Repeat (iterate) Steps 6 to 10 until both transfer and pressure drops are met as specified. It should be emphasized that since we have imposed no constraints on the exchanger dimensions, the above procedure will yield L_1, L_2, and L_3 for the selected surfaces such that the design will meet exactly the heat duty and pressure drops on both fluid sides.

Flow Maldistribution

In the previously presented heat transfer (ε-NTU, MTD, etc. methods) and pressure drop analyses, it is presumed that the fluid is uniformly distributed through the core. In practice, flow maldistribution does occur to some extent and often severely, and may result in a significant reduction in exchanger heat transfer performance and an increase in the pressure drop. Hence, it may be necessary for the designer to take into account the effect of flow maldistribution causing undesirable performance deterioration up front while designing a heat exchanger.

Some maldistributions are geometry-induced (i.e., the result of exchanger fabrication conditions, such as header design or manufacturing tolerances, or the duct geometry/structure upstream of the exchanger), and other maldistributions are the result of exchanger operating conditions. Gross, passage-to-passage and manifold-induced flow maldistributions are examples of the former category, while viscosity, natural convection, and density-difference-induced flow maldistributions are of the latter category. Flow maldistributions associated with two-phase and multiphase flow are too complex, with only limited information available in the literature. The analysis methods and results for some of the above flow maldistributions for single-phase flows are given by Shah (1985), Mueller and Chiou (1987), and Putnam and Rohsenow (1985).

Fouling in Heat Exchangers

Fouling, Its Effect, and Mechanisms.

Fouling refers to undesired accumulation of solid material (by-products of the heat transfer processes) on heat exchanger surfaces which results in additional thermal resistance to heat transfer, thus reducing exchanger performance. The fouling layer also blocks the flow passage/area and increases surface roughness, thus either reducing the flow rate in the exchanger or increasing the pressure drop or both. The foulant deposits may be loose such as magnetite particles or hard and tenacious such as calcium carbonate scale; other deposits may be sediment, polymers, coking or corrosion products, inorganic salts, biological growth, etc. Depending upon the fluids, operating conditions, and heat exchanger construction, the maximum fouling layer thickness on the heat transfer surface may result in a few hours to a number of years.

Fouling could be very costly depending upon the nature of fouling and the applications. It increases capital costs: (1) oversurfacing heat exchanger, (2) provisions for cleaning, and (3) use of special materials and constructions/surface features. It increases maintenance costs: (1) cleaning techniques, (2) chemical additives, and (3) troubleshooting. It may cause a loss of production: (1) reduced capacity and (2) shutdown. It increases energy losses: (1) reduced heat transfer, (2) increased pressure drop, and (3) dumping dirty streams. Fouling promotes corrosion, severe plugging, and eventual failure of uncleaned

heat exchangers. In a fossil-fired exhaust environment, gas-side fouling produces a potential fire hazard in heat exchangers.

The following are the major fouling mechanisms:

- *Crystallization or precipitation fouling* results from the deposition/formation of crystals of dissolved substances from the liquid onto heat transfer surface due to solubility changes with temperature beyond the saturation point. If the deposited layer is hard and tenacious, it is often referred to as scaling. If it is porous and mushy, it is called sludge.
- *Particulate fouling* results from the accumulation of finely divided substances suspended in the fluid stream onto heat transfer surface. If the settling occurs as a result of gravity, it is referred to as sedimentation fouling.
- *Chemical reaction fouling* is defined as the deposition of material produced by chemical reaction (between reactants contained in the fluid stream) in which the heat transfer surface material does not participate.
- *Corrosion fouling* results from corrosion of the heat transfer surface that produces products fouling the surface and/or roughens the surface, promoting attachment of other foulants.
- *Biological fouling* results from the deposition, attachment, and growth of biological organisms from liquid onto a heat transfer surface. Fouling due to microorganisms refers to microbial fouling and fouling due to macroorganisms refers to macrobial fouling.
- *Freezing fouling* results from the freezing of a single-component liquid or higher-melting-point constituents of a multicomponent liquid onto a subcooled heat transfer surface.

Biological fouling occurs only with liquids since there are no nutrients in gases. Also crystallization fouling is not too common with gases since most gases contain few dissolved salts (mainly in mists) and even fewer inverse-solubility salts. All other types of fouling occur in both liquid and gas. More than one mechanism is usually present in many fouling situations, often with synergetic results. Liquid-side fouling generally occurs on the exchanger side where the liquid is being heated, and gas-side fouling occurs where the gas is being cooled; however, reverse examples can be found.

Importance of Fouling.

Fouling in liquids and two-phase flows has a significant detrimental effect on heat transfer with some increase in pressure drop. In contrast, fouling in gases reduces heat transfer somewhat (5 to 10% in general) in compact heat exchangers, but increases pressure drop significantly up to several hundred percent. For example, consider U = 1400 W/m^{2}K as in a process plant liquid-to-liquid heat exchanger. Hence, $R = 1/U = 0.00072$ m^{2}K/W. If the fouling factors ($r_{s,h} + r_{s,c}$) together amount to 0.00036 (considering a typical TEMA value of the fouling factor as 0.00018), 50% of the heat transfer area requirement A for given q is chargeable to fouling. However, for gas flows on both sides of an exchanger, $U \approx 280$ W/m^{2}K, and the same fouling factor of 0.00036 would represent only about 10% of the total surface area. Thus, one can see a significant impact on the heat transfer surface area requirement due to fouling in heat exchangers having high U values (such as having liquids or phase-change flows).

Considering the core frictional pressure drop (Equation (4.5.32) as the main pressure drop component, the ratio of pressure drops of fouled and cleaned exchangers is given by

$$\frac{\Delta p_F}{\Delta p_C} = \frac{f_F}{f_C}\left(\frac{D_{h,C}}{D_{h,F}}\right)\left(\frac{u_{m,F}}{u_{m,C}}\right)^2 = \frac{f_F}{f_C}\left(\frac{D_{h,C}}{D_{h,F}}\right)^5 \tag{4.5.82}$$

where the term after the second equality sign is for a circular tube and the mass flow rates under fouled and clean conditions remain the same. Generally, $f_F > f_C$ due to the fouled surface being rough. Thus, although the effect of fouling on the pressure drop is usually neglected, it can be significant, particularly for compact heat exchangers with gas flows. If we consider $f_F = f_C$, and the reduction in the tube inside

diameter due to fouling by only 10 and 20%, the resultant pressure drop increase will be 69 and 205%, respectively, according to Equation (4.5.82) regardless of whether the fluid is liquid or gas!

Accounting of Fouling in Heat Exchangers.

Fouling is an extremely complex phenomenon characterized by a combined heat, mass, and momentum transfer under transient condition. Fouling is affected by a large number of variables related to heat exchanger surfaces, operating conditions, and fluids. Fouling is time dependent, zero at $\tau = 0$; after the induction or delay period τ_d, the fouling resistance is either pseudolinear, falling rate, or asymptotic.

Fouling is characterized by all or some of the following sequential events: initiation, transport, attachment, removal, and aging (Epstein, 1983). Research efforts are concentrated on quantifying these events by semitheoretical models (Epstein, 1978) with very limited success on specific fouling situations. Hence, the current heat exchanger design approach is to use a constant (supposedly an asymptotic) value of the fouling factor $r_s = 1/h_s$. Equation (4.5.6) presented earlier includes the fouling resistances on the hot and cold sides for a nontubular extended-surface exchanger. Here $1/h_s = r_s$ is generally referred to as the *fouling factor.* Fouling factors for some common fluids are presented in Tables 4.5.8 and 4.5.9.

The specification of fouling effects in a process heat exchanger is usually represented in the following form, wherein the combined fouling factor $r_{s,t}$ is the sum of the fouling factors on the hot and cold sides:

$$\text{Combined fouling factor} \qquad r_{s,t} = \frac{1}{U_C} - \frac{1}{U_F} \tag{4.5.83}$$

$$\text{Cleanliness factor} \qquad \text{CF} = U_F / U_C \tag{4.5.84}$$

$$\text{Percentage oversurface} \qquad \%\text{OS} = \left(\frac{A_F}{A_C} - 1 \right) 100 \tag{4.5.85}$$

Here the subscripts F and C denote fouled and clean exchanger values. From Equation (4.5.6) with $A_h = A_c = A$, $\eta_o = 1$, $\Delta T_{m,F} = \Delta T_{m,C}$, it can be shown that

$$\frac{A_F}{A_C} = \frac{U_C}{U_F} = 1 + U_C r_{s,t} \tag{4.5.86}$$

where $r_{s,t} = r_{s,h} + r_{s,c}$. In heat exchanger design, constant (supposedly an asymptotoic) values of $r_{s,h}$ and $r_{s,c}$ are used. Accordingly, extra heat transfer surface area is provided to take into account the deleterious effect of fouling. Thus, the heat exchanger will be "oversized" for the initial clean condition, "correctly sized" for asymptotic fouling (if it occurs in practice), and "undersized" just before the cleaning operation for nonasymptotic fouling.

Influence of Operating and Design Variables.

Based on operational experience and research over the last several decades, many variables have been identified that have a significant influence on fouling. The most important variables are summarized next.

Flow velocity. Flow velocity is one of the most important variables affecting fouling. Higher velocities increase fluid shear stress at the fouling deposit–fluid interface and increase the heat transfer coefficient; but, at the same time, increased pressure drop and fluid pumping power may erode the surface and may accelerate the corrosion of the surface by removing the protective oxide layer. The fouling buildup in general is inversely proportional to $u_m^{1.5}$. For water, the velocity should be kept above 2 m/sec to suppress fouling, and the absolute minimum should be above 1 m/sec to minimize fouling.

Surface temperature. Higher surface temperatures promote chemical reaction, corrosion, crystal formation (with inverse solubility salts), and polymerization, but reduce biofouling for temperatures above the optimum growth, avoid potential freezing fouling, and avoid precipitation of normal-solubility salts.

TABLE 4.5.8 Fouling Factors for Various Fluid Streams Used in Heat Exchangers

Water Type	Fouling Factors ($m^2 \cdot K/W$)
Seawater (43°C maximum outlet)	0.000275 to 0.00035
Brackish water (43°C maximum outlet)	0.00035 to 0.00053
Treated cooling tower water (49°C maximum outlet)	0.000175 to 0.00035
Artificial spray pond (49°C maximum outlet)	0.000175 to 0.00035
Closed-loop treated water	0.000175
River water	0.00035 to 0.00053
Engine jacket water	0.000175
Distilled water or closed-cycle condensate	0.00009 to 0.000175
Treated boiler feedwater	0.00009
Boiler blowdown water	0.00035 to 0.00053
Liquids	
No. 2 fuel oil	0.00035
No. 6 fuel oil	0.0009
Transformer oil	0.000175
Engine lube oil	0.000175
Refrigerants	0.000175
Hydraulic fluid	0.000175
Industrial organic HT fluids	0.000175 to 0.00035
Ammonia	0.000175
Ammonia (oil bearing)	0.00053
Methanol solutions	0.00035
Ethanol solutions	0.00035
Ethylene glycol solutions	0.00035
MEA and DEA solutions	0.00035
DEG and TEG solutions	0.00035
Stable side draw and bottom products	0.000175 to 0.00035
Caustic solutions	0.00035
Gas or Vapor	
Steam (non-oil-bearint)	0.0009
Exhaust steam (oil-bearing)	0.00026 to 0.00035
Refrigerant (oil-bearing)	0.00035
Compressed air	0.000175
Ammonia	0.000175
Carbon dioxide	0.00035
Coal flue gas	0.00175
Natural gas flue gas	0.00090
Acid gas	0.00035 to 0.00053
Solvent vapor	0.000175
Stable overhead products	0.000175
Natural Gas and Petroleum Streams	
Natural gas	0.000175 to 0.00035
Overhead products	0.000175 to 0.00035
Lean oil	0.00035
Rich oil	0.000175 to 0.00035
Natural gasoline and liquefied petroleum gases	0.000175 to 0.00035
Oil Refinery Streams	
Crude and vacuum unit gases and vapors	
Atmospheric tower overhead vapors	0.00017
Light naphthas	0.00017
Vacuum overhead vapors	0.00035

TABLE 4.5.8 (continued) Fouling Factors for Various Fluid Streams Used in Heat Exchangers

Oil Refinery Streams	
Crude and vacuum liquids	
Gasoline	0.00035
Naphtha and light distillates	0.00035 to 0.00053
Kerosene	0.00035 to 0.00053
Light gas oil	0.00035 to 0.00053
Heavy gas oil	0.00053 to 0.0009
Heavy fuel oil	0.00053 to 0.00123
Vacuum tower bottoms	0.00176
Atmospheric tower bottoms	0.00123
Cracking and coking unit streams	
Overhead vapors	0.00035
Light cycle oil	0.00035 to 0.00053
Heavy cycle oil	0.00053 to 0.0007
Light coker gas oil	0.00053 to 0.0007
Heavy coker gas oil	0.00070 to 0.0009
Bottoms slurry oil (1.5 m/sec minimum)	0.00053
Light liquid products	0.00035
Catalytic reforming, hydrocracking, and hydrodesulfurization streams	
Reformer charge	0.00026
Reformer effluent	0.00026
Hydrocharger charge and effluent[a]	0.00035
Recycle gas	0.000175
Liquid product over 50°C (API)[b]	0.000175
Liquid product 30 to 50°C (API)[b]	0.00035
Light ends processing streams	
Overhead vapors and gases	0.000175
Liquid products	0.000175
Absorption oils	0.00035 to 0.00053
Alkylation trace acid streams	0.00035
Reboiler streams	0.00035 to 0.00053

[a] Depending on charge characteristics and storage history, charge fouling resistance may be many times this value.

[b] American Petroleum Institute.

Source: Chenoweth, J., Final Report, HTRI/TEMA Joint Committee to Review the Fouling Section of TEMA Standards, HTRI, Alhambra, CA, 1988. With permission.

It is highly recommended that the surface temperature be maintained below the reaction temperature; it should be kept below 60°C for cooling tower water.

Tube material. The selection of the tube material is important from the corrosion point of view which in turn could increase crystallization and biological fouling. Copper alloys can reduce certain biofouling, but their use is limited by environmental concerns with river, ocean, and lake waters.

There are many other variables that affect fouling. It is beyond the scope here, but the reader may refer to TEMA (1988).

Fouling Control and Cleaning Techniques.

Control of fouling should be attempted first before any cleaning method is attempted. For gas-side fouling, one should verify that fouling exists, identify the sequential event that dominates the foulant accumulation, and characterize the deposit. For liquid-side fouling, fouling inhibitors/additives should be employed while the exchanger is in operation; for example, use antidispersant polymers to prevent sedimentation fouling, "stabilizing" compounds to prevent polymerization and chemical reaction fouling, corrosion inhibitors to prevent corrosion fouling, biocide/germicide to prevent biofouling, softeners, acids, and polyphosphates to prevent crystallization fouling.

TABLE 4.5.9 Fouling Factors and Design Parameters for Finned Tubes in Fossil Fuel Exhaust Gases

Type of Flue Gas	Fouling Factor, m^2K/W	Minimum Spacing between Fins, m	Maximum Gas Velocity to Avoid Erosion, m/sec
Clean Gas (Cleaning Devices Not Required)			
Natural Gas	0.0000881–0.000528	0.00127–0.003	30.5–36.6
Propane	0.000176–0.000528	0.00178	—
Butane	0.000176–0.000528	0.00178	—
Gas turbine	0.000176	—	—
Average Gas (Provisions for Future Installation of Cleaning Devices)			
No. 2 oil	0.000352–0.000704	0.00305–0.00384	25.9–30.5
Gas turbine	0.000264	—	—
Diesel engine	0.000528	—	—
Dirty Gas (Cleaning Devices Required)			
No. 6 oil	0.000528–0.00123	0.00457–0.00579	18.3–24.4
Crude oil	0.000704–0.00264	0.00508	—
Residual oil	0.000881–0.00352	0.00508	—
Coal	0.000881–0.00881	0.00587–0.00864	15.2–21.3

Source: Weierman, R.C., 1982. Design of Heat Transfer Equipment for Gas-Side Fouling Service, Workshop on an Assessment of Gas-Side Fouling in Fossil Fuel Exhaust Environments, W.J. Marner and R.L. Webb, Eds., JPL Publ. 82-67, Jet Propulsion Laboratory, California Institute of Technology, Pasadena. With permission.

If the foulant control is not effective, the exchanger must be cleaned either on-line or off-line. On-line cleaning includes flow-driven brushes/sponge balls inside tubes, power-driven rotating brushes inside tubes, acoustic horns/mechanical vibrations for tube banks with gases, soot blowers, and shutting off of the cold gas supply, flowing hot gas, or reversing of the fluids. Off-line cleaning methods, without dismantling the exchanger include chemical cleaning (circulate acid/detergent solutions), circulating of particulate slurry (such as sand and water), and thermal melting of frost layers. Off-line cleaning with a heat exchanger opened includes high-pressure steam or water cleaning, and thermal baking of an exchanger and then rinsing for small heat exchanger modules removed from the container of the modular exchangers.

Nomenclature

A total heat transfer area (primary + fin) on one fluid side of a heat exchanger, A_p: primary surface area, A_f: fin surface area, m^2

A_{fr} frontal area on one side of an exchanger, m^2

A_k total wall cross-sectional area for heat conduction in fin or for longitudinal conduction in the exchanger, m^2

A_o minimum free-flow area on one fluid side of a heat exchanger, m^2

b plate spacing, $h' + \delta_f$, m

C flow stream heat capacity rate with a subscript c or h, $\dot{m}c_p$, W/°C

C^* heat capacity rate ratio, C_{min}/C_{max}, dimensionless

c_p specific heat of fluid at constant pressure, J/kg K

D_h hydraulic diameter of flow passages, $4A_oL/A$, m

d_e fin tip diameter of an individually finned tube, m

d_i, d_o tube inside and outside diameters, respectively, m

Eu N-row average Euler number, $\Delta p/(\rho\, u_m^2\, N/2g_c)$, $\rho\Delta p g_c/(NG^2/2)$, dimensionless

F log-mean temperature difference correction factor, dimensionless
f Fanning friction factor, $\rho \Delta p g_c D_h/(2LG^2)$, dimensionless
f_{tb} Fanning friction factor per tube row for crossflow over a tube bank outside, $\rho \Delta p g_c/(2NG^2)$
G mass velocity based on the minimum free flow area, $\dot{m}/A_o$, kg/m^{2}sec
g gravitational acceleration, m^2/sec
g_c proportionality constant in Newton's second law of motion, $g_c = 1$ and dimensionless in SI units
H fin length for heat conduction from primary surface to either fin tip or midpoint between plates for symmetric heating, m
h heat transfer coefficient, W/m^{2}K
h' height of the offset strip fin (see Figure 4.5.15), m
j Colburn factor, $NuPr^{-1/3}/Re$, $StPr^{2/3}$, dimensionless
k fluid thermal conductivity, W/m K
k_f thermal conductivity of the fin material, W/m K
k_w thermal conductivity of the matrix (wall) material, W/m K
L fluid flow (core or tube) length on one side of an exchanger, m
l fin length for heat conduction from primary surface to the midpoint between plates for symmetric heating, see Table 4.5.5 for other definitions of l, m
l_f offset trip fin length or fin height for individually finned tubes, l_f represents the fin length in the fluid flow direction for an uninterrupted fin with $l_f = L$ in most cases, m
m fin parameter, 1/m
N number of tube rows
N_f number of fins per meter, 1/m
N_t total number of tubes in an exchanger
NTU number of heat transfer units, $UA/C_{\min}$, it represents the total number of transfer units in a multipass unit, $NTU_s = UA/C_{shell}$, dimensionless
Nu Nusselt number, hD_h/k, dimensionless
ntu_c number of heat transfer units based on the cold side, $(\eta_o hA)_c/C_c$, dimensionless
ntu_h number of heat transfer units based on the hot side, $(\eta_o hA)_h/C_h$, dimensionless
$\dot{m}$ mass flow rate, kg/sec
P temperature effectiveness of one fluid, dimensionless
$\wp$ fluid pumping power, W
Pr fluid Prandtl number, $\mu c_p/k$, dimensionless
p fluid static pressure, Pa
Δp fluid static pressure drop on one side of heat exchanger core, Pa
p_f fin pitch, m
q heat duty, W
q_e heat transfer rate (leakage) at the fin tip, W
q'' heat flux, W/m^2
R heat capacity rate ratio used in the P-NTU method, $R_1 = C_1/C$, $R_2 = C_2/C_1$, dimensionless
R thermal resistance based on the surface area A, compare Equations 4.5.5 and 4.5.6 for definitions of specific thermal resistances, K/W
Re Reynolds number, GD_h/μ, dimensionless
Re_d Reynolds number, $\rho u_m d_o/\mu$, dimensionless
r_h hydraulic radius, $D_h/4$, $A_o L/A$, m
r_s fouling factor, $1/h_s$, m^{2}K/W
St Stanton number, h/Gc_p, dimensionless
s distance between adjacent fins, $p_f - \delta_f$ m
T fluid static temperature to a specified arbitrary datum, °C
T_a ambient temperature, °C
T_o fin base temperature, °C
T_l fin tip temperature, °C
U overall heat transfer coefficient, W/m^2 K
u_m mean axial velocity in the minimum free flow area, m/sec

V heat exchanger total volume, m^3
X_d diagonal tube pitch, m
X_l longitudinal tube pitch, m
X_t transverse tube pitch, m
α ratio of total heat transfer area on one side of an exchanger to the total volume of an exchanger, A/V, m^2/m^3
β heat transfer surface area density, a ratio of total transfer area on one side of a plate-fin heat exchanger to the volume between the plates on that side, m^2/m^3
ε heat exchanger effectiveness, it represents an overall exchanger effectiveness for a multiplass unit, dimensionless
δ wall thickness, m
δ_f fin thickness, m
η_f fin efficiency, dimensionless
η_o extended surface efficiency, dimensionless
λ longitudinal wall heat conduction parameter based on the total conduction area, $\lambda = k_w A_{k,t}/C_{\min}L$, $\lambda_c = k_w A_{k,c}/C_c L$, $\lambda_h = k_w A_{k,h}/C_h L$, dimensionless
μ fluid dynamic viscosity, Pa·s
ρ fluid density, kg/m^3
σ ratio of free flow area to frontal area, A_o/A_{fr}, dimensionless

Subscripts

C clean surface value
c cold fluid side
F fouled surface value
f fin
h hot fluid side
i inlet to the exchanger
o outlet to the exchanger
s scale or fouling
w wall or properties at the wall temperature
1 one section (inlet or outlet) of the exchanger
2 other section (outlet or inlet) of the exchanger

References

Bhatti, M.S. and Shah, R.K. 1987. Turbulent and transition flow convective heat transfer in ducts, in *Handbook of Single-Phase Convective Heat Transfer,* S. Kakaç, R. K. Shah, and W. Aung, Eds., John Wiley & Sons, New York, chap. 4, 166 pp.

Chai, H.C. 1988. A simple pressure drop correlation equation for low finned tube crossflow heat exchangers, *Int. Commun. Heat Mass Transfer*, 15, 95–101.

Chenoweth, J. 1988. Final Report, HTRI/TEMA Joint Committee to Review the Fouling Section of TEMA Standards, HTRI, Alhambra, CA.

Cowell, T.A., Heikal, M.R., and Achaichia, A. 1995. Flow and heat transfer in compact louvered fin surfaces, *Exp. Thermal Fluid Sci.*,10, 192–199.

Epstein, N. 1978. Fouling in heat exchangers, in *Heat Transfer 1978*, Vol. 6, Hemisphere Publishing, New York, 235–254.

Epstein, N. 1983. Thinking about heat transfer fouling: a 5×5 matrix, *Heat Transfer Eng.*, 4(1), 43–56.

Foumeny, E.A. and Heggs, P.J. 1991. *Heat Exchange Engineering*, Vol. 2, *Compact Heat Exchangers: Techniques for Size Reduction*, Ellis Horwood Ltd., London.

Ganguli, A. and Yilmaz, S.B. 1987. New heat transfer and pressure drop correlations for crossflow over low-finned tube banks, *AIChE Symp. Ser. 257*, 83, 9–14.

Ghajar, A.J. and Tam, L.M. 1994. Heat transfer measurements and correlations in the transition region for a circular tube with three different inlet configurations, *Exp. Thermal Fluid Sci.*, 8, 79–90.

Huang, L.J. and Shah, R.K. 1992. Assessment of calculation methods for efficiency of straight fins of rectangular profiles, *Int. J. Heat Fluid Flow*, 13, 282–293.

Idelchik, I.E. 1994. *Handbook of Hydraulics Resistance*, 3rd ed., CRC Press, Boca Raton, FL.

Kakaç, S., Ed. 1991. *Boilers, Evaporators, and Condensers*, John Wiley & Sons, New York.

Kakaç, S., Bergles, A.E., and Mayinger, F. 1981. *Heat Exchangers: Thermal-Hydraulic Fundamentals and Design*, Hemisphere Publishing, Washington, D.C.

Kakaç, S., Shah, R.K., and Bergles, A.E. 1983. *Low Reynolds Number Flow Heat Exchangers*, Hemisphere Publishing, Washington, D.C.

Kakaç, S., Bergles, A.E., and Fernandes, E.O. 1988. *Two-Phase Flow Heat Exchangers: Thermal Hydraulic Fundamentals and Design*, Kluwer Academic Publishers, Dordrecht, Netherlands.

Kays, W.M. and London, A.L. 1984. *Compact Heat Exchangers*, 3rd ed., McGraw-Hill, New York.

Manglik, R.M. and Bergles, A.E. 1995. Heat transfer and pressure drop correlations for the rectangular offset-strip-fin compact heat exchanger, *Exp. Thermal Fluid Sci.*, 10, 171–180.

Miller, D.S. 1990. *Internal Flow Systems, 2nd ed., BHRA (Information Services)*, Cranfield, Bedford, U.K.

Mueller, A.C. and Chiou, J.P. 1987. *Review of Various Types of Flow Maldistribution in Heat Exchangers*, Book No. H00394, HTD-Vol. 75, ASME, New York, 3–16.

Putnam, G.R. and Rohsenow, W.M. 1985. Viscosity induced nonuniform flow in laminar flow heat exchangers, *Int. J. Heat Mass Transfer*, 28, 1031–1038.

Rabas, T.J. and Taborek, J. 1987. Survey of turbulent forced-convection heat transfer and pressure drop characteristics of low-finned tube banks in cross flow, *Heat Transfer Eng.*, 8(2), 49–62.

Roetzel, W., Heggs, P.J., and Butterworth, D., Eds. 1991. *Design and Operation of Heat Exchangers*, Springer-Verlag, Berlin.

Rozenman, T. 1976. Heat transfer and pressure drop characteristics of dry cooling tower extended surfaces, Part I: Heat transfer and pressure drop data, Report BNWL-PFR 7-100; Part II: Data analysis and correlation, Report BNWL-PFR 7-102, Battelle Pacific Northwest Laboratories, Richland, WA.

Shah, R.K. 1981. Compact heat exchangers, in *Heat Exchangers: Thermal-Hydraulic Fundamentals and Design*, S. Kakaç, A.E. Bergles, and F. Mayinger, Eds., Hemisphere Publishing, Washington, D.C., 111–151.

Shah, R.K. 1983. Heat Exchanger Basic Design Methods, in *Low Reynolds Number Flow Heat Exchanger*, S. Kakaç, R.K. Shah and A.E. Bergles, Eds., pp. 21–72, Hemisphere, Washington, D.C.

Shah, R.K. 1985. Compact heat exchangers, in *Handbook of Heat Transfer Applications*, 2nd ed., W.M. Rohsenow, J.P. Hartnett, and E.N. Ganic, Eds., McGraw-Hill, New York, Chap. 4, Part 3.

Shah, R.K. 1988a. Plate-fin and tube-fin heat exchanger design procedures, in *Heat Transfer Equipment Design*, R.K. Shah, E.C. Subbarao, and R.A. Mashelkar, Eds., Hemisphere Publishing, Washington, D.C., 255–266.

Shah, R.K. 1988b. Counterflow rotary regenerator thermal design procedures, in *Heat Transfer Equipment Design*, R.K. Shah, E.C. Subbarao, and R.A. Mashelkar, Eds., Hemisphere Publishing, Washington, D.C., 267–296.

Shah, R.K. 1991. Multidisciplinary approach to heat exchanger design, in *Industrial Heat Exchangers*, J.-M. Buchlin, Ed., Lecture Series No. 1991-04, von Kármán Institute for Fluid Dynamics, Rhode Saint Genèse, Belgium.

Shah, R.K. 1993. Nonuniform heat transfer coefficients for heat exchanger thermal design, in *Aerospace Heat Exchanger Technology 1993*, R.K. Shah and A. Hashemi, Eds., Elsevier Science, Amsterdam, Netherlands, 417–445.

Shah, R.K. 1994. Heat exchangers, in *Encyclopedia of Energy Technology and The Environment*, A. Bision and S.G. Boots, Eds., John Wiley & Sons, New York, 1651–1670.

Shah, R.K., Bell, K.J., Mochizuki, S., and Wadekar, V. V., Eds., 1997. *Compact Heat Exchangers for the Process Industries*, Begell House, New York.

Shah, R.K. and Bhatti, M.S. 1987. Laminar convective heat transfer in ducts, in *Handbook of Single-Phase Convective Heat Transfer*, S. Kacaç, R.K. Shah, and W. Aung, Eds., John Wiley, New York, Chap. 3, 137 pp.

Shah, R.K. and Bhatti, M.S. 1988. Assessment of correlations for single-phase heat exchangers, in *Two-Phase Flow Heat Exchangers: Thermal-Hydraulic Fundamentals and Design*, S. Kakaç, A.E. Bergles, and E.O. Fernandes, Eds., Kluwer Academic Publishers, Dordrecht, The Netherlands, 81–122.

Shah, R.K. and Giovannelli, A.D. 1988. Heat pipe heat exchanger design theory, in *Heat Transfer Equipment Design*, R.K. Shah, E.C. Subbarao, and R.A. Mashelkar, Eds., Hemisphere Publishing, Washington, D.C., 609–653.

Shah, R.K. and Hashemi, A., Eds. 1993. *Aerospace Heat Exchanger Technology*, Elsevier Science, Amsterdam.

Shah, R.K., Kraus, A.D., and Metzger, D.E., Eds., 1990. *Compact Heat Exchangers — A Festschrift for Professor A.L. London*, Hemisphere, Washington, D.C.

Shah, R.K. and London, A.L. 1978. Laminar flow forced convection in ducts, Suppl. 1 to *Advances in Heat Transfer*, Academic Press, New York.

Shah, R.K. and Mueller, A.C. 1988. Heat Exchange, in *Ullmann's Encyclopedia of Industrial Chemistry*, Unit Operations II, vol. B3, chap. 2, 108 pages, VCH, Weinheim, Germany.

Shah, R.K. and Pignotti, A. 1997. The influence of a finite number of baffles on the shell-and-tube heat exchanger performance, *Heat Transfer Eng.*, 18.

Shah, R.K., Subbarao, E.C., and Mashelkar, R.A., Eds. 1988. *Heat Transfer Equipment Design*, Hemisphere Publishing, Washington, D.C.

Shah, R.K. and Wanniarachchi, A.S. 1991. Plate heat exchanger design theory, in *Industrial Heat Exchangers*, J.-M. Buchlin, Ed., Lecture Series No. 1991-04, von Kármán Institute for Fluid Dynamics, Rhode Saint Genèse, Belgium.

Taylor, M.A. 1987. *Plate-Fin Heat Exchangers: Guide to Their Specificaitons and Use*, 1st ed., HTFS, Harwell Laboratory, Oxon, U.K., rev. 1990.

TEMA, 1988. *Standards of the Tubular Exchanger Manufacturers Association*, 7th ed., Tubular Exchanger Manufacturers Association, New York.

Webb, R.L. 1994. *Principles of Enhanced Heat Transfer*, John Wiley & Sons, New York.

Weierman, R.C. 1982. Design of Heat Transfer Equipment for Gas-Side Fouling Service, Workshop on an Assessment of Gas-Side Fouling in Fossil Fuel Exhaust Environments, W.J. Marner and R.L. Webb, Eds., JPL Publ. 82-67, Jet Propulsion Laboratory, California Institute of Technology, Pasadena.

Zukauskas, A. 1987. Convective heat transfer in cross flow, in *Handbook of Single-Phase Convective Heat Transfer*, S. Kacaç, R.K. Shah, and W. Aung, John Wiley, New York, Chap. 6.

Further Information

Heat exchangers play a crucial and dominant role in many developments related to energy conservation, recovery, utilization, economic development of new energy sources, and environmental issues such as air and water pollution control, thermal pollution, waste disposal, etc. Many new and innovative heat exchangers have been developed for these and many other applications worldwide. A broad overview is provided for various heat exchangers and basic design theory for single-phase heat exchangers. For further details and study, the reader may refer to the following references: Kakaç et al. (1981; 1983; 1988), Taylor (1987), Shah et al. (1990), Foumeny and Heggs (1991), Kakaç (1991), Roetzel et al. (1991), Shah and Hashemi (1993), and Shah et al. (1997).

Shell-and-Tube Heat Exchangers

Kenneth J. Bell

Introduction

A shell-and-tube heat exchanger is essentially a bundle of tubes enclosed in a shell and so arranged that one fluid flows through the tubes and another fluid flows across the outside of the tubes, heat being transferred from one fluid to the other through the tube wall. A number of other mechanical components are required to guide the fluids into, through, and out of the exchanger, to prevent the fluids from mixing, and to ensure the mechanical integrity of the heat exchanger. A typical shell-and-tube heat exchanger is shown in Figure 4.5.18 (TEMA, 1988), but the basic design allows many modifications and special features, some of which are described below.

1. Stationary Head-Channel
2. Stationary Head Flange-Channel or Bonnet
3. Channel Cover
4. Stationary Head Nozzle
5. Stationary Tubesheet
6. Tubes
7. Shell
8. Shell Cover
9. Shell Flange-Stationary Head End
10. Shell Flange-Rear Head End
11. Shell Nozzle
12. Shell Cover Flange
13. Floating Tubesheet
14. Floating Head Cover
15. Floating Head Cover Flange
16. Floating Head Backing Device
17. Tierods and Spacers
18. Transverse Baffles or Support Plates
19. Impingement Plates
20. Pass Partition
21. Vent Connection
22. Drain Connection
23. Instrument Connection
24. Support Saddle
25. Lifting Lug

Nomenclature of Heat Exchanger Components

For the purpose of establishing standard terminology, Figure 4.5.18 illustrates various types of heat exchangers. Typical parts and connections, for illustrative purposes only, are numbered for identification:

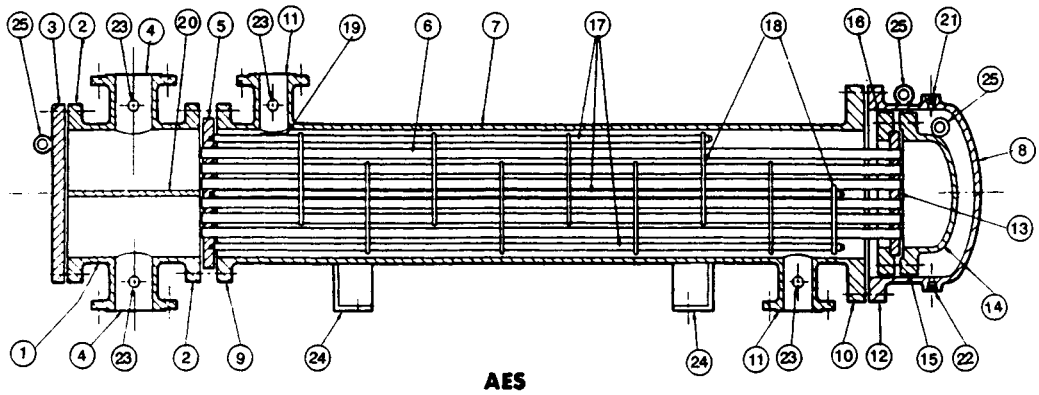


FIGURE 4.5.18 Longitudinal section of a typical shell-and-tube heat exchanger (TEMA AES) with nomenclature. (Modified from TEMA, *Standards* 7th ed., Tubular Exchanger Manufacturers Association, Tarrytown, NY, 1988.)

Shell-and-tube heat exchangers have been constructed with heat transfer areas from less than 0.1 m^2 (1 ft^2) to over 100,000 m^2 (1,000,000 ft^2), for pressures from deep vacuum to over 1000 bar (15,000 psi), for temperatures from near 0 to over 1400 K (2000°F), and for all fluid services including single-phase heating and cooling and multiphase vaporization and condensation. The key to such flexibility is the wide range of materials of construction, forming and joining methods, and design features that can be built into these exchangers (see Schlünder, Vol. 4, 1983; Saunders, 1988; and Yokell, 1990). Most

shell-and-tube heat exchangers are manufactured in conformity with TEMA *Standards* (1988) and the *ASME Boiler and Pressure Vessel Code* (latest edition), but other codes and standards may apply.

Construction Features

In the design process, it is important to consider the mechanical integrity under varying operational conditions and the maintainability (especially cleaning) of the exchanger as equally important with the thermal-hydraulic design.

Tubes. Tubes used in shell-and-tube exchangers range from 6.35 mm ($^1/_4$ in.) to 50.8 mm (2 in.) and above in outside diameter, with the wall thickness usually being specified by the Birmingham wire gauge (BWG). Tubes are generally available in any desired length up to 30 m (100 ft) or more for plain tubes. While plain tubes are widely used, a variety of internally and/or externally enhanced tubes is available to provide special heat transfer characteristics when economically justified (see subsection on enhancement in Section 4.8). Low fin tubes having circumferential fins typically 0.8 to 1.6 mm (0.032 to 0.062 in.) high, spaced 630 to 1260 fins/m (16 to 32 fins/in.) are often employed, especially when the shell-side heat transfer coefficient is substantially smaller than the tube-side coefficient. The outside heat transfer area of a low fin tube is three to six times the inside area, resulting in a smaller heat exchanger shell for the same service, which may offset the higher cost of the tube per unit length.

The tubes are inserted into slightly oversized holes drilled (or, occasionally, punched) through the tubesheets (items 5 and 13, Figure 4.5.18). The tubes are secured by several means, depending upon the mechanical severity of the application and the need to avoid leakage between the streams. In some low-severity applications, the tubes are roller-expanded into smooth holes in the tubesheet. For a stronger joint, two shallow circumferential grooves are cut into the wall of the hole in the tubesheet and the tube roller-expanded into the grooves; to eliminate the possibility of leakage, a seal weld can be run between the outer end of the tube and the tubesheet. Alternatively, the tubes may be strength-welded into the tubesheet.

Tube Supports. It is essential to provide periodic support along the length of the tubes to prevent sagging and destructive vibration caused by the fluid flowing across the tube bank. A secondary role played by the tube supports is to guide the flow back and forth across the tube bank, increasing the velocity and improving the heat transfer on the shell side (but also increasing the pressure drop). The tube support is usually in the form of single segmental baffles (item 18 in Figure 4.5.18) — circular plates with holes drilled to accommodate the tubes and with a segment sheared off to form a “window” or “turnaround” to allow the shell-side fluid to pass from one cross-flow section to the next. The baffles must overlap at least one full row of tubes to give the bundle the necessary rigidity against vibration. When minimizing shell-side pressure drop is not a priority, a baffle cut of 15 to 25% of the shell inside diameter is customary. Baffle spacing is determined first by the necessity to avoid vibration and secondarily to approximately match the free cross-flow area between adjacent baffles to the flow area in the window; i.e., small baffle cuts correspond to closer baffle spacing.

In situations such as low-pressure gas flows on the shell side where pressure drop is severely limited, double segmental and strip baffle arrays can be used. More recently, a helical baffle arrangement has been introduced (Kral et al., 1996) which causes the shell-side fluid to spiral through the exchanger giving improved heat transfer vs. pressure drop characteristics. Where vibration prevention and/or minimum pressure drop are the main concerns, grids of rods or strips can be used (Gentry et al., 1982).

Shells. The shell is the cylinder which confines the shell-side fluid (item 7 in Figure 4.5.18), fitted with nozzles for fluid inlet and exit. Diameters range from less than 50 mm (2 in.) to 3.05 m (10 ft) commonly, and at least twice that value for special applications. In diameters up to 610 mm (24 in.), shells are usually made from standard pipe or tubular goods by cutting to the desired length; in larger sizes, metal plates are rolled to the desired diameter and welded.

A variety of nozzle arrangements are used for special purposes, and TEMA has a standard code to identify the major types, as well as the various front and rear head configurations on the tube side. Figure 4.5.19 shows these configurations with the corresponding code letters.

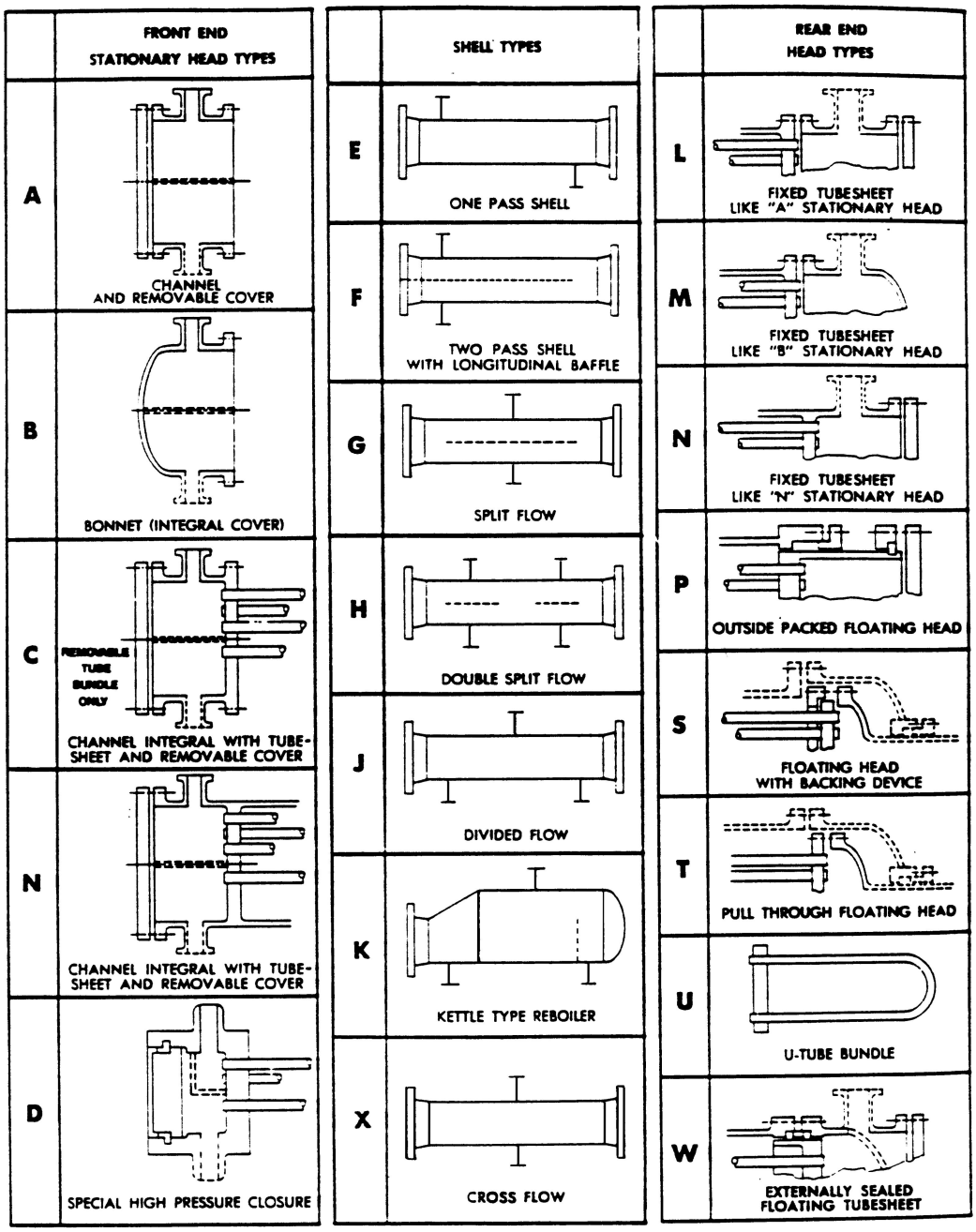


FIGURE 4.5.19 TEMA nomenclature for shell and tube configurations. (From TEMA, *Standards*, 7th ed., Tubular Exchanger Manufacturers Association, Tarrytown, NY, 1988. With permission.)

The E shell (center column, top) has the nozzles on opposite ends of the shell and is the most common configuration. It is used for any of the thermal services (single-phase heating or cooling, vaporization, and condensation). The nozzles may be on opposite sides of the shell as shown, or on the same side; the choice is largely determined by plumbing convenience. The E shell allows countercurrent flow (see below) of the two streams if there is one tube-side pass (i.e., the tube-side fluid flows through all of the tubes in parallel).

The F shell has both nozzles at one end of the shell and uses a longitudinal baffle on the shell side (shown dashed in the drawing) to force the shell-side fluid to flow to the far end of the heat exchanger and then back to the exit nozzle on the other side of the longitudinal baffle. Ideally, this allows countercurrent flow of the two streams if there are two tube-side passes (i.e., the tube-side fluid flows through half of the tubes in one direction, is turned around in the rear head, and returns through the other half of the tubes — see discussion of head types below). However, the longitudinal baffle must be carefully sealed to the shell to prevent leakage of the shell-side fluid across it; this is done by welding the longitudinal baffle to the shell and front tubesheet (which limits some design options) or by using mechanical seals. The F shell is mainly used for sensible heat transfer services.

The G shell has both nozzles at the center of the shell, with a centrally located longitudinal baffle to force the fluid to the ends of the shell before returning. While the G shell is used for all services, its main application is as a shellside vaporizer with either forced or natural (thermosiphon) convection of the boiling fluid; in the latter service, limited leakage across the baffle generally does not greatly degrade the thermal performance and the longitudinal baffle does not need to be perfectly sealed against the shell.

The H shell is effectively a double G shell and is employed in the same services. It is considered when the calculated shell-side pressure drop for a G arrangement is too high and threatens to limit the circulation rate.

The J shell, with one nozzle on top of the shell and two on the bottom, or vice versa, is commonly used in vacuum-condensing applications because of its low pressure drop. Two J shells (one inverted) may be mated in series for long-condensing-range mixtures. The nozzles are usually different diameters, with the large diameter accommodating the inlet vapor. The baffles are vertically cut.

The K shell (or kettle reboiler or flooded chiller) is exclusively intended for vaporization of liquid on the shell side, with a condensing vapor (usually steam) or a hot liquid on the tube side as the heating medium. The tubesheet diameter is large enough to accommodate the tube bundle, but the shell transitions to a larger diameter to allow the vapor to disengage from the liquid pool and exit from the top nozzle. A weir or other level control is used to maintain the liquid level, usually just above the top tubes in the bundle.

The X shell is intended to provide a well-distributed cross flow of the shell-side fluid, the fluid usually entering at the top and exiting at the bottom but occasionally used for upflow or horizontal cross flow. To obtain good distribution, multiple nozzles from a properly designed manifold may be required. Alternatively, the upper tubes in the bundle may be omitted to allow internal redistribution, or a large plenum chamber may be welded to the top of the shell ("vapor dome" or "bathtub nozzle"), or a diverging transition section may be placed between the inlet piping and the top of the shell. The tube supports may be complete circles since there is little or no longitudinal shell-side flow. The X shell gives the lowest shell-side pressure drop of any configuration and is often used for low-pressure vapor condensers.

Front Head. TEMA recognizes several front head designs as shown in the first column of Figure 4.5.19. Any of these designs will get the tube-side fluid into the tubes, but each has special features which recommend it to meet special needs. In Figure 4.5.19 the dashed lines indicate optional features depending upon need.

The A head, a channel with removable cover, bolts directly to the shell flange as shown in Figure 4.5.18, the tubesheet in that case being held between them and sealed with gaskets. Alternatively, the tubesheet may be integral with the shell (see the L rear head in Figure 4.5.19). A removable channel cover permits inspection, cleaning, removal, and replacement of tubes without disturbing the piping. The dashed lines at the center and the lower nozzle indicate that a pass partition plate may be welded

in the channel (and gasketed against the tubesheet and channel cover) to provide for two tube-side passes (as shown in Figure 4.5.18 and required by the F shell design). Additional pass partitions may be provided to allow four, six, or any even number of tube-side passes. This permits the designer to utilize the available tube-side pressure drop to increase velocity, improve the heat transfer coefficient, and possibly reduce fouling. A second nozzle is required on the channel for multipass designs.

The B, or bonnet, front head reduces the number of gasketed joints and thus the opportunity for leakage, but it does not permit inspection of the tubes without breaking the piping connection. It is generally less expensive than the A head.

C and N heads retain the removable cover feature of the A head but, respectively, replace the channel-to-tubesheet and the tubesheet-to-shell gasketed connections with welds to minimize leakage possibilities. The N head is particularly used in nuclear services.

The D head is mainly used in feed-water heater applications where tube-side pressures are in the 100 to 400 bar range. The internal partition (optional) need only withstand the 1 to 2 bar pressure drop through the tubes so it can be of lightweight construction. The high-pressure closure against the atmosphere uses a shear key ring to lock the main closure in place.

Rear Head. A variety of rear head designs are used in shell-and-tube exchangers, primarily because of the need to accommodate thermally induced stresses. During operation, the tubes and the shell have different temperatures and therefore will expand (or try to) different amounts, even if there were no residual stresses in the exchanger before start-up and even if the entire exchanger is made out of the same material. The thermal stress problem is exacerbated if there are residual stresses, or if the exchanger is made of different materials, or during transient operation (including start-up and shutdown). If the temperature differences are small, the structure may be able to accommodate the thermal stresses safely; usually, however, it is necessary to make specific provision to allow the shell and the tubes to expand or contract independently. Failure to do so can result in buckling, bending, or even rupture of the shell or the tubes, or destruction of the tube-to-tubesheet joint.

A simple solution is to incorporate an expansion joint or a bellows into the shell (or in certain special applications, into the tube-side piping internal to the shell cover). However, this solution cannot cover the entire range of pressures and temperature differences encountered in practice. Further, it is usually possible to incorporate other desirable features, such as removable bundles, with thermal stress relief in the variety of rear head designs available. These are shown in the last column of Figure 4.5.19.

The L and M rear heads correspond to the A and B front heads previously described. As shown, they require a fixed tubesheet design; that is, the tubesheets are rigidly fastened to the shell, and thermal stress relief, if necessary, must be provided by a shell-side expansion joint or bellows. The tube bundle cannot be removed for inspection or mechanical cleaning on the shell side. However, the outer tube limit (OTL) — the diameter of the tube field circumscribing the outermost tubes in the bundle — can be as little as 0.4 in. (10 mm) less than the inside diameter of a pipe shell and 0.5 in. (12.7 mm) for a rolled shell. Therefore, the tube field can be very full, giving more tubes and minimizing bypass flow. Similar comments apply to the N rear head, except that more clearance must be left between the outermost tubes and the shell.

The type P head uses packing between the skirt on the rear tubesheet and the shell extension to seal the shell-side fluid against leakage. The compression on the packing has to be adjusted to prevent excessive leakage on the one hand and to allow limited movement of the tube-side head on the other, so the shell-side fluid must be benign and cheap (not surprisingly, it is often cooling water). On the other hand, leakage between the two fluids can occur only through tube hole leaks. Because of the tubesheet skirt, clearance between the outermost tubes and the shell must increase compared with types L or M; accordingly, fewer tubes are possible in a given shell, and sealing strips to partially block the bundle-to-shell bypass stream are recommended. When the floating head cover and packing gland are removed, the tube bundle can be pulled out of the shell for inspection and cleaning.

The TEMA S split-ring floating head design uses a split backing ring to hold the floating head cover and its gasket to the tubesheet. The split backing ring is bolted to the cover with a bolt circle outside

the diameter of the tubesheet. Therefore, when the split ring is removed, the entire tube bundle may be pulled out of the shell. Tube count is similar to type P design and sealing strips are recommended. Usually, the split-ring floating head is used with an even number of tube passes so that a plain bonnet-type shell cover can be used. However, as shown by the dashed lines in Figure 4.5.19, single tube-side pass design (and countercurrent flow) can be achieved by use of a packing gland on the exit piping through the bonnet; alternatively, a deep bonnet can be used together with an expansion joint or bellows on the tube-side exit piping.

The pull-through floating head, type T, uses a floating head cover that flanges directly to the tubesheet, reducing the possibility of internal leakage compared with type S, but also eliminating more tubes around the periphery. Sealing strips are a virtual necessity. Single tube-side pass design is similar to type S, but is rarely used.

TEMA type U uses a bundle of U tubes and hence requires no rear head at all. The U-tube bundle effectively eliminates the thermal stress problem between shell and tubes, because each tube is free to expand or contract independently. The U bundle is also the cheapest construction because the cost of a second tubesheet is avoided. However, there are a number of drawbacks: designs must have an even number of tube-side passes, mechanical cleaning of the smaller bend radius tubes in the U bend is impossible, individual tubes cannot be replaced except in the outer row, some tube count is lost because of minimum bend limits, and the U bend must be carefully supported against vibration or kept out of the cross-flow stream by placing the shell nozzle upstream of the bend. The tube side in the U bend is susceptible to erosion, especially with two-phase or particulate-containing fluids.

Type W uses two sets of packing, often with a lantern ring in between. This construction is generally limited to benign fluids and low to very moderate pressures and temperatures.

Other Features. Numerous other components are necessary or optional to construction of shell-and-tube exchangers. Probably the most complete discussion is given by Yokell (1990).

Principles of Design

Design Logic. The design of a shell-and-tube exchanger involves the following steps:

1. Selection of a set of design features which are required for mechanical integrity and ease of maintenance, and which will likely lead to satisfying the thermal requirements within the allowable pressure drops, and at lowest cost.
2. Selection of a set of dimensions for the actual exchanger.
3. For the dimensions selected in (2), calculation of the thermal performance of the heat exchanger and both tube-side and shell-side pressure drops, using available rating procedures.
4. Comparison of the thermal performance calculated in (3) with that required and examination of the pressure drops calculated in (3) to ensure that the allowed pressure drops are reasonably used but not exceeded.
5. Adjustment of the dimensions selected in (2) and repetition of steps (3) and (4) until the criteria are satisfied.
6. Completion of the mechanical design to satisfy code requirements.
7. Cost estimation.

Basic Design Equations. The basic design equation for a shell-and-tube exchanger in steady-state service is

$$A^* = \int_0^{q_T} \frac{dq}{U^*\left(T_h - T_c\right)} \tag{4.5.87}$$

where A^* is the heat transfer area required in the heat exchanger, m^2 (ft^2); q_T is the heat transfer rate of the heat exchanger, W (Btu/hr); U^* is the local overall heat transfer coefficient referenced to area A^*, W/m^2 K (Btu/hr ft^2 °F); and T_h and T_c are the local hot and cold stream temperatures, K (°F). The *

superscript on A^* and U^* only means that a consistent reference area must be used in defining these terms. For example, for an exchanger with plain tubes, it is customary to use the total outside heat transfer area of all of the tubes in the exchanger, A_o, as the reference area, and then U_o is the overall heat transfer coefficient referenced to A_o. If the exchanger has low-finned tubes, A^* may refer either to the total outside area including fins or to the inside tube heat transfer area; the choice is optional, but must be spelled out. Since T_h and T_c generally vary with the amount of heat transferred (following the first law of thermodynamics, and excepting isobaric phase transition of a pure component) and U^* may vary with local heat transfer conditions, in principle Equation 4.5.87 must be numerically integrated with T_h, T_c, and U^* calculated along the path of integration, and this process is performed by the most-advanced computer-based design methods.

For many applications, certain reasonable assumptions can be made allowing the analytical integration of Equation 4.5.87 to give (Schlünder, Vol. 1, 1983; Hewitt et al., 1994)

$$A^* = \frac{q_T U^*}{(\text{MTD})} \tag{4.5.88}$$

where MTD is the mean temperature difference for the particular flow conditions and configuration. The key assumptions are that there is no significant bypassing of fluid around the heat transfer surface, that the overall heat transfer coefficient is constant, and that the specific heats of the two streams are constant over their temperature ranges in the exchanger; isothermal phase transitions, such as vaporizing or condensing a pure component at constant pressure, are also allowed.

If the two streams are in countercurrent flow, i.e., if they flow always in the opposite direction to one another,

$$\text{MTD} = (\text{LMTD})_{\text{countercurrent}} = \frac{\left(T_{h,i} - T_{c,o}\right) - \left(T_{h,o} - T_{c,i}\right)}{\ln\left(\dfrac{T_{h,i} - T_{c,o}}{T_{h,o} - T_{c,i}}\right)} \tag{4.5.89}$$

where $(\text{LMTD})_{\text{countercurrent}}$ is the "logarithmic mean temperature difference for countercurrent flow" and the subscripts i and o indicate "inlet" and "outlet," respectively. E shells with a single tube-side pass and F shells with two tube-side passes are almost always designed for countercurrent flow. (While the flow between adjacent baffles is basically cross flow, it can be shown that the total shell-side flow pattern is equivalent to countercurrent flow if there are more than three or four baffles).

Very occasionally, usually when close control of tube wall temperatures is required, cocurrent flow is specified, with the two streams flowing in the same direction through the exchanger. For this case,

$$\text{MTD} = (\text{LMTD})_{\text{cocurrent}} = \frac{\left(T_{h,i} - T_{c,i}\right) - \left(T_{h,o} - T_{c,o}\right)}{\ln\left(\dfrac{T_{h,i} - T_{c,i}}{T_{h,o} - T_{c,o}}\right)} \tag{4.5.90}$$

where the symbols have the same meaning as before. $(\text{LMTD})_{\text{countercurrent}}$ is always equal to or greater than $(\text{LMTD})_{\text{cocurrent}}$, so wherever possible, countercurrent design and operation is preferred.

However, most shell-and-tube exchangers have nozzle and tube pass configurations which lead to mixed countercurrent and cocurrent flow regions (as well as cross flow in the X shell). For these cases,

$$\text{MTD} = F(\text{LMTD})_{\text{countercurrent}} \tag{4.5.91}$$

where $(LMTD)_{countercurrent}$ is calculated from Equation (4.5.89) and F is the "configuration correction factor" for the flow configuration involved. F has been found as a function of dimensionless temperature ratios for most flow configurations of interest and is given in analytical and/or graphical form in the earlier part of this section by Shah and in many heat transfer references (e.g., Schlünder, Vol. 1, 1983). F is equal to unity for pure countercurrent flow and is less than unity for all other cases; practical considerations limit the range of interest to values above 0.7 at the lowest and more comfortably to values above 0.8. Values of zero or below indicate conditions that violate the second law of thermodynamics.

The Overall Heat Transfer Coefficient. The overall heat transfer coefficient U^*, referenced to the heat transfer area A^*, is related to the individual (film) heat transfer coefficients and the fouling resistances by

$$U^* = \frac{1}{\dfrac{A^*}{h_i A_i} + R_{fi}\dfrac{A^*}{A_i} + \dfrac{A^* \ln\left(d_o/d_i\right)}{2\pi N_t L k_w} + R_{fo}\dfrac{A^*}{A_o} + \dfrac{A^*}{h_o A_o}} \tag{4.5.92}$$

where h_i and h_o are, respectively, the tube-side and shell-side film heat transfer coefficients, W/m^{2}K (Btu/hr ft^2 °F), each referenced to its corresponding heat transfer area; R_{fi} and R_{fo} the corresponding fouling resistances (see below), m^{2}K/W (hr ft^2 °F/Btu); N_t the total number of tubes in the heat exchanger; L the effective tube length between the inside surfaces of the tubesheets, m (ft); d_o and d_i the outside and inside tube diameters, m (ft); and k_w the thermal conductivity of the tube wall material, W/m K (Btu/hr ft°F). For the special but important case of plain tubes

$$A^* = A_o = N_t\left(\pi d_o L\right) \tag{4.5.93}$$

and Equation (4.5.92) reduces to

$$U_o = \frac{1}{\dfrac{d_o}{h_i d_i} + R_{fi}\dfrac{d_o}{d_i} + \dfrac{d_o \ln\left(d_o/d_i\right)}{2k_w} + R_{fo} + \dfrac{1}{h_o}} \tag{4.5.94}$$

If finned tubes are used, the root diameter d_r of the fins replaces d_o in Equation (4.5.92) and A_o includes the surface area of the fins as well as the bare tube surface between the fins; it is also necessary to include a fin efficiency (typically about 0.8 to 0.95) multiplier in the numerators of the last two terms on the right side of Equation (4.5.92) to account for resistance to conduction in the fins. The treatment of fin efficiency is fully developed in Kern and Kraus (1972). Efficiencies of some of the important geometries are given in the earlier half of this section.

Film Heat Transfer Coefficients. Calculation of single-phase tube-side heat transfer coefficients for plain tubes is discussed in Section 4.1; special correlations are required for internally enhanced tubes, see discussion of enhancement in Section 4.8. Intube condensation and vaporization are covered in the subsection on boiling and condensation in Section 4.4.

Shell-side heat transfer calculations are more complex owing to the large number and range of design variables and process conditions that can occur. The most accurate methods are proprietary and computer based. The best known of these methods are those of Heat Transfer Research, Inc. (HTRI), College Station, TX; Heat Transfer and Fluid Flow Services (HTFS), Harwell, U.K.; and B-JAC, Midlothian, VA. For single-phase flow, the Delaware method appears to be the best in the open literature, and it is feasible for both had and computer use; various presentations of the method appear in many references, including Schlünder, Vol. 3 (1983) and Hewitt et al. (1994). These references also give methods for

shell-side vaporizing and condensing design. An approximate design procedure is given in the next subsection.

Fouling. Fouling is the formation of any undesired deposit on the heat transfer surface, and it presents an additional resistance to the flow of heat. Several different types of fouling are recognized:

Sedimentation: deposition of suspended material on the surface.
Crystallization: precipitation of solute from supersaturated solutions.
Corrosion: formation of corrosion products on the surface.
Thermal degradation/polymerization: formation of insoluble products by oxidation, charring, and/or polymerization of a process stream.
Biofouling: growth of large organisms (e.g., barnacles) that interfere with flow to or past a heat transfer surface ("macrobiofouling") or small organisms (e.g., algae) that form a fouling layer on the surface ("microbiofouling").

The effect of fouling on design is twofold: Extra surface must be added to the heat exchanger to overcome the additional thermal resistance, and provision must be made to allow cleaning either by chemical or mechanical means. The fouling resistances included in Equation (4.5.92) result in requiring extra surface by reducing U^* (though they do not properly account for the time-dependent nature of fouling) and should be chosen with care. Table 4.5.8, based on the TEMA *Standards* provides some guidance, but prior experience with a given service is the best source of values. Ranges of typical values for major classes of service are included in Table 4.5.10.

Other things being equal, a fouling stream that requires mechanical cleaning should be put in the tubes because it is easier to clean the tube side. If this is not possible or desirable, then a removable bundle with a rotated square tube layout should be chosen to facilitate cleaning.

Pressure Drop. Tube-side pressure drop in plain tubes is discussed in Section 3.4. These calculations are straightforward and quite accurate as long as the tubes are smooth and clean; however, even a small amount of roughening due to corrosion or fouling (sometimes with a significant reduction of flow area) can double or triple tube-side pressure drop. Special correlations are required for internally enhanced tubes.

Calculation of shell-side pressure drop is implicit in the design methods mentioned above for heat transfer. Roughness has less effect on shell-side pressure drop than on tube side, but fouling still may have a very substantial effect if the deposits fill up the clearances between the baffles and the shell and between the tubes and the baffles, or if the deposits are thick enough to narrow the clearances between adjacent tubes. Existing design methods can predict these effects if the thickness of the fouling layer can be estimated.

Limitations of Design. It should be recognized that even under the best of conditions — new, clean exchangers with conventional construction features — heat exchanger design is not highly accurate. The best methods, when compared with carefully taken test data, show deviations of ±20% on overall heat transfer and ±40% on shell-side pressure drop (Palen and Taborek, 1969). These ranges are considerably worsened in fouling services. In these cases, the thermal *system* should be designed for operational flexibility, including carefully chosen redundancy of key components, and easy maintenance.

Approximate Design Method

Because of the complexity of rigorous design methods, it is useful to have an estimation procedure that can quickly give approximate dimensions of a heat exchanger for a specified service. Such a method is given here for purposes of preliminary cost estimation, plant layout, or checking the results of computer output. This method is based upon Equation (4.5.88) with $A^* = A_o$ and $U^* = U_o$ and depends upon rapidly estimating values for q_T, MTD, and U_o. The procedure is as follows:

TABLE 4.5.10 Typical Film Heat Transfer Coefficients for Shell-and-Tube Heat Exchangers

Fluid Conditions		h, W/m²K[a,b]	Fouling resistance, m² K/W[a]
Sensible heat transfer			
Water[c]	Liquid	5000–7500	$1–2.5 \times 10^{-4}$
Ammonia	Liquid	6000–8000	$0–1 \times 10^{-4}$
Light organics[d]	Liquid	1500–2000	$0–2 \times 10^{-4}$
Medium organics[e]	Liquid	750–1500	$1–4 \times 10^{-4}$
Heavy organics[f]	Liquid		
	Heating	250–750	$2–10 \times 10^{-4}$
	Cooling	150–400	$2–10 \times 10^{-4}$
Very heavy organics[g]	Liquid		
	Heating	100–300	$4–30 \times 10^{-3}$
	Cooling	60–150	$4–30 \times 10^{-3}$
Gas[h]	Pressure 100–200 kN/m² abs	80–125	$0–1 \times 10^{-4}$
Gas[h]	Pressure 1 MN/m² abs	250–400	$0–1 \times 10^{-4}$
Gas[h]	Pressure 10 MN/m² abs	500–800	$0–1 \times 10^{-4}$
Condensing heat transfer			
Steam, ammonia	Pressure 10 kN/m² abs, no noncondensables[i,j]	8000–12000	$0–1 \times 10^{-4}$
Steam, ammonia	Pressure 10 kN/m² abs, 1% noncondensables[k]	4000–6000	$0–1 \times 10^{-4}$
Steam, ammonia	Pressure 10 kN/m² abs, 4% noncondensables[k]	2000–3000	$0–1 \times 10^{-4}$
Steam, ammonia	Pressure 100 kN/m² abs, no noncondensables[i,j,k,l]	10000–15000	$0–1 \times 10^{-4}$
Steam, ammonia	Pressure 1 MN/m² abs, no noncondensables[i,j,k,l]	15000–25,000	$0–1 \times 10^{-4}$
Light organics[d]	Pure component, pressure 10 kN/m² abs, no noncondensables[i]	1500–2000	$0–1 \times 10^{-4}$
Light organics[d]	Pressure 10 kN/m² abs, 4% noncondensables[k]	750–1000	$0–1 \times 10^{-4}$
Light organics[d]	Pure component, pressure 100 kN/m² abs, no noncondensables	2000–4000	$0–1 \times 10^{-4}$
Light organics[d]	Pure component, pressure 1 MN/m² abs	3000–4000	$0–1 \times 10^{-4}$
Medium organics[e]	Pure component or narrow condensing range, pressure 100 kN/m² abs[m,n]	1500–4000	$1–3 \times 10^{-4}$
Heavy organics	Narrow condensing range pressure 100 kN/m² abs[m,n]	600–2000	$2–5 \times 10^{-4}$
Light multicomponent mixtures, all condensable[d]	Medium condensing range, pressure 100 kN/m² abs[k,m,o]	1000–2500	$0–2 \times 10^{-4}$
Medium multicomponent mixtures, all condensable[e]	Medium condensing range, pressure 100 kN/m² abs[k,m,o]	600–1500	$1–4 \times 10^{-4}$
Heavy multicomponent mixtures, all condensable[f]	Medium condensing range, pressure 100 kN/m² abs[k,m,o]	300–600	$2–8 \times 10^{-4}$
Vaporizing heat transfer[p,q]			
Water[r]	Pressure < 0.5 MN/m² abs, $\Delta T_{SH,max}$ = 25 K	3000–10000	$1–2 \times 10^{-4}$
Water[r]	Pressure < 0.5 MN/m² abs, pressure < 10 MN/m² abs, $\Delta T_{SH,max}$ = 20 K	4000–15000	$1–2 \times 10^{-4}$

TABLE 4.5.10 (continued) Typical Film Heat Transfer Coefficients for Shell-and-Tube Heat Exchangers

Fluid Conditions		h, W/m²K[a,b]	Fouling resistance, m² K/W[a]
Ammonia	Pressure < 3 MN/m² abs, $\Delta T_{SH,max}$ = 20 K	3000–5000	$0–2 \times 10^{-4}$
Light organics[d]	Pure component, pressure < 2 MN/m² abs, $\Delta T_{SH,max}$ = 20 K	1000–4000	$1–2 \times 10^{-4}$
Light organics[d]	Narrow boiling range,[s] pressure < 2 MN/m² abs, $\Delta T_{SH,max}$ = 15 K	750–3000	$0–2 \times 10^{-4}$
Medium organics[e]	Pure component, pressure < 2 MN/m² abs, $\Delta T_{SH,max}$ = 20 K	1000–3500	$1–3 \times 10^{-4}$
Medium organics[e]	Narrow boiling range,[s] pressure < 2 MN/m² abs, $\Delta T_{SH,max}$ = 15 K	600–2500	$1–3 \times 10^{-4}$
Heavy organics[f]	Pure component, pressure < 2 MN/m² abs, $\Delta T_{SH,max}$ = 20 K	750–2500	$2–5 \times 10^{-4}$
Heavy organics[g]	Narrow boiling range,[s] pressure < 2 MN/m² abs, $\Delta T_{SH,max}$ = 15 K	400–1500	$2–8 \times 10^{-4}$
Very heavy organics[h]	Narrow boiling range,[s] pressure < 2 MN/m² abs, $\Delta T_{SH,max}$ = 15 K	300–1000	$2–10 \times 10^{-4}$

Source: Schlünder, E.U., Ed., *Heat Exchanger Design Handbook*, Begell House, New York, 1983. With permision.

[a] Heat transfer coefficients and fouling resistances are based on area in contact with fluid. Ranges shown are typical, not all encompassing. Temperatures are assumed to be in normal processing range; allowances should be made for very high or low temperatures.

[b] Allowable pressure drops on each side are assumed to be about 50–100 kN/m² except for (1) low-pressure gas and two-phase flows, where the pressure drop is assumed to be about 5% of the absolute pressure; and (2) very viscous organics, where the allowable pressure drop is assumed to be about 150–250 kN/m².

[c] Aqueous solutions give approximately the same coefficients as water.

[d] Light organics include fluids with liquid viscosities less than about 0.5×10^{-3} Nsec/m², such as hydrocarbons through C_8, gasoline, light alcohols and ketones, etc.

[e] Medium organics include fluids with liquid viscosities between about 0.5×10^{-3} and 2.5×10^{-3} Nsec/m², such as kerosene, straw oil, hot gas oil, and light crudes.

[f] Heavy organics include fluids with liquid viscosities greater than 2.5×10^{-3} Nsec/m², but not more than 50×10^{-3} Nsec/m², such as cold gas oil, lube oils, fuel oils, and heavy and reduced crudes.

[g] Very heavy organics include tars, asphalts, polymer melts, greases, etc., having liquid viscosities greater than about 50×10^{-3} Nsec/m². Estimation of coefficients for these materials is very uncertain and depends strongly on the temperature difference, because natural convection is often a significant contribution to heat transfer in heating, whereas congelation on the surface and particularly between fins can occur in cooling. Since many of these materials are thermally unstable, high surface temperatures can lead to extremely severe fouling.

[h] Values given for gases apply to such substances as air, nitrogen, carbon dioxide, light hydrocarbon mixtures (no condensation), etc. Because of the very high thermal conductivities and specific heats of hydrogen and helium, gas mixtures containing appreciable fractions of these components will generally have substantially higher heat transfer coefficients.

[i] Superheat of a pure vapor is removed at the same coefficient as for condensation of the saturated vapor if the exit coolant temperature is less than the saturation temperature (at the pressure existing in the vapor phase) and if the (constant) saturation temperature is used in calculating the MTD. But see note k for vapor mixtures with or without noncondensable gas.

[j] Steam is not usually condensed on conventional low-finned tubes; its high surface tension causes bridging and retention of the condensate and a severe reduction of the coefficient below that of the plain tube.

TABLE 4.5.10 (continued) Typical Film Heat Transfer Coefficients for Shell-and-Tube Heat Exchangers

Fluid Conditions	h, W/m^{2}K[a,b]	Fouling resistance, m^2 K/W[a]

[k] The coefficients cited for condensation in the presence of noncondensable gases or for multicomponent mixtures are only for very rough estimation purposes because of the presence of mass transfer resistances in the vapor (and to some extent, in the liquid) phase. Also, for these cases, the vapor-phase temperature is not constant, and the coefficient given is to be used with the MTD estimated using vapor-phase inlet and exit temperatures, together with the coolant temperatures.

[l] As a rough approximation, the same relative reduction in low-pressure condensing coefficients due to noncondensable gases can also be applied to higher pressures.

[m] Absolute pressure and noncondensables have about the same effect on condensing coefficients for medium and heavy organics as for light organics. For large fractions of noncondensable gas, interpolate between pure component condensation and gas cooling coefficients.

[n] Narrow condensing range implies that the temperature difference between dew point and bubble point is less than the smallest temperature difference between vapor and coolant at any place in the condenser.

[o] Medium condensing range implies that the temperature difference between dew point and bubble point is greater than the smallest temperature difference between vapor and coolant, but less than the temperature difference between inlet vapor and outlet coolant.

[p] Boiling and vaporizing heat transfer coefficients depend very strongly on the nature of the surface and the structure of the two-phase flow past the surface in addition to all of the other variables that are significant for convective heat transfer in other modes. The flow velocity and structure are very much governed by the geometry of the equipment and its connecting piping. Also, there is a maximum heat flux from the surface that can be achieved with reasonable temperature differences between surface and saturation temperatures of the boiling fluid; any attempt to exceed this maximum heat flux by increasing the surface temperature leads to partial or total coverage of the surface by a film of vapor and a sharp decrease in the heat flux.
Therefore, the vaporizing heat transfer coefficients given in this table are only for very rough estimating purposes and assume the use of plain or low-finned tubes without special nucleation enhancement.
$\Delta T_{SH,max}$ is the maximum allowable temperature difference between surface and saturation temperature of the boiling liquid. No attempt is made in this table to distinguish among the various types of vapor-generation equipment, since the major heat transfer distinction to be made is the propensity of the process stream to foul. Severely fouling streams will usually call for a vertical thermosiphon or a forced-convection (tube-side) reboiler for ease of cleaning.

[q] Subcooling heat load is transferred at the same coefficient as latent heat load in kettle reboilers, using the saturation temperature in the MTD. For horizontal and vertical thermosiphons and forced-circulation reboilers, a separate calculation is required for the sensible heat transfer area, using appropriate sensible heat transfer coefficients and the liquid temperature profile for the MTD.

[r] Aqueous solutions vaporize with nearly the same coefficient as pure water if attention is given to boiling-point elevation, if the solution does not become saturated, and if care is taken to avoid dry wall conditions.

[s] For boiling of mixtures, the saturation temperature (bubble point) of the final liquid phase (after the desired vaporization has taken place) is to be used to calculate the MTD. A narrow-boiling-range mixture is defined as one for which the difference between the bubble point of the incoming liquid and the bubble point of the exit liquid is less than the temperature difference between the exit hot stream and the bubble point of the exit boiling liquid. Wide-boiling-range mixtures require a case-by-case analysis and cannot be reliably estimated by these simple procedures.

Estimation of q_T. For sensible heat transfer,

$$q_T = \dot{m}_h c_{p,h}\left(T_{h,i} - T_{h,o}\right) = \dot{m}_c c_{p,c}\left(T_{c,o} - T_{c,i}\right) \tag{4.5.95}$$

where $\dot{m}$ is the mass flow rate, c_p the specific heat, and T the stream temperature, with subscripts h and c denoting the hot and cold streams, respectively, and i and o inlet and outlet, respectively.

For isothermal phase change,

$$q_T = \dot{m} h_{fg} \tag{4.5.96}$$

where $\dot{m}$ is the mass rate of condensation or vaporization and h_{fg} is the latent heat of phase transformation.

For more complex cases, such as partial or multicomponent condensation, more elaborate analyses are required, although this method can still be used with care to give rough estimates.

Estimation of MTD. The first step is to calculate or estimate $\text{LMTD}_{\text{countercurrent}}$ from Equation (4.5.89) and then estimate F as follows:

1. If the two streams are in countercurrent flow, $F = 1$.
2. If the two streams are in a combination of countercurrent and cocurrent flows (i.e., multiple tube passes) and the outlet temperatures of the two streams are equal, $F = 0.8$.
3. If the exchanger has multiple passes and $T_{h,o} > T_{c,o}$, then $0.8 < F < 1.0$, with the actual value depending upon the temperature ranges of the two streams and $(T_{h,o} - T_{c,o})$. It is usually sufficiently accurate to take $F = 0.9$ in this case, but a more accurate value can be obtained from the earlier half of this section by Shah.
4. Design of a multiple tube pass exchanger with $T_{h,o} < T_{c,o}$ (i.e., a temperature cross) leads to $F < 0.8$, which is inefficient, of uncertain inaccuracy, and perhaps even thermodynamically impossible. The problem can be handled with multiple shells in series. Consult Shah's discussion.
5. Then, MTD = $F(\text{LMTD})_{\text{countercurrent}}$, (Equation 4.5.91).

Estimation of U_o. The best way to estimate U_o is to use Equation (4.5.94), together with values of h_o, h_i, $R_{f,o}$, and $R_{f,i}$, chosen from Table 4.5.10. This table includes ranges of values that are typical of the fluids and services indicated assuming normally allowable pressure drops, exchanger construction, and fouling. However, care should be taken in selecting values to consider possible unusual conditions, e.g., especially high or low velocities (implying correspondingly high or low allowable pressure drops), and especially fouling. In selecting values from the table, the user should carefully read the footnotes for each entry.

Calculation of A_o. The total outside tube heat transfer area required in the heat exchanger is now found from Equation (4.5.88).

Estimation of Exchanger Dimensions. Figure 4.5.20 shows the relationship among A_o, effective tube length L, and inside shell diameter for a fully tubed, fixed tubesheet heat exchanger with one tube-side pass, with $^3/_4$ in. (19.05 mm) plain tubes on a $^{15}/_{16}$ in. (23.8 mm) pitch equilateral triangular tube layout. These curves are constructed using tube count tables (e.g., Saunders, 1988). The dashed lines marked 3:1, 6:1, 8:1, 10:1, and 15:1 indicate ratios of tube length to shell inside diameter for guidance in selection. Exchangers of less than 3:1 ratio are expensive because of the large-diameter shell and tubesheet, with more holes to be drilled and tubes rolled and/or welded, and shell-side flow distribution is likely to be poor and lead to excessive fouling. Exchangers greater than 15:1 ratio are probably beyond the point of saving money by reducing shell diameter and number of tubes and may require excessive clear way for pulling the bundle; the bundles may be springy and difficult to handle during maintenance. Most heat exchangers fall into the 6:1 to 10:1 range.

Figure 4.5.20 is a very specific case which is used as a reference. In order to extend its usefulness to other tube diameters, layouts, bundle constructions, etc., Equation (4.5.97) is used:

$$A_o' = A_o F_1 F_2 F_3 \tag{4.5.97}$$

where A_o' is the value to be used with Figure 4.5.20, A_o is the required area calculated from Equation (4.5.88), and

F_1 is the correction factor for the tube layout. $F_1 = 1.00$ for $^3/_4$ in. (19.05 mm) outside diameter tubes on a $^{15}/_{16}$ in. (23.8 mm) triangular pitch. Values of F_1 for other tube diameters and pitches are given in Table 4.5.11.

F_2 is the correction factor for the number of tube-side passes. $F_2 = 1.00$ for one tube-side pass, and Table 4.5.12 gives values of F_2 for more passes.

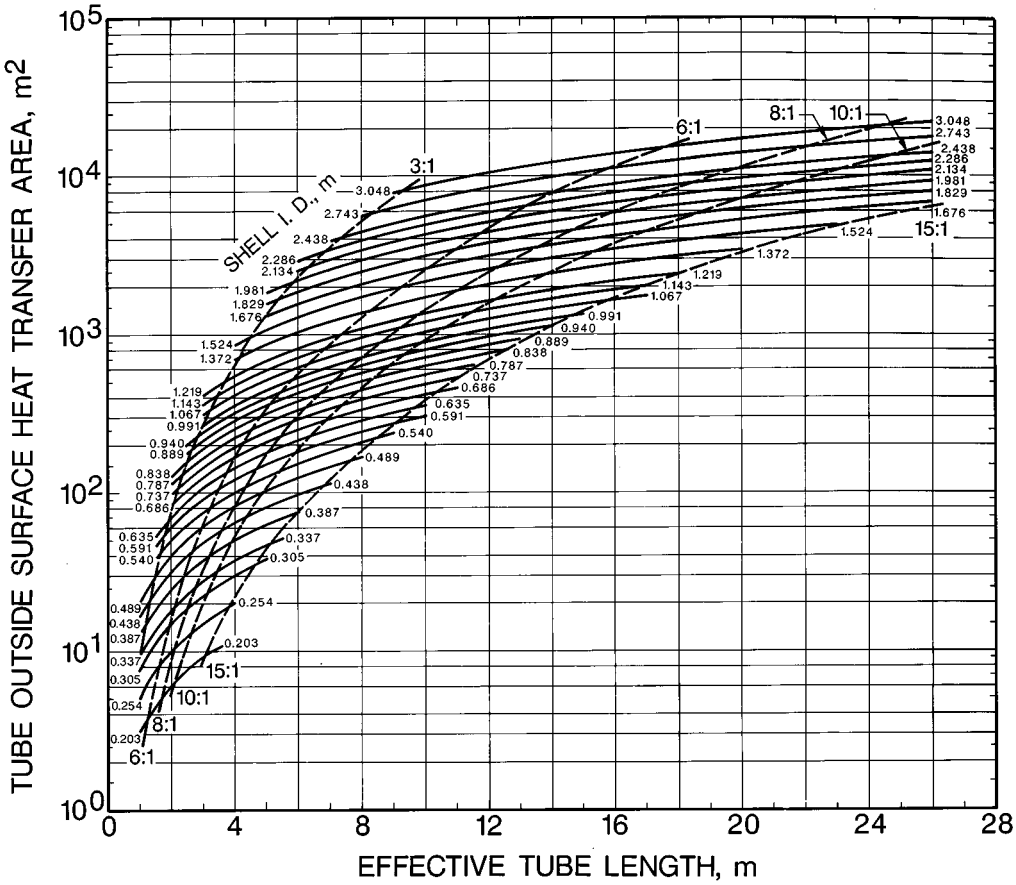


FIGURE 4.5.20 Heat transfer area as a function of shell inside diameter and effective tube length for 19.05 mm ($^3/_4$ in.) tubes on a 23.8 mm ($^{15}/_{16}$ in.) equilateral triangular tube layout, fixed tubesheet, one tube-side pass, fully tubed shell. (From Schlünder, E. U., Ed. *Heat Exchanger Design Handbook*, Begell House, New York, 1983. With permission.)

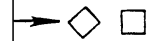
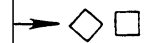
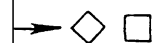
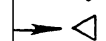
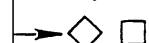
F_3 is the correction factor for shell construction/tube bundle configuration. $F_3 = 1.00$ for fixed tubesheet, fully tubed shells, and Table 4.5.13 gives values of F_3 for the standard TEMA types.

Once a value of A_o' has been calculated from Equation (4.5.97), enter the ordinate of Figure 4.5.20 at that value and move horizontally, picking off the combinations of shell inside diameter and tube length that meet that requirement. The final choice can then be made from among those possibilities.

Example of the Approximate Design Method

Problem Statement. Estimate the dimensions of a shell-and-tube heat exchanger to cool 100,000 lb_m/hr (12.6 kg/sec) of liquid toluene from 250 to 110°F (121.1 to 43.3°C) using cooling tower water available at 80°F (26.7°C). Use split-ring floating head construction (TEMA S) with $^3/_4$ in. (19.05 mm) outside diameter × 14 BWG (0.083 in. = 2.11 mm wall) low-carbon steel tubes on $^{15}/_{16}$ in. (23.8 mm) equilateral triangular pitch. This construction implies one shell-side pass and an even number of tube-side passes — assume two for the present. Choose cooling water exit temperature of 100°F (37.8°C). Specific heat of toluene is 0.52 Btu/lb_m°F (2177 J/kgK) and viscosity at 180°F (82.2°C) is 0.82 $lb/_m$/ft hr (0.34×10^{-3} Nsec/m^2 or 0.34 cP).

TABLE 4.5.11 Values of F_1 for Various Tube Diameters and Layouts

Tube Outside Diameter, in. (mm)	Tube Pitch, in. (mm)	Layout	F_1
5/8 (15.88)	13/16 (20.6)		0.90
5/8 (15.88)	13/16 (20.6)		1.04
3/4 (19.05)	15/16 (23.8)		1.00
3/4 (19.05)	15/16 (23.8)		1.16
3/4 (19.05)	1 (25.4)		1.14
3/4 (19.05)	1 (25.4)		1.31
1 (25.4)	1 1/4 (31.8)		1.34
1 (25.4)	1 1/4 (31.8)		1.54

$$F_1 = \frac{\text{(Heat transfer area / cross–sectional area of unit cell)}_{\text{Reference}}}{\text{(Heat transfer area / cross–sectional area of unit cell)}_{\text{New Case}}}$$

This table may also be used for low-finned tubing in the following way. The value estimated for h_o from Table 4.5.10 should be multiplied by the fin efficiency (usually between 0.75 and 1 for a good application; 0.85 is a good estimate) and used in Equation 4.5.92 with $A^* = A_o$, the total outside heat transfer area including fins. Then this value of A_o is divided by the ratio of the finned tube heat transfer area to the plain tube area (per unit length). The result of this calculation is used as A_o in Equation 4.5.96 to find A_o' to enter Figure 4.5.20.

Source: Schlünder, E.U., Ed., *Heat Exchanger Design Handbook*, Begell House, New York, 1983. With permission.

TABLE 4.5.12 Values of F_2 for Various Numbers of Tube Side Passes[a]

	F_2 Number of Tube-Side Passes			
Inside Shell Diameter, in. (mm)	**2**	**4**	**6**	**8**
Up to 12 (305)	1.20	1.40	1.80	—
$13^1/_4$ to $17^1/_4$ (337 to 438)	1.06	1.18	1.25	1.50
$19^1/_4$ to $23^1/_4$ (489 to 591)	1.04	1.14	1.19	1.35
25 to 33 (635 to 838)	1.03	1.12	1.16	1.20
35 to 45 (889 to 1143)	1.02	1.08	1.12	1.16
48 to 60 (1219 to 1524)	1.02	1.05	1.08	1.12
Above 60 (above 1524)	1.01	1.03	1.04	1.06

[a] Since U-tube bundles must always have at least two passes, use of this table is essential for U-tube bundle estimation. Most floating head bundles also require an even number of passes.

Source: Schlünder, E.U., Ed., *Heat Exchanger Design Handbook*, Begell House, New York, 1985. With permission.

Solution.

$$q_T = (100{,}000\ \text{lb}_\text{m}/\text{hr})(0.52\ \text{Btu}/\text{lb}_\text{m}\,^\circ\text{F})(250-110)^\circ\text{F}$$

$$= 7.28\times10^6\ \text{Btu}/\text{hr} = 2.14\times10^6\ \text{W}$$

TABLE 4.5.13 F_3 for Various Tube Bundle Constructions

Type of Tube Bundle Construction	F_3 Inside Shell Diameter, in. (mm)				
	Up to 12 (305)	13–22 (330–559)	23–36 (584–914)	37–48 (940–1219)	Above 48 (1219)
Split backing ring (TEMA S)	1.30	1.15	1.09	1.06	1.04
Outside packed floating heat (TEMA P)	1.30	1.15	1.09	1.06	1.04
U-Tube* (TEMA U)	1.12	1.08	1.03	1.01	1.01
Pull-through floating head (TEMA T)	—	1.40	1.25	1.18	1.15

[a] Since U-tube bundles must always have at least two tube-side passes, it is essential to use Table 4.5.12 also for this configuration.

Source: Schlünder, E.U., Ed., *Heat Exchanger Design Handbook*, Begell House, New York, 1983. With permission.

$$\text{LMTD}_{\text{countercurrent}} = \frac{(250-100)-(110-80)}{\ln\dfrac{250-100}{110-80}} = 74.6°\text{F} = 41.4°\text{C}$$

Since there are at least two tube-side passes, flow is not countercurrent, and $T_{h_o} > T_{c_o}$, estimate $F \approx 0.9$. Therefore, MTD = 0.9 (74.6°F) = 67.1°F = 37.3°C.

Estimation of U_o. Light organic liquid cooled by liquid water. (Note that 1 Btu/hr ft² °F = 5.678 W/m²K).

Water (in tubes) h_i	1000 Btu/hr ft² °F	5700 W/m²K
Toluene (in shell) h_o	300 Btu/hr ft² °F	1700 W/m²K
Tube-side fouling R_{f_i}	0.001 hr ft² °F/Btu	1.8×10^{-4} m²K/W
Shell-side fouling R_{f_o}	0.0005 hr ft² °F/Btu	8.8×10^{-5} m²K/W

Tube wall resistance (for estimation purposes, this term can be approximated by x_w/k_w, where x_w is the wall thickness):

$$\frac{x_w}{k_w} = \frac{0.083 \text{ in.}}{(12 \text{ in./ft})(26 \text{ Btu/hr ft °F})} = 2.7\times10^{-4}\,\frac{\text{hr ft}^2\,°\text{F}}{\text{Btu}} = 4.6\times10^{-5}\,\frac{\text{m}^2\text{K}}{\text{W}}$$

Then,

$$U_o = \frac{1}{\dfrac{0.750}{1000(0.584)} + \dfrac{0.001(0.750)}{0.584} + 2.7\times10^{-4} + 0.0005 + \dfrac{1}{300}}$$

$$= 150 \text{ Btu/hr ft °F} = 848 \text{ W/m}^2\text{K}$$

$$A_o = \frac{7.28\times10^6 \text{ Btu/hr}}{(150 \text{ Btu/hr ft}^2 °\text{F})(67.1°\text{F})} = 723 \text{ ft}^2 = 67.7 \text{ m}^2$$

Correct for changes in construction features (preliminary examination of Figure 4.5.20 indicates shell inside diameter will be in the range of 500 mm, or 20 in.):

F_1: F_1 = 1.00 since the same tube size and layout is used;
F_2: F_2 = 1.04, assuming two passes;
F_3: F_3 = 1.15, TEMA S construction;

A'_o = (723 ft^2) (1.00)(1.04)(1.15) = 865 ft^2 = 81 m^2.

From Figure 4.5.20, entering at A'_o, pick off the following combinations of shell inside diameter and tube length:

Shell Inside Diameter		Effective Tube Length		L/D_s
in.	mm	ft	m	
27	686	6.6	2.0	2.9
25	635	7.5	2.3	3.6
$23^1/_4$	591	9.2	2.8	4.7
$21^1/_4$	540	10.8	3.3	6.1
$19^1/_4$	489	13.1	4.0	8.2
$17^1/_4$	438	16.7	5.1	11.6

Any of these combinations would supply the desired area; the $21^1/_4$ in. (540 mm) and $19^1/_4$ in. (489 mm) would appear to be likely choices.

References

American Society of Mechanical Engineers. 1995. *ASME Boiler and Pressure Vessel Code*, Section VIII. New editions published every 3 years. ASME, New York.

Gentry, C.C., Young, R.K., and Small, W.M. 1982. RODbaffle heat exchanger thermal-hydraulic predictive methods, in *Proceedings of the Seventh International Heat Transfer Conference*, Munich, Germany, 6, 197–202.

Hewitt, G.F., Shires, G.L., and Bott, T.R. 1994. *Process Heat Transfer*, CRC/Begell House, Boca Raton, FL.

Kern, D.Q. and Kraus, A.D. 1972. *Extended Surface Heat Transfer*, McGraw-Hill, New York.

Kral, D., Stehlik, P., Van der Ploeg, H.J., and Master, B.I., 1996. Helical baffles in shell and tube heat exchangers. Part I: Experimental verification, *Heat Transfer Eng.*, 17(1), 93–101.

Palen, J.W. and Taborek, J. 1969. Solution of shell side flow pressure drop and heat transfer by stream analysis method, *Chem. Eng. Prog. Symp. Ser. No. 92, Heat Transfer-Philadelphia*, 65, 53–63.

Saunders, E.A.D. 1988. *Heat Exchangers: Selection, Design, and Construction*, Longman Scientific & Technical/John Wiley & Sons, New York.

Schlünder, E.U., Ed. 1983. *Heat Exchanger Design Handbook*, Begell House, New York.

Singh, K.P. and Soler, A.I. 1984. *Mechanical Design of Heat Exchangers and Pressure Vessel Components*, Arcturus, Cherry Hill, NJ.

TEMA. 1988. *Standards*, 7th ed., Tubular Exchanger Manufacturers Association, Tarrytown, NY.

Yokell, S. 1990. *A Working Guide to Shell and Tube Heat Exchangers*, McGraw-Hill, New York.

4.6 Temperature and Heat Transfer Measurements

Robert J. Moffat

There are two different kinds of material to consider with respect to experimental methods: the unit operations of measurement (transducers and their environmental errors) and the strategy of experimentation. This section deals only with the unit operations: transducers, their calibrations, and corrections for environmental errors.

Temperature Measurement

An International Practical Temperature Scale (IPTS) has been defined in terms of a set of fixed points (melting points of pure substances) along with a method for interpolating between the fixed points. The IPTS agrees with the thermodynamic temperature scale within a few degrees Kelvin over most of its range. The IPTS is the basis for all commerce and science, and all calibrations are made with respect to the IPTS temperature. The scale is revised periodically.

Accurate calibrations are not enough to ensure accurate data, however. If a sensor has been installed to measure a gas temperature or a surface temperature, any difference between the sensor temperature and the measurement objective due to heat transfer with the environment of the sensor is an "error." In most temperature-measuring applications, the environmental errors are far larger than the calibration tolerance on the sensor and must be dealt with just as carefully as the calibration.

Thermocouples

Any pair of thermoelectrically dissimilar materials can be used as a thermocouple. The pair need only be joined together at one end and connected to a voltage-measuring instrument at the other to form a usable system. A thermocouple develops its signal in response to the temperature difference from one end of the pair to the other. The temperature at one end, known as the *reference junction* end, must be known accurately before the temperature at the other end can be deduced from the voltage.

Thermocouples are the most commonly used electrical output sensors for temperature measurement. With different materials for different ranges, thermocouples have been used from cryogenic temperatures (a few Kelvin) to over 3000 K. In the moderate temperature range, ambient to 1200°C, manufacturer's quoted calibration accuracy can be as good as ±3/8% of reading (referred to 0°C) for precision-grade base metal thermocouples. Broader tolerances apply at very high temperature and very low temperatures. Thermocouple signals are DC voltages in the range from a few microvolts to a few tens of microvolts per degree C. Because of their low signal levels, thermocouple circuits must be protected from ground loops, galvanic effects, and from pickup due to electrostatic or electromagnetic interactions with their surroundings. Thermocouples are low-impedance devices. Multiple channels of thermocouples can be fed to a single voltage reader using low-noise-level scanners or preamplifiers and electronic multiplexers.

The alloys most frequently used for temperature measurement are listed in Table 4.6.1. These alloys have been developed, over the years, for the linearity, stability, and reproducibility of their EMF vs. temperature characteristics and for their high-temperature capability.

Calibration data for thermocouples are periodically reviewed by the National Institutes of Science and Technology based on the then-current IPTS. Values in Table 4.6.1 illustrate the approximate levels which can be expected, and are from the National Bureau of Standards Monograph 125. Maximum temperatures listed in this table are estimates consistent with a reasonable service lifetime. Allowable atmosphere refers to the composition in contact with the thermoelements themselves. Accuracy estimates are provided for two levels of precision: standard grade and precision grade where these data are available.

Noble metal and refractory metal thermocouples are often used with substitute lead wires, as a cost-saving measure. These lead wires, described in Table 4.6.2 are cheaper and easier to handle than the high temperature thermocouples. They have the same temperature–EMF characteristics as their primary thermoelements, but only over the range of temperatures the lead wires will usually encounter (up to a few hundred degrees C). Except for the substitute alloys, thermocouple extension wires have the same

TABLE 4.6.1 Application Characteristics of Some Common Thermocouple Alloys

Max T °F	Max T °C	Allowable Atmos. (Hot)	Material Names	ANSI Type[a]	Color Code	Output mV/100°F	Accuracy, %	
							Standard[a]	Precision[a]
5072	2800	Inert, H_2, vacuum	Tungsten/tungsten 26% rhenium	—	—	0.86	—	—
5000	2760	Inert, H_2, vacuum	Tungsten 5% rhenium/tungsten 26% rhenium	—	—	0.76	—	—
4000	2210	Inert, H_2	Tungsten 3% rhenium/tungsten 35% rhenium	—	—	0.74	—	—
3720	1800	Oxidizing[b]	Platinum 30% rhodium/platinum 6% rhodium	B	—	0.43	1/2	1/4
2900	1600	Oxidizing[b]	Platinum 13% rhodium/platinum	R	—	0.64	1/4	1/4
2800	1540	Oxidizing[b]	Platinum 10% rhodium/platinum	S	—	0.57	1/4	1/4
2372	1300	Oxidizing[b,c]	Platinel II (5355)/Platinel II (7674)	—	—	2.20	5/8	—
2300	1260	Oxidizing	Chromel/Alumel,[d] Tophel/Nial,[e] Advance T1/T2,[f] Thermo-Kanathal P/N[g]	K	Yellow red	2.20	4°F, or 3/4%	2°F, or 3/8%
1800	980	Reducing[a]	Chromel/constantan	E	Purple red	4.20	1/2	3/8
1600	875	Reducing	Iron/constantan	J	White red	3.00	4°F, or 3/4%	2°F, or 3/8%
750	400	Reducing	Copper/constantan	T	Blue red	2.50	3/4	3/8

[a] Per ANSI C96.1 Standard.
[b] Avoid contact with carbon, hydrogen, metallic vapors, silica, reducing atmosphere.
[c] @ Engelhard Corp.
[d] @ Hoskins Mfg. Co.
[e] Wilber B. Driver Co.
[f] Driver-Harris Co.
[g] The Kanthal Corp.

TABLE 4.6.2 Substitute Material Extension Wires for Thermocouples

Thermocouple Material	Thermocouple Type[a]	Extension Wire, Type[a]	Color for (+) Wire	Color for (–) Wire	Overall Color
Tungsten/tungsten 26% rhenium	—	Alloys 200/226[b]	—	—	—
Tungsten 5% rhenium/tungsten 26% rhenium	—	Alloys (405/426)[b]	White	Red	Red[b]
Tungsten 3% rhenium/tungsten 25% rhenium	—	Alloys (203/225)[b]	White/yellow	White/red	Yellow/red[b]
Platinum/platinum rhodium	S, R	SX, SR	Black	Red	Green
Platinel II-5355/Platinel II-7674	—	P2X[d]	Yellow	Red	Black[d]
Chromel/Alumel, Tophel/Nial, Advance, Thermokanthal[c]	K	KX	Yellow	Red	Yellow
Chromel/constantan	E	EX	Purple	Red	Purple
Iron/constantan	J	JX	White	Red	Black
Copper/constantan	T	TX	Blue	Red	Blue

[a] ANSI, except where noted otherwise.
[b] Designations affixed by Hoskins Mfg. Co.
[c] Registered trade mark names.
[d] Engelhard Mfg. Co.

composition as thermocouple wires, differing only in the type of insulation and the accuracy of calibration, which is not held as closely for extension wire as for thermocouple-grade wire.

Any instrument capable of reading low DC voltages (on the order of millivolts) with 5 to 10 μV resolution will suffice for temperature measurements. *Galvanometric measuring instruments* can be used, but, since they draw current, the voltage available at the terminals of the instrument depends not only on the voltage output of the thermocouple loop but also on the resistance of the instrument and the loop together. Such instruments are normally marked to indicate the external resistance for which they have been calibrated. *Potentiometric instruments*, either manually balanced or automatically balanced, draw no current when in balance, hence can be used with thermocouple loops of any arbitrary resistance without error. High-input impedance *voltmeters* draw very low currents and, except for very high resistance circuits, are not affected by the loop resistance.

Thermocouple Theory. Equation (4.6.1) is the general form describing the EMF generated in a two-wire thermocouple (Moffat, 1962). The same form can be derived from either the free-electron theory of metals or from thermodynamic arguments alone: the output of a thermocouple can be described as the sum of a set of terms, one arising in each wire in the circuit.

The junctions do not generate the EMF: they are merely electrical connections between the wires. For a two-wire circuit,

$$\text{EMF} = \int_0^L \varepsilon_1 \frac{dT}{dx} dx + \int_L^0 \varepsilon_2 \frac{dT}{dx} dx \tag{4.6.1}$$

where

ε_1 and ε_2 = the total thermoelectric power of materials 1 and 2, respectively, mV/C. The value of ε is equal to the sum of the Thomson coefficient and the temperature derivative of the Peltier coefficient for the material.
T = temperature, C
x = distance along the wire, m
L = length of the wire, m

This form for expressing the output of a two-wire circuit applies regardless of whether the wires are uniform in composition or not. If a circuit contained four wires (two thermocouple wires and two extension wires), then Equation (4.6.1) would be written with four terms, one for each length of wire.

When the wire is uniform in composition and both wires begin at (T_o) and both end at (T_L) the two terms can be collected into one integral:

$$\text{EMF} = \int_{T_0}^{T_L} \left(\varepsilon_1 - \varepsilon_2\right) dT \tag{4.6.2}$$

The EMF–temperature (E–T) tables produced by NIST and others are "solutions" to Equation (4.6.2) and can be used only when the following three conditions are met:

1. The thermoelectric power, ε, is not a function of position; i.e., the wires are homogeneous;
2. There are only two wires in the circuit;
3. Each wire begins at T_o and ends at T_L

When the circuit consists entirely of pairs of materials, Equation 4.6.2 can be used directly as the basis for understanding the source of the EMF. As an example, consider the three-pair system shown in Figure 4.6.1. For that circuit, Equation (4.6.2) would have three terms: one for each pair. The total EMF generated by the circuit would be the sum of the EMFs generated in the thermocouple pair and in the extension wire pair. The pair of copper wires would not contribute to the net EMF, assuming the two copper wires were perfectly matched. The EMF contributed by each pair would be proportional to the temperature difference from end to end of that pair, as shown in Equation (4.6.3) and (4.6.4).

$$\text{EMF} = \int_{T_1}^{T_2} \left(\varepsilon_{\text{cu}} - \varepsilon_{\text{cu}}\right) dT + \int_{T_2}^{T_3} \left(\varepsilon_+ - \varepsilon_-\right)_{\text{LEADS}} dT + \int_{T_3}^{T_4} \left(\varepsilon_+ - \varepsilon_-\right)_{\text{TC}} dT \tag{4.6.3}$$

$$\text{EMF} = 0 + (T3 - T2)\left(\varepsilon_+ - \varepsilon_-\right)_{\text{LEADS}} + (T4 - T3)\left(\varepsilon_+ - \varepsilon_-\right)_{\text{TC}} \tag{4.6.4}$$

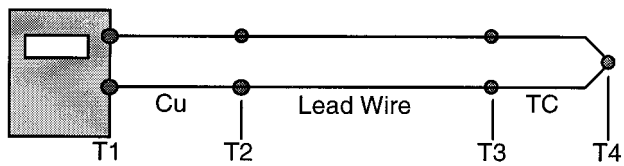


FIGURE 4.6.1 A three-pair circuit.

Most thermocouple circuits consist only of pairs of wires and can be understood in terms of these two equations, but some require a more detailed treatment. A graphical method of analysis is available, based on Equation (4.6.1).

The temperature–EMF calibrations of the more common materials are shown in Figure 4.6.2 derived from NBS Monograph 125 and other sources. This figure provides the input data for a simple graphical technique for describing the EMF generation in a circuit. Each curve in Figure 4.6.2 represents the output which would be derived from a thermocouple made of material X used with platinum when the cold end is held at 0°C and the hot end is held at *T*.

Those elements commonly used as "first names" for thermocouple pairs, i.e., Chromel (Chromel-Alumel), iron (-constantan), copper (-constantan), have positive slopes in Figure 4.6.2.

The simplest thermocouple circuit for temperature measurement consists of two wires joined together at one end (forming the "measuring junction") with their other ends connected directly to a measuring instrument, as shown in the upper portion of Figure 4.6.3. The EMF generation in this circuit is graphically represented in the lower portion, an E–T diagram, using the data in Figure 4.6.2. The E–T

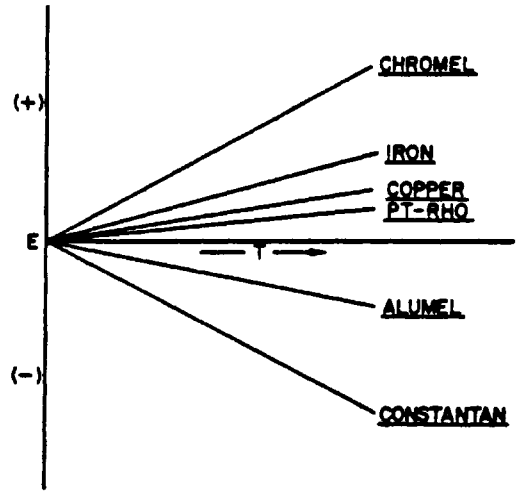


FIGURE 4.6.2 E–T calibrations for several common thermocouple materials.

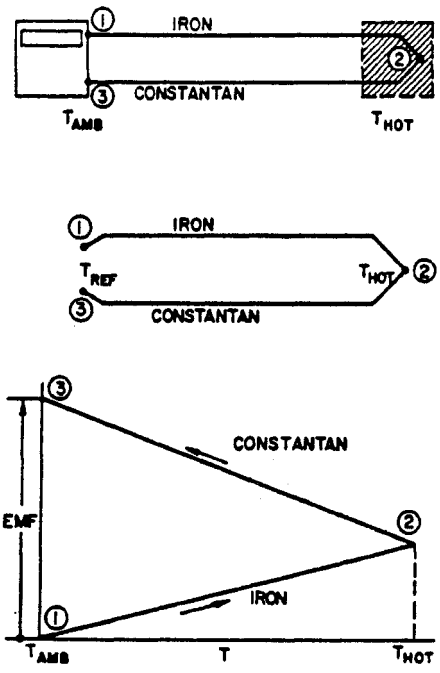


FIGURE 4.6.3 E–T diagram of a thermocouple using an ambient reference.

diagram is used for examining the EMF generated in the circuit, to be certain that it arises only from the desired thermocouple materials, and that all segments of the circuit are properly connected. The E–T diagram is not used for evaluating the output — that is done using the tables after the circuit has been shown to be correctly wired.

To construct an E–T diagram, first sketch the physical system considered and assign a number to each "point of interest" on that sketch and a temperature. On E–T coordinates, locate point 1 at 0/mV and at its assigned temperature. Then start a line from point 1, moving toward the temperature of point 2, and copying the shape of the calibration curve of the iron wire (see Figure 4.6.2). From 2 to 3, use the constantan calibration curve. The difference in elevation of points 1 and 3 describes the net EMF

generated in the circuit between points 1 and 3, and describes the polarity. When point 3 lies physically above point 1 in the E–T diagram, it is, by convention, electrically negative with respect to point 1.

The simple triangular shape shown in Figure 4.6.3 identifies a proper circuit. Any thermocouple circuit whose E–T diagram is equivalent to that is appropriate for temperature measurement and its EMF may be interpreted using the conventional tables. Any circuit whose E–T diagram is not equivalent to the pattern circuit should be rewired.

Thermocouples generate their signal in response to the temperature difference between the two ends of the loop. For accurate measurements, the temperature of the "reference junction" must be known. Laboratory users often use an ice bath made from a good-quality Dewar flask or vacuum-insulated bottle of at least 1 pt capacity, as shown in Figure 4.6.4. The flask should be filled with finely crushed ice and then flooded with water to fill the interstices between the ice particles. The reference thermocouple is inserted into a thin-walled glass tube containing a small amount of silicone oil and submerged six or eight diameters into the ice pack. The oil assures good thermal contact between the thermocouple junction and the ice/water mixture. The tube should be sealed at the top to prevent atmospheric moisture from condensing inside it, which would cause corrosion when using iron-constantan thermocouples. Figure 4.6.5 shows in iron-constantan thermocouple circuit with an ice bath. The individual thermocouple wires are connected to copper wires in the ice bath, and the two copper wires taken to the voltmeter. The lower portion of this figure shows the E–T diagram for this circuit, and proves that the output of this circuit is entirely due to the temperature difference from end to end of the iron-constantan loop: the two copper wires do not contribute to the output.

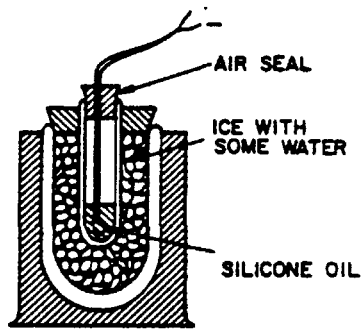


FIGURE 4.6.4 Characteristics of a good ice bath.

Calibration. Thermocouple calibrations are provided by the wire manufacturers to tolerances agreed upon industry-wide, as summarized in Table 4.6.1. These tolerances include two components: the uncertainty in the average slope of the calibration curve of the wire, and the effects of local inhomogeneities in the wire. It is difficult to improve the accuracy of a thermocouple by calibrating it. For a truly significant calibration, the thermocouple would have to be exposed to the same temperature during calibration, at every point along it, that it would encounter in service. In an oven calibration, most of the signal is generated in the material at the mouth of the oven, as could be recognized by considering the temperature gradient distribution along the wire. The material inside the oven contributes little or nothing to the signal.

Thermistors

Thermistors are electrical resistance temperature transducers whose resistance varies inversely, and exponentially, with temperature. The resistance of a 5000 Ω thermistor may go down by 200 Ω for each degree C increase in temperature in the vicinity of the initial temperature. Interrogated by a 1.0 mA current source, this yields a signal of 200 mV/°C. As a consequence of this large signal, thermistors are frequently used in systems where high sensitivity is required. It is not uncommon to find thermistor data logged to the nearest 0.001°C. This does not mean that the data are accurate to 0.001°C, simply

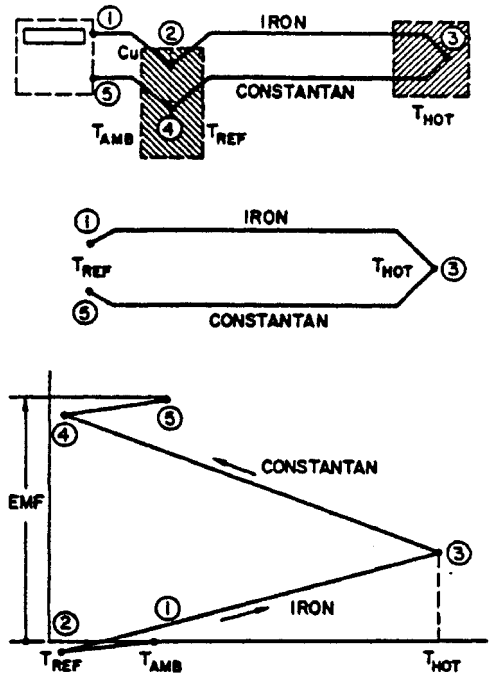


FIGURE 4.6.5 A thermocouple circuit using an ice bath reference, and its E–T diagram.

that the data are readable to that precision. A thermistor probe is sensitive to the same environmental errors which afflict any immersion sensor: its accuracy depends on its environment.

Thermistor probes can be used between –183°C (the oxygen point) and +327°C (the lead point) but most applications are between –80 and +150°C. The sensitivity of a thermistor (i.e., the percent change in resistance per degree C change in thermistor temperature) varies markedly with temperature, being highest at cryogenic temperatures.

Thermistor probes range in size from 0.25-mm spherical beads (glass covered) to 6-mm-diameter steel-jacketed cylinders. Lead wires are proportionately sized. Disks and pad-mounted sensors are available in a wide range of shapes, usually representing a custom design "gone commercial." Aside from the unmounted spherical probes and the cylindrical probes, there is nothing standard about the probe shapes.

Calibration. Thermistor probes vary in resistance from a few hundred ohms to megohms. Probe resistance is frequently quoted at 25°C, with no power dissipation in the thermistor. The commercial range is from about 2000 to 30,000 Ω. Representative values of the sensitivity coefficient (% change in resistance per degree C) is given in Table 4.6.3 and resistance values themselves, in Table 4.6.4.

TABLE 4.6.3 Thermistor Temperature Coefficient Variations with Temperature

Temp. °C	Condition	ΔR/R, %
–183	Liquid oxygen	–61.8
–80	Dry ice	–13.4
–40	Frozen mercury	–9.2
0	Ice point	–6.7
25	Room temperature	–5.2
100	Boiling water	–3.6
327	Melting lead	–1.4

TABLE 4.6.4 Thermistor Resistance Variation with Temperature

Temp., °C	Res., Ω	Temp., °C	Res., Ω
–80	1.66 M	0	7355
–40	75.79 K	25	2252
–30	39.86 K	100	152.8
–20	21.87 K	120	87.7
–10	12.46 K	150	41.9

Proprietary probes are available which "linearize" thermistors by placing them in combination with other resistors to form a circuit whose overall resistance varies linearly with temperature over some range. These compound probes can be summed, differenced, and averaged as can any linear sensor. Modern manufacturing practices allow matched sets to be made, interchangeable within +/–0.1°C.

Thermal Characteristics. Thermistor probes are generally interrogated using a low current, either AC or DC. A level of about 10 μA would be typical. With a probe resistance of 10 K Ω, 0.01 W must be dissipated into its surrounding material. This current results in the probe running slightly above the temperature of the medium into which it is installed: the "self-heating" effect. Since thermistors are often used where very small changes in temperature are important, even small amounts of self-heating may be important.

The self-heating response is discussed in terms of the "dissipation constant" of the probe, in milliwatts per degree C. The dissipation constant depends on the thermal resistance between the thermistor and its surroundings. For fluid-sensing probes, the self-heating varies with velocity and thermal conductivity, while for solid immersion probes, it varies with the method of attachment and type of substrate.

Dissipation constants for representative probes are given in Table 4.6.5. The self-heating effect must be considered in calibration as well as in use.

The transient response of a thermistor is more complex than that of a thermocouple and, size for size, they are not as well suited to transient measurements.

TABLE 4.6.5 Representative Thermal Dissipation Constants for Two Thermistor Probe Designs

Environment	1.0-cm Disk	5.0-cm Cylinder
Still air	8 mW/C	1 mW/C
Still oil	55	—
Still water	—	3.5
Oil at 1 m/sec	250	—

Thermistor probes are sold with calibration tables of resistance vs. temperature at some specified accuracy, on the order of +/–0.1 or 0.2 K, depending on the grade of probe purchased. These tables are typically in increments of 1 K. For computer interpretation, they should be fit to the Steinhart-Hart form[2] and the coefficients determined for least error.

$$\frac{1}{T} = A_0 + A_1 \ln(R) + A_3 \ln\left(R^3\right) \tag{4.6.5}$$

Resistance Temperature Detectors

The terms *resistance temperature detector* (RTD) and *resistance thermometer* are used interchangeably to describe temperature sensors containing either a fine wire or a thin film metallic element whose resistance increases with temperature. In use, a small current (AC or DC) is passed through the element, and its resistance measured. The temperature of the element is then deduced from the measured resistance using a calibration equation or table lookup.

RTDs are used both for standards and calibration laboratories and for field service. Field-service probes are generally encased in stainless steel protective tubes with either wire or film elements bonded to sturdy support structures. They are made to take considerable physical abuse. Laboratory standard-grade probes are often enclosed in quartz tubes, with the resistance wire mounted in a strain-free manner on a delicate mandrel.

High-quality resistance thermometers have been used as defining instruments over part of the range of the IPTS. Because of this association with high-precision thermometry, resistance thermometers in general have acquired a reputation for high precision and stability. Commercial probes, however, are far different in design from the standards-grade probes, and their stability and precision depend on their design and manufacture.

RTDs are often recommended for single-point measurements in steady-state service at temperatures below 1000°C where longtime stability and traceable accuracy are required and where reasonably good heat transfer conditions exist between the probe and its environment.

They are not recommended for use in still air, or in low-conductivity environments. RTDs self-heat, which causes an error when the probes are used in a situation with poor heat transfer. They are not recommended for transient service or dynamic temperature measurements unless specifically designed for such service. The probes tend to have complex transient characteristics and are not amenable to simple time-constant compensation.

Physical Characteristics. The physical characteristics of any given resistance thermometer represent a compromise between two opposing sets of requirements. For accuracy, repeatability, and speed of response, a delicate, low-mass sensing element is desired, supported in a strain-free manner in good thermal contact with its surroundings. For durability, a rugged sensor is indicated, mounted firmly to a sturdy structure inside a robust, sealed protection tube.

Both the short-term calibration (resistance vs. specimen temperature) and the long-term stability (drift) are directly affected by the mechanical configuration of the probe. The electrical resistance of the sensing element is a function of its temperature and state of mechanical strain (Figure 4.6.6).

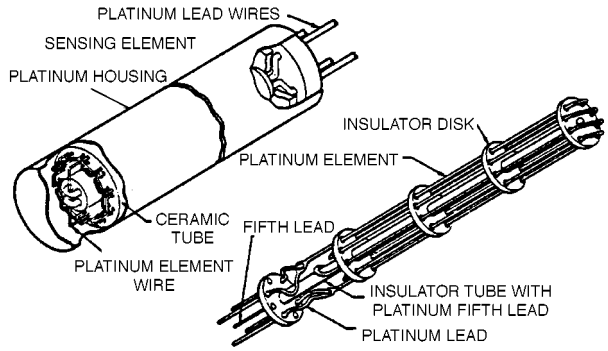


FIGURE 4.6.6 Slack-wire platinum resistance thermometer.

The sensing elements used in field-service RTD probes range from thin metallic films deposited on rectangular ceramic wafers ($0.5 \times 1.0 \times 2.0$ mm) with pigtail leads (0.25 mm diameter and 2.5 cm long) to glass-encapsulated, wire-wound mandrels (4 mm in diameter and 2.0 cm long), again with pigtail leads. Bonding the sensor to its support provides good mechanical protection to the element, but subjects the element to strain due to thermal expansion. As long as this process is repeatable, the calibration is stable.

Electrical Characteristics. RTDs are available commercially with resistances from 20 to 20,000 Ω with 100 Ω being common. Bifilar windings are frequently used in wire-wound elements, to reduce the

electrical noise pickup. This is more important in the quartz-jacketed probes than in those with stainless steel protection tubes. Twisted pair lead wires are recommended.

Thermal Characteristics. Figure 4.6.7 shows a simplified cross section of a typical resistance thermometer and a thermal circuit which can be used to discuss its behavior. In forming such a thermal circuit model, each element of the probe is described by its resistive and capacitive attributes following conventional heat transfer practice. The principal components are

- The external thermal resistance per unit length;
- The thermal capacitance of the protective tube per unit length, C_T;
- The radial internal thermal resistance between the sensor and the protective tube, R_{int};
- The capacitance of the sensor element and its support, C_{sensor};
- The axial internal thermal resistance of the stem, per unit length, R_T.

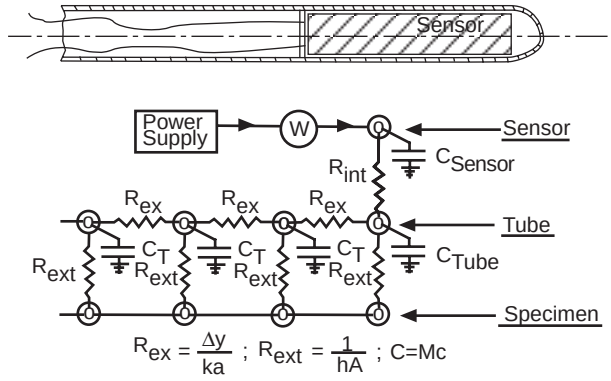


FIGURE 4.6.7 Thermal circuit representation of a typical resistance thermometer.

This circuit can be used to predict the temperature distribution within the probe both at steady state and during transients and can be refined, if needed, by subdividing the resistance and capacitance entities.

Steady-State Self-Heating. Interrogating an RTD by passing a current through it dissipates power in the element, shown in Figure 4.6.7 as W, which goes off as heat transfer through the internal and external resistances. This self-heating causes the sensing element to stabilize at a temperature higher than its surroundings and constitutes an "error" if the intent is to measure the surrounding temperature. The amount of the self-heating error depends on three factors:

- The amount of power dissipated in the element,
- The internal thermal resistance of the probe, as a consequence of its design, and
- The external thermal resistance between the surface of the probe and the surrounding material.

The self-heating temperature rise is given by Equation (4.6.6):

$$T_{sens} - T_{surr} = W\left(R_{int} + R_{ext}\right) \tag{4.6.6}$$

The internal thermal resistance of a probe, R_{int}, measured in degree C per watt, describes the temperature rise of the sensing element above the surface temperature of the probe, per unit of power dissipated. The internal thermal resistance can be deduced from measurements of the sensor temperature at several different current levels when the probe is maintained in a well-stirred ice bath, where the external thermal resistance is very low. The slope of the apparent temperature vs. power dissipated line, °C/W, is the internal thermal resistance. When an RTD is used in a gas or liquid, the external resistance between the

probe and its surroundings must be estimated from standard heat transfer data. The external resistance is 1/(hA), °C/W.

A typical cylindrical probe exposed to still air will display self-heating errors on the order of 0.1 to 1.0°C per mW (commercial probes of 1.5 to 5 mm in diameter). At 1 m/sec air velocity, the self-heating error is reduced to between 0.03 and 0.3°C. In water at 1 m/sec velocity, the self-heating effect would be reduced by a factor of four or five compared to the values in moving air, depending on the relative importance of the internal and the external thermal resistances.

Calibration and Drift. The relationship between resistance and temperature must be determined for each probe or acquired from the manufacturer. Generally speaking, the reported values will require interpolation.

The resistance–temperature characteristic of a probe may drift (i.e., change with time) while the probe is in service. Manufacturers of laboratory-grade probes will specify the expected drift rate, usually in terms of the expected error in temperature over an interval of time. Two sample specifications are "0.01 C per 100 hours" for a low-resistance, high-temperature probe (0.22 Ω at 0°C, 1100°C maximum service temperature) and "0.01 C per year" for a moderate-resistance, moderate-temperature probe (25.5 Ω at 0°C, 250°C maximum service temperature). Drift of the resistance-temperature relationship takes place more rapidly at high temperatures.

Radiation Devices

Surface temperatures and gas temperatures can be deduced from radiation measurements. Surface-temperature measurements are based on the emitted infrared energy, while gas-temperature measurements use specific emission lines from the gas itself or from a tracer inserted into the gas.

Commercial surface-temperature measurement systems (single-point) are available, at low cost, which can measure temperature to +/–1% of reading, above 38°C, if the emissivity of the surface is known. The device referenced requires a spot size of 1.25 cm diameter, viewed from 75 cm. Spectroscopic gas-temperature measurements can be accurate to +/–3 or 4% of reading, but require a significant investment in effort as well as equipment (on the order of 1 to 2 years and $100,000 to $200,000). Several techniques based on Raman scattering have been used in combustion systems. Planar-lasar-induced fluorescence has shown considerable promise as one of the newer methods.

Infrared emission from a surface is described by two laws: the Stefan Boltzmann law describing the total emitted radiation as a function of temperature, and Planck's law describing its distribution as a function of temperature. These laws form the basis for all radiation-based surface-temperature detectors.

Early radiometers focused the total infrared energy on a thermopile bolometer and used the temperature rise across its calibrated heat loss path to measure the incident energy flux. Solid-state photon detectors have replaced thermopile bolometers as the detector of choice. Such a detector will respond to any photon having energy above a certain level (specific to the detector). Since the energy of a photon is inversely proportional to its wavelength, detectors respond to all wavelengths below some value. Modern detectors use band-pass filters to limit the wavelength band of photons admitted to the detector and rely on Planck's law to infer the temperature from the energy flux:

$$E_{b,\lambda} = \frac{C_1 \lambda^{-5}}{e^{C_2/\lambda T} - 1} \tag{4.6.7}$$

where $E_{b,\lambda}$ = radiated power at the wavelength λ, W/m²
T = temperature, K
C_1 = 3.743×10^8, Wμm⁴/m²
C_2 = 1.4387×10^4, μmK

Commercial radiation temperature detectors use different wave bands for different temperature ranges, with different detectors for each band. The emissivity of the surface must be known, as a function of temperature, in the wavelength band used by the detector.

Radiation detectors are vulnerable to interference from four sources: low signal-to-noise ratio at low temperatures (below a few hundred degrees C); radiation from the surroundings reflecting into the detector (also usually more important at low temperatures); low spatial resolution (also more evident at low temperatures); uncertainty in the emissivity of the surface (at all temperatures); and absorption of radiation into water vapor and CO_2 in the line of sight (at any temperature).

A fiber-optic blackbody temperature detector system is offered by the Luxtron Corporation for standards room and field service above 300°C. The unit consists of a blackbody capsule fiber-optically coupled to a filtered, band-limited photon detector. Accuracy of 0.01 to 0.1°C is claimed, depending on temperature level.

A fluoroptic temperature-measuring system is also offered by the same company, for use only at lower temperatures (–200 to +450°C). This system uses an ultraviolet-stimulated phosphor on the end of an optical fiber as its sensor. The fluorescent signal from the phosphor decays with time, and its "time constant" is a function of temperature. Accuracy of +/–0.5°C is claimed for measurements within +/–50°C of a calibration point, or +/–1°C within 100°C.

Temperature-Sensitive Paints, Crayons, and Badges

Temperature-sensitive paints, crayons, and badges are available from several suppliers (Omega Engineering, Inc., Stamford, CT, and others in Germany and Japan). Each undergoes an irreversible change (e.g., a change in color or a change from solid to liquid) at one specified temperature. With a range of paints, temperatures from ambient to about 1500°C can be covered. The accuracy generally quoted is about +/–1% of level, although melting standards are available to +/–0.5°C.

The phase-change materials melt at well-defined temperatures, yielding easily discernible evidence that their event temperature has been exceeded. When more than one phase-change paint is applied to the same specimen, there can be interference if the melt from the low-melting paint touches the high-melting material. Color change materials do not interfere, but are more difficult to interpret. The calibration of high-temperature paints (both phase change and color change) may shift when they are used on heavily oxidized materials, due to alloying of the oxide with the paint. Recommended practice is to calibrate the paints on specimens of the application material. The event temperature which will cause transformation depends on the time at temperature: short exposure to a high temperature often has the same effect as long exposure to a lower temperature.

The paints and crayons are nonmetallic and, therefore, tend to have higher emissivities for thermal radiation than metals. They should be used only over small areas of metallic surfaces, compared with the metal thickness, or else their different emissivities may lead to a shift in the operating temperature of the parts.

The principal disadvantages of the paints and crayons are that they require visual interpretation, which can be highly subjective, and they are one-shot, irreversible indicators which respond to the highest temperature encountered during the test cycle. They cannot record whether the peak was reached during normal operation or during soak-back.

Liquid crystals can be divided into three groups, depending on their molecular arrangements: (1) smectic, (2) nematic, and (3) cholesteric. Most of the temperature-sensitive liquid crystals now in use are cholesteric: made from esters of cholesterol. Their molecules are arranged in planar layers of molecules with their long axes parallel and in the plane of the layer. The molecules in each layer are rotated with respect to those in its neighboring layers by about 15 min of arc in a continuous, helical pattern along an axis normal to the layers.

The colors reflected from cholesteric liquid crystals are thought to be due to Bragg diffraction from the aligned layers. The "wrap angle" between adjacent layers increases with temperature; hence, the color of the liquid crystal shifts toward short wavelengths (toward blue) as the temperature is raised. The color can also be affected by electric fields, magnetic fields, pressure, shear stress, and some chemical vapors.

Warm cholesterics are colorless liquids and they pass through a series of bright colors as they are heated through their "color-play" temperature band. The first color to appear is a deep red, followed by

yellow, green, blue, and violet. Further heating yields a colorless liquid again. This cycle is reversible and repeatable, and the color–temperature relationship can be calibrated.

Liquid crystals selectively reflect only a small fraction of the incident light; hence, to enhance the brightness of the color image, they must be backed up with black paint or a nonreflecting surface.

A typical calibration is shown in Figure 4.6.8 for liquid crystals painted over black paint on an aluminum calibration strip. The upper part of Figure 4.6.8 describes the color variation, while the lower part shows the imposed linear temperature distribution. The hot end is blue, the cold end is red. Color-play intervals range from 0.5 to 10.0°C. Liquid crystals whose color-play interval is on the order of 0.5 to 2.0°C are often referred to as *narrow-band* materials, while those whose interval extends to 5.0 to 10°C are called *wide band*. Narrow-band images are easy to interpret by eye. Wide-band images show only subtle variations of color for small changes in temperature, and accurate work requires digital image handling or multiple images taken with different filters.

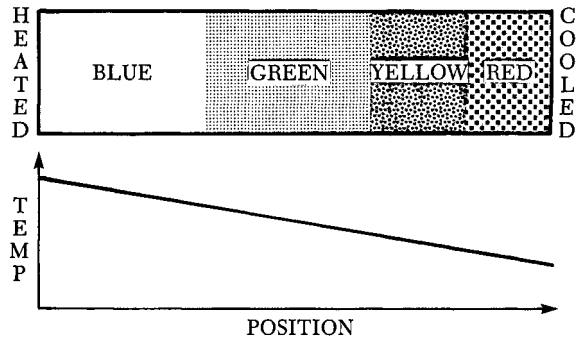


FIGURE 4.6.8 Schematic representation of a calibration strip.

Several different narrow-band liquid crystals can be mixed together to make a single, multi-event paint covering a wide range of temperatures, provided their color-play intervals do not overlap. Such a mixture yields a set of color-play bands, one for each component.

Calibration. Liquid crystals are sold by event temperature and color-play bandwidth, with a nominal accuracy of +/–1°C on the event temperature. In many applications, especially if the image is to be visually interpreted, no further calibration is needed.

The accuracy attainable with a liquid crystal is related to the width of the color-play interval. With narrow-band material (a color-play interval of about 1.0°C), visual interpretation can be done with an uncertainty of 0.25 to 0.5°C. With digital image interpretation, spectrally controlled lighting and appropriate corrections for reflected light interference, the uncertainty can be held below 0.25°C.

Early users reported that the perceived color of a liquid crystal depended on both the lighting angle and the viewing angle. This dependence can be eliminated by using a light source along the line of sight (coaxial viewing and illumination).

Multiple-Event Paints. Several narrow-band paints can be mixed together to make a single paint with all the characteristics of each component, if their color-play intervals do not overlap. Each component retains its original calibration and acts independently of the other components.

Figure 4.6.9 shows the image from a five-event paint used to map the adiabatic wall temperature isotherms around a heated block in mixed convection. The outermost isotherm is 30°C, and the events are spaced apart at 5°C intervals. Determination of the temperatures from a multiple-event image requires that the temperature be known at one point in the image.

Liquid Crystals in Water. Liquid crystals can be used to mark the temperature distribution in water and some other liquids by adding a small quantity of encapsulated liquid crystal material to the liquid and photographing the color distribution using planar lighting. Velocity and temperature distributions can be

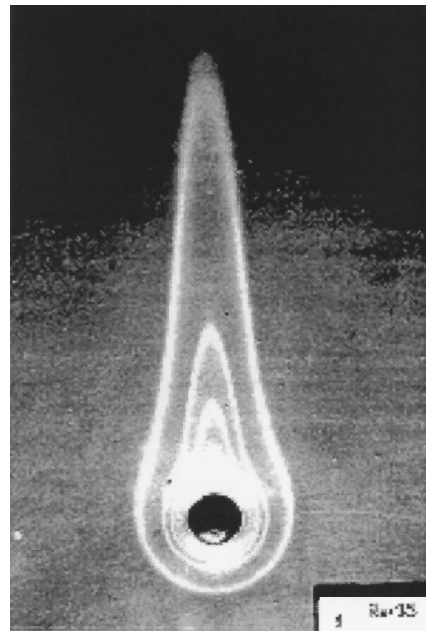


FIGURE 4.6.9 Multi-event liquid crystal used to visualize the isotherm pattern above a heated spot in mixed convection.

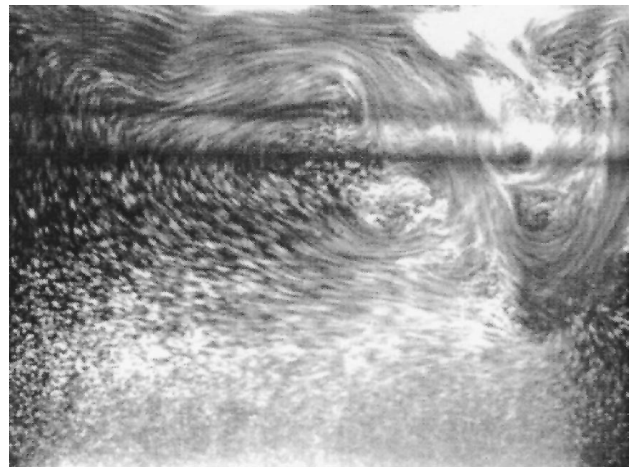


FIGURE 4.6.10 Liquid crystal visualization of the velocity and temperature distribution in a water-filled tank.

determined by photographing the liquid crystal particles using a known exposure time. The temperature is deduced from the particle color, and the velocity by the length of the streak its image forms. Figure 4.6.10 shows the velocity and temperature distributions in a shear-driven, water-filled cavity 30 sec after the impulsive start of belt motion. In this view, the belt is at the top of the image, and moved from left to right. The water was stably stratified initially, with the top being 4°C hotter than the bottom. This technique was demonstrated by Rhee et al. (1984) and has been used by several workers.

Image Processing. Several schemes have been proposed to remove the subjectivity from interpretation of liquid crystal images. Akino et al. (1989), and others, have processed RGB video images of narrow-band images using multiple filters to extract images of specified isochromes, related to temperatures through a calibration. Hollingsworth et al. (1989) processed RGB images of wide-band images using

chromaticity coordinates (hue, saturation, and intensity) and extracted temperature at each pixel, rather than along isochromes.

Heat Flux

Heat flux to or from a surface can be measured directly, using heat flux meters, or inferred from an overall energy balance, or inferred from temperature–time measurements at the surface or within the body. There are no primary standards for heat flux measurement.

Three general classes of heat flux meters are in common use: slug calorimeters, planar heat flux gauges (sometimes called Schmidt–Boelter gauges), and circular foil gauges (sometimes called Gardon gauges). Sensitivities range from microvolts per kW/m^2 to millivolts per W/m^2. Planar gauges can be used for radiant or convective heat loads. Circular foil gauges should be used only for radiant loads.

Slug Calorimeter

The slug calorimeter is an energy balance transducer consisting of a known mass of material instrumented so that its temperature can be measured. A simple version is shown in Figure 4.6.11. If losses are negligibly small and the mass and the specific heat are constant, the instantaneous heat flux is deduced from

$$\dot{q}''_{\text{in}} A = Mc \frac{\partial T}{\partial t} \tag{4.6.8}$$

where T = Average temperature of the slug, C
M = Mass of the slug, kg
c = Specific heat, J/kg·C
A = Face area, m^2
τ = Time

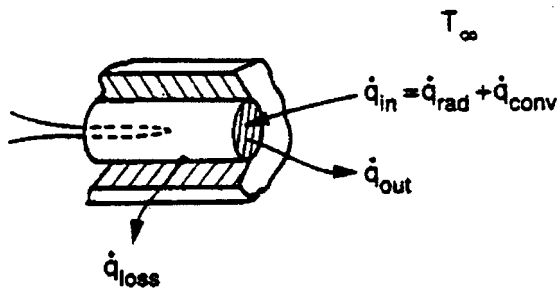


FIGURE 4.6.11 A simple slug calorimeter.

The variation of slug temperature with time is used to infer net heat transfer rate to the gauge. Slug calorimeters are used mainly when the heat flux, or the heat transfer coefficient, is expected to be relatively constant. They are of less value when the input flux changes arbitrarily because of the inaccuracies inherent in differentiating the signals.

Planar Heat Flux Gauge

Planar heat flux gauges use Fourier's law to deduce the heat flux from a steady-state measurement of the temperature difference across a thin sheet of thermally insulating material. The planar gauge geometry is shown in Figure 4.6.12. The working equation for a planar gauge is

FIGURE 4.6.12 A typical planar heat flux gauge.

$$\text{EMF} = N\varepsilon\,\Delta T = \frac{N\varepsilon t}{k}\dot{q}'' \tag{4.6.9}$$

where N = number of junction pairs,
e = thermoelectric power of the thermoelement, mV/C
t = thickness of the insulator, m
k = conductivity of the insulator, W/m·C
q'' = heat flux through the gauge, W/m^2

The figure shows one thermocouple junction on the top and one on the bottom surface of the insulator. Most gauges use multiple junctions. The thermoelements may be wire (down to 0.025 mm diameter) or thin films deposited on the insulator (10 to 20 Å). The assembly is usually sandwiched between two sheets of protective material to form an integral unit. Up to 150°C application temperature, these units are often made of Kapton, and provided with a contact adhesive. They may be as thin as 0.15 mm overall.

Gauges should not be removed and reinstalled without recalibration, as the act of removing them from the surface may delaminate the gauge, changing its thermal resistance, and therefore its calibration.

Circular Foil Gauges

A circular foil gauge consists of a thin circular disk of metal supported by its edge from a structure of constant and uniform temperature. The circular foil gauge is often called a Gardon gauge. A constantan foil is often used, with a copper support structure. Two copper wires complete the circuit: one attached to the center of the foil disk and one to the support structure. The copper–constantan thermocouple thus formed produces an EMF determined by the temperature difference from the center of the foil disk to its rim. That temperature difference is directly proportional to the average heat flux on the disk. A cross-sectional view of a circular foil gauge is shown in Figure 4.6.13.

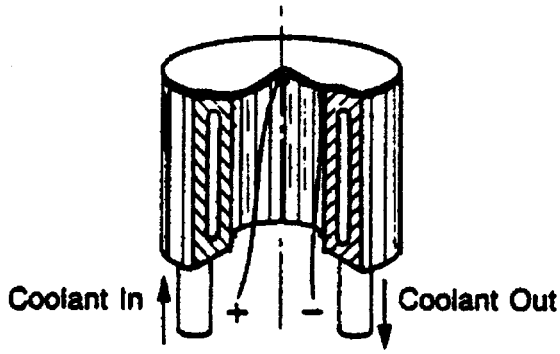


FIGURE 4.6.13 A water-cooled circular foil gauge (Gardon gauge).

The working equation for a circular foil gauge is

$$\text{EMF} = \varepsilon\frac{R^2}{4kt}\dot{q}'' \tag{4.6.10}$$

where ε = thermoelectric power, mV/C

R = radius of the disk, m
k = thermal conductivity of the disk, W/m·C
t = thickness of the disk, m
q'' = heat flux absorbed by the disk, W/m^2 (must be uniform)

The output signal is thus directly proportional to the heat flux on the disk. Cooling passages are frequently built into the support structure to maintain the edge of the disk (the heat sink for the foil disk) at constant temperature.

Calibration

Calibration of the Gardon-type heat flux meters is most easily done by comparison, using a radiation calibrator.

Planar gauges can be calibrated either by conduction or radiation, but the results will depend on the calibration method for some guages.

Sensor Environmental Errors

Temperature sensors generate signals in response to their own temperatures, but are usually installed to measure the temperature of some fluid or solid. There is heat transfer between the sensor and all of its surroundings, with the result that the sensor usually equilibrates at some temperature different from the fluid or solid it is installed in. This difference is considered an error in the measurement.

Similarly, heat flux gauges are generally installed so one can infer the heat flux which would have been there had the gauge not altered the system behavior. But heat flux gauges do disturb the system, and the heat flux at the gauge location, when the gauge is there, may be significantly different from that which would have been there without the gauge. This system disturbance effect must also be considered an error.

Steady-State Errors in Gas-Temperature Measurement

All immersion-type temperature sensors (thermocouples, resistance detectors, and thermistors) are subject to the same environmental errors, which are frequently larger than the calibration errors of the sensors. Large probes are usually affected more than small ones; hence, RTDs and thermistors (selected by investigators who wish to claim high accuracy for their data) are more vulnerable to environmental errors (due to their larger size and their self-heating errors). This aspect of accuracy is sometimes overlooked.

Sensor installations for gas-temperature measurements should be checked for all three of the usual steady-state environmental errors: velocity error, radiation error, and conduction error. The same equations apply to all sensors, with appropriate dimensions and constants.

$$\text{velocity error:} \quad E_v = (1-\alpha)\frac{V^2}{2g_c Jc_p} \tag{4.6.11}$$

$$\text{radiation error:} \quad E_r = \frac{\sigma\varepsilon}{h}\left(T_{\text{sens}}^4 - T_{\text{surr}}^4\right) \tag{4.6.12}$$

$$\text{conduction error:} \quad E_c = \frac{T_{\text{gas}} - T_{\text{mount}}}{\cosh\left[L\sqrt{\dfrac{hA_c}{kA_k}}\right]} \tag{4.6.13}$$

where E_v = velocity error, °
α = recovery factor, —

V = velocity, ft/sec
g_c = universal gravitational constant
J = Joules constant, ff/bf/Btu
c_p = specific heat, Btu/lbm, °F

and E_r = radiation error, °R
σ = Stefan-Boltzmann constant
ε = emissivity
h = heat transfer coefficient, Btu/secft^2, °F
T_{sens} = indicated temperature, °R
T_{surr} = surrounding temperature, °R
E_c = conduction error, °R
T_{gas} = gas temperature, °R
T_{mount} = mount temperature, °R
L = length of exposed junction, ft
h = heat transfer coefficient, Btu/secft^2, °F
A_c = heat transfer area, ft^2
k = thermal conductivity, Btu/secft
A_k = conduction area, ft^2

Velocity error depends upon the recovery factor, which varies with the Prandtl number of the fluid. The Prandtl numbers of most liquids are greater than 1; hence, the recovery factor α is greater than 1 and probes tend to read higher than the stagnation temperature in high-speed liquid flows. With thermistors and RTDs in liquids, the self-heating effect and the velocity error both tend to cause high readings. In gases, where the Prandtl number is less than 1, the two effects are of opposite sign and may partly cancel each other.

Radiation and conduction errors vary inversely with the heat transfer coefficient. Both tend to be larger for larger-diameter probes since, all other factors remaining the same, the heat transfer coefficient will be lower for a large-diameter probe. This results in larger radiation and conduction errors. In liquids, radiation error is not a problem, but velocity error and conduction error may both be significant. Conduction error becomes a problem in liquid-temperature measurements when thermowells are used. The depth of immersion of the well is frequently too short to eliminate conduction error.

Steady-State Errors in Solid and Surface-Temperature Measurements

When probes are used to measure solid temperature by inserting them into a hole in the specimen, they are subject to conduction errors proportional to their size and conductivity. A general rule of thumb is to keep the insertion depth at least 20 times the diameter (or wall thickness) of the probe. This assumes a close-fitting hole, backfilled with a material with higher thermal conductivity than air. For more-exact advice regarding a specific installation, a careful thermal circuit analysis of the installation should be developed, and its results used to guide the selection of diametrical clearance, backfill materials, and penetration depth.

A thermocouple attached to a hot surface surrounded by cooler fluid will exchange heat with the fluid by convection and with the surrounding solids by radiation. Heat lost from the thermocouple must be made up from the surface by conduction, which will result in a cold spot at the point of attachment.

Figure 4.6.14 shows the system disturbance error caused by a surface-attached thermocouple, as a fraction of the maximum possible error for the installation. If the surface is irradiated (e.g., by heating lamps), the irradiation will raise the surface temperature, but will also affect the system disturbance error. The effect on the system disturbance error caused by turning on the irradiation is similar to that of raising the temperature of the surrounding fluid to a new value, T_∞°,

where T_{ind} = indicated temperature from an otherwise error-free thermocouple, °C
T_s = undisturbed substrate temperature, °C

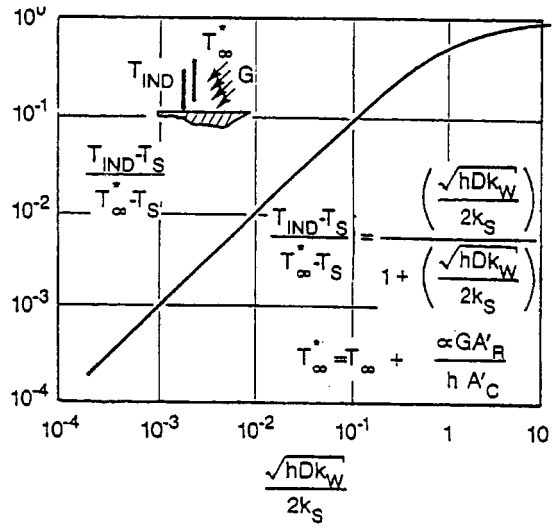


FIGURE 4.6.14 System disturbance errors caused by an attached thermocouple (worst case).

T°_∞ = effective fluid temperature, °C
h = heat transfer coefficient, TC to fluid, W/m, °C
D = outside diameter of TC, m
k_w = effective thermal conductivity of TC, W/m, °C
k_s = thermal conductivity of substrate, W/m, °C

The effective gas temperature is defined in terms of the incident irradiation and the heat transfer coefficient as

$$T^*_\infty = T_\infty + \frac{\alpha G A_R}{h A_C} T^*_\infty \tag{4.6.14}$$

where T_∞ = actual gas temperature, °C
α = absorptivity of the TC for thermal radiation
G = incident thermal radiation flux, W/m^2
A_R/A_c = ratio of irradiated surface to convective surface
h = heat transfer coefficient between the TC and the gas, W/m°C

Steady-State Errors in Heat Flux Gauges for Convective Heat Transfer

If the gauge is not flush with the surface, it may disturb the flow, and if it is not at the same temperature as the surface, it will disturb the heat transfer. Thus, the gauge may properly report the heat flux which is present when the gauge is present, but that may be significantly different from the heat flux which would have been there if the gauge had not been there.

For planar gauges, both effects are usually small. The thermal resistance of such a gauge is generally small, and they are thin enough to avoid disturbing most flows. Circular foil gauges pose a more serious problem, since they are often cooled significantly below the temperature of the surrounding surface. Dropping the wall temperature at the gauge location can significantly increase the local heat load in two ways: one due to the fact that, for a given value of h, a cold spot receives a higher heat load from the gas stream. The second effect arises because the value of the heat transfer coefficient itself depends on the local wall temperature distribution: a local cold spot under a hot gas flow will experience a higher heat transfer coefficient than would have existed had the surface been of uniform temperature.

Evaluating the Heat Transfer Coefficient

The heat transfer coefficient is a defined quantity, given by

$$h = \frac{\dot{q}''_{\text{conv}}}{\left(T_o - T_{\text{ref}}\right)} \tag{4.6.15}$$

where h = heat transfer coefficient, W/m^2, °C
$\dot{q}''_{\text{conv}}$ = convective heat flux, W/m^2
T_o = temperature of the considered surface, °C
T_{ref} = temperature used as reference for this definition, °C

Different reference temperatures are conventionally used for different situations:

- T_∞: The free-stream temperature. Used for isolated objects of uniform temperature in a uniform free stream, where an average value of h is desired which describes the overall heat transfer between the object and the flow. Also used in boundary layer heat transfer analyses where local values of h are needed to deal with locally varying conditions.
- T_m: The mixed mean fluid temperature. Used for internal flows where the intent of the calculation is to describe the changes in mixed mean fluid temperature (e.g., heat exchangers).
- $T_{\text{adiabatic}}$: The adiabatic surface temperature. Used for isolated objects or small regions of uniform temperature in either internal or external flows, where the overall thermal boundary conditions are neither uniform heat flux nor uniform temperature.

For a given data set, the value of the heat transfer coefficient will depend on the reference temperature chosen, and h should be subscripted to inform later users which reference was used: e.g., h_∞, h_m, or $h_{\text{adiabatic}}$.

Direct Methods

The two most commonly used methods for measuring the heat transfer coefficient are both derived from the same energy balance equation:

$$hA\left(T_o - T_{\text{ref}}\right) = \dot{e}_{\text{in}} + \dot{q}_{\text{cond,in}} + \dot{q}_{\text{rad,in}} - Mc\frac{dT}{d\tau} \tag{4.6.16}$$

where h = the heat transfer coefficient, W/m, °C
A = the area available for convective transport, m^2
T_{ref} = the reference temperature used in defining h, °C
T_o = the average surface temperature over the area A, °C
$\dot{e}_{\text{in}}$ = externally provided input, W
$\dot{q}_{\text{cond,in}}$ = net energy conducted in, W
$\dot{q}_{\text{rad,in}}$ = net energy radiated in, W
$Mc\ dT/d\tau$ = rate of increase of thermal energy stored within the system, W

Steady State. In the steady-state method, the transient term is zero (or nearly so), and h is determined by measuring the input power and the operating temperature, and correcting for losses. Equation (4.6.16) can be applied to differentially small elements or to whole specimens. The considered region must be reasonably uniform in temperature, so the energy storage term and the convective heat transfer term use the same value.

For tests of isolated objects, or embedded calorimeter sections, steady-state tests usually use high-conductivity specimens (e.g., copper or aluminum) with embedded electric heaters. The resulting value of h is the average over the area of the specimen. The Biot number, hL/k, for the specimen should be low (on the order of 0.01 or less) if only one temperature sensor is used in the specimen, so the surface temperature can be determined from the embedded sensor.

If a single heated element is used within an array of unheated elements, the resulting heat transfer coefficient is implicitly defined as $h_{adiabatic}$ and should be identified as such. Heat transfer coefficients measured with single-active-element tests cannot be used with the mixed mean fluid temperature.

When the variation of h over a surface is required, one common steady-state technique is to stretch a thin foil (stainless steel, or carbon impregnated paper, or gold deposited on polycarbonate) over an insulating substrate, and electrically heat the foil surface. Liquid crystals or infrared techniques can be used to map the surface temperature, from which the heat transfer coefficient distribution can be determined. The "heated foil with liquid crystal" approach was used by Cooper et al. in 1975 to measure heat transfer coefficients, and has since been used by many others. Hippensteele et al. (1985) have made extensive use of the foil technique in studies of gas turbine heat transfer. An example of their work on the end wall of a turbine cascade is shown in Figure 4.6.15.

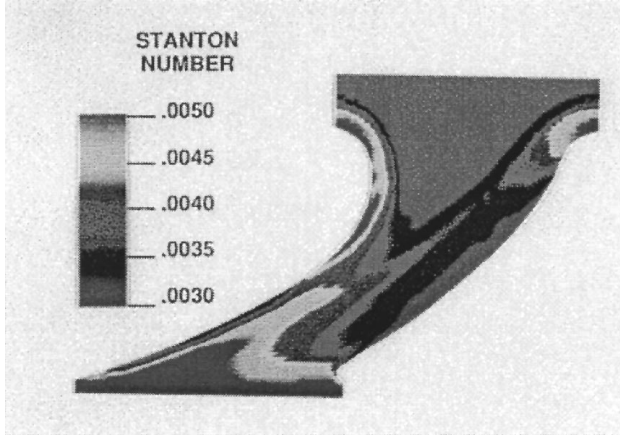


FIGURE 4.6.15 Heat transfer coefficient distribution on the end wall of a turbine cascade. (From Hippensteele, S.A. et al., NASA Technical Memorandum 86900, March, 1985. With permission.)

Hollingsworth et al. (1989) used a stainless steel foil heater for a study in air for an electronics cooling application, illustrated in Figure 4.6.16.

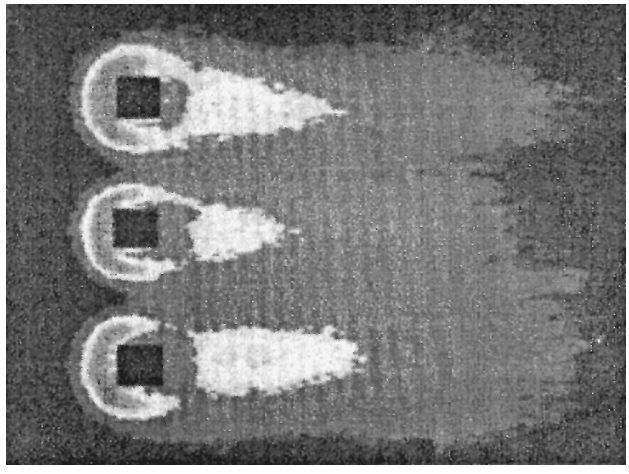


FIGURE 4.6.16 Visualization of the heat transfer coefficient distribution on a heated plate around three unheated cubes.

Another steady-state technique which reveals the distribution of *h* on the surface was introduced by den Ouden and Hoogendoorn (1974) and is currently in use by Meinders (1996). It uses a uniform and constant-temperature substrate (originally, a tank of warm water, now a copper block) covered with a layer of known thermal resistance (originally, a plate of glass, now a thin layer of epoxy). The surface was painted with liquid crystals (now visualized using infrared imaging) and the surface-temperature distribution determined. The inner (uniform) and outer (measured) temperature distributions are then used as boundary conditions to a three-dimensional conduction solver which calculates the total heat flux at each point on the surface. The total heat flux is corrected for radiation to yield the net convective transport at each point, from which *h* can be determined.

This method appears to have an advantage in accuracy over the heated foil technique because of the more accurate handling of substrate conduction.

Transient Lumped Parameter Systems. In the lumped parameter transient method, the specimen is assumed to be uniform in temperature at every instant through the transient. The power, $\dot{e}_{in}$, in Equation (4.6.16) is usually zero, although that is not necessary (one could simply change the power level at time zero to initiate the transient). At time zero, a transient is initiated, and the response curve recorded.

The data can be interpreted, and the validity of the first-order assumption tested at the same time by plotting $(T - T_{final})/(T_{initial} - T_{final})$ on the log scale of semilog coordinates, with time on the algebraic scale. If the line is straight, then the system is first order and the characteristic time can be determined from any two points on the line by

$$\tau = \frac{\left(t_2 - t_1\right)}{\ln\left\{\dfrac{T_{fin} - T_1}{T_{fin} - T_2}\right\}} \tag{4.6.17}$$

where τ = characteristic time, Mc/hA, sec
t_1 = time at the first instant
t_2 = time at the second instant
T_1 = specimen temperature at the first instant, °C
T_2 = specimen temperature at the second instant, °C
T_{fin} = specimen temperature after a long time (fluid temperature), °C

The heat transfer coefficient is extracted from the time-constant definition.

Indirect Methods

An increasingly popular method is the extraction of *h* from surface-temperature variations after a step in flow temperature using an inverse calculation method (see the section on inferential methods of heat flux measurement). The simplest inverse method assumes one-dimensional conduction into an infinitely thick plate of constant material properties. Even highly irregular geometries can be studied with this technique, if the streamwise extent of the specimen is small and the testing time is short. A short time interval is necessary so the penetration of the thermal wave is limited to a thin layer near the surface. The short streamwise extent is necessary so the temperature response of the surface upstream does not alter the thermal step applied to the downstream surface. This technique has been used to determine the heat transfer coefficient distribution on the inside walls of passages of irregular shape, by making the passage in a transparent material.

Naphthalene Sublimination. The equations for mass diffusion are similar to those for heat transfer, except for replacing the Prandtl number in the heat transfer equation by the Schmidt number in the diffusion equation. Thus, one could expect that the distribution of the mass transfer coefficients on a surface would mimic the distribution of the heat transfer coefficients.

The most commonly used analog technique is naphthalene sublimination. As early as 1940, the mass transfer/heat transfer similarity was used to estimate the heat transfer coefficient distribution. Naphthalene

is a solid material which sublimes at a reasonable rate in air at ambient temperature. Specimens can be cast in naphthalene with good precision, and the recession of the surface mapped as a function of position and time using automated or semiautomated measuring equipment. The surface recession over a known interval of time is a measure of the mass transfer rate, from which the mass transfer coefficient can be deduced.

Naphthalene experiments are generally done at uniform temperature; hence, a uniform vapor pressure exists at the surface. This corresponds to the heat transfer situation of heat transfer from a uniform temperature surface. No counterpart of the uniform heat flux situation has been produced using naphthalene, nor have there been experiments corresponding to variable wall temperature.

Naphthalene sublimation experiments do not suffer from any counterpart of the conduction heat transfer in the substrate. Conduction makes it difficult to work near discontinuities in wall temperature in a heat transfer experiment. Details of the fine structure of mass transfer near obstructions and discontinuities can be resolved in naphthalene experiments, but those details might not exist in a heat transfer process. The Prandtl number of air is much lower than the Schmidt number of naphthalene diffusing in air; hence, thermal conduction would tend to blur out sharp gradients in the temperature field more than diffusion would blur out gradients in naphthalene concentration.

The Schmidt number of naphthalene in air is about 2.5, far different than the Prandtl number of air (0.71); hence, the mass transfer coefficient deduced from a naphthalene experiment is not numerically equal to the heat transfer coefficient which would have existed at those conditions. The usual recommendation is to adjust for the Prandtl number of Schmidt number using a relation of the form:

$$\mathrm{St}\,\mathrm{Pr}^{2/3} = f\{\mathrm{Re}\} = \mathrm{Sh}_j \mathrm{Sc}_j^{2/3} \tag{4.6.18}$$

based on laminar results. That recommendation has not been seriously tested by experiments in turbulent and separated flows. By using nominal values of the Schmidt number and Prandtl number, the heat transfer Stanton number would be 2.3 times higher than the measured Sherwood number and an uncertainty of 10% in that ratio would alter the inferred heat transfer coefficient by 23%.

System Performance Matching. Sometimes the "effective average heat transfer coefficient" for a system is inferred from the overall behavior of the system, e.g., estimating h from the effectiveness of a heat exchanger. Values deduced by this means cannot be expected to agree well with direct measurements unless a very sophisticated system description model is used.

References

Moffat, R.J., The gradient approach to thermocouple circuitry, *Temperature, Its Measurement and Control in Science and Industry*, Rienhold, New York, 1962.

Steinhart, J.S. and Hart, S.R., Calibration curves for thermistors, *Deep Sea Res.*, 15, 497, 1968.

Rhee, H.S., Koseff, J.R., and Street, R.L., Flow visualization of a recirculating flow by rheoscopic liquid and liquid crystal techniques, *Exp. Fluids*, 2, 57–64, 1984.

Hollingsworth, K., Boehman, A.L., Smith, E.G., and Moffat, R.J., Measurement of temperature and heat transfer coefficient distributions in a complex flow using liquid crystal thermography and true-color image processing, in *Coll. Pap. Heat Transfer, ASME HTD*, 123, 35–42, Winter Annual Meeting, 1989.

Cooper, T.E., Field, R.J., and Meyer, J.F., Liquid crystal thermography and its application to the study of convective heat transfer, *J. Heat Transfer*, 97, 442–450, 1975.

Hippensteele, S.A., Russell, L.M., and Torres, F.J., Local Heat Transfer Measurements on a Large Scale Model Turbine Blade Airfoil Using a Composite of a Heater Element and Liquid Crystals, NASA Technical Memorandum 86900, March 1985.

den Ouden, C. and Hoogendoorn, C.J., Local convective heat transfer coefficients for jets impinging on a plate: experiments using a liquid crystal technique, in *Proc. of the 5th Int. Heat Transfer Conf.*, Vol. 5, AIChE, New York, 1974, 293–297.

Personal Communication from Erwin Meinders, March 1996. Work in progress at the Technical University of Delft under Prof. Hanjalic.

Akino, N. and Kunugi, T., *ASME HTD*, Vol. 112, 1989.

4.7 Mass Transfer

Anthony F. Mills

Introduction

Mass transfer may occur in a gas mixture, a liquid solution, or a solid solution. There are several physical mechanisms that can transport a chemical species through a phase and transfer it across phase boundaries. The two most important mechanisms are ordinary diffusion and convection. Mass diffusion is analogous to heat conduction and occurs whenever there is a gradient in the concentration of a species. Mass convection is essentially identical to heat convection: a fluid flow that transports heat may also transport a chemical species. The similarity of mechanisms of heat transfer and mass transfer results in the mathematics often being identical, a fact that can be exploited to advantage. But there are some significant differences between the subjects of heat and mass transfer. One difference is the much greater variety of physical and chemical processes that require mass transfer analysis. Another difference is the extent to which the essential details of a given process may depend on the particular chemical system involved, and on temperature and pressure.

In the next subsection, concentrations, velocities, and fluxes are defined, and special attention is paid to phase interfaces where the concentration of a chemical species is almost always discontinuous. Fick's law of ordinary diffusion is introduced in the third section, where other diffusion phenomena are also discussed. The fourth section presents various forms of the species conservation equation. Results for diffusion in a stationary medium are given in the fifth section, and include steady diffusion across a plane wall, transient diffusion in a semi-infinite solid, and diffusion in a porous catalyst. Results for diffusion in a moving medium are given in the sixth section, and the Stefan flow is introduced for diffusion with one component stationary. Also considered are particle combustion, droplet evaporation, and combustion of a volatile liquid hydrocarbon fuel droplet. The last section deals with mass convection. Low mass transfer rate theory is presented and how to exploit the analogy between convective heat and mass transfer is shown. Particular attention is given to situations involving simultaneous heat and mass transfer associated with evaporation or condensation. The section closes by presenting high mass transfer rate theory for convection, and gives engineering calculation methods for boundary layer flows that account for variable property effects.

Concentrations, Velocities, and Fluxes

Definitions of Concentrations

In a gas mixture, or liquid or solid solution, the local *concentration* of a mass species can be expressed in a number of ways. The *number density* of species i in a mixture or solution of n species is defined as

$$\begin{aligned}\text{Number density of species } i &\equiv \text{Number of molecules of } i \text{ per unit volume} \\ &\equiv \mathcal{N}_i \text{ molecules/m}^3\end{aligned} \tag{4.7.1}$$

Alternatively, if the total number of molecules of all species per unit volume is denoted as $\mathcal{N}$, then we define the *number fraction* of species i as

$$n_i \equiv \frac{\mathcal{N}_i}{\mathcal{N}}; \qquad \mathcal{N} = \sum \mathcal{N}_i \tag{4.7.2}$$

where the summation is over all species present, $i = 1,2,\ldots,n$. Equations (4.7.1) and (4.7.2) describe *microscopic* concepts and are used, for example, when the kinetic theory of gases is used to describe transfer processes.

Whenever possible, it is more convenient to treat matter as a continuum. Then the smallest volume considered is sufficiently large for macroscopic properties such as pressure and temperature to have their usual meanings. For this purpose we also require *macroscopic* definitions of concentration. First, on a mass basis,

$$\begin{aligned} &\textit{Mass concentration} \text{ of species } i \equiv \text{partial density of species } i \\ &\equiv \rho_i \ \mathrm{kg/m^3} \end{aligned} \tag{4.7.3}$$

The total mass concentration is the total mass per unit volume, that is, the density $\rho = \sum \rho_i$. The *mass fraction* of species i is defined as

$$m_i = \frac{\rho_i}{\rho} \tag{4.7.4}$$

Second, on a molar basis,

$$\begin{aligned} &\textit{Molar concentration} \text{ of species } i \equiv \text{number of moles of } i \text{ per unit volume} \\ &\equiv c_i \ \mathrm{kmol/m^3} \end{aligned} \tag{4.7.5}$$

If M_i (kg/kmol) is the molecular weight of species i, then

$$c_i = \frac{\rho_i}{M_i} \tag{4.7.6}$$

The total molar concentration is the molar density $c = \sum c_i$. The *mole fraction* of species i is defined as

$$x_i \equiv \frac{c_i}{c} \tag{4.7.7}$$

A number of important relations follow directly from these definitions. The mean molecular weight of the mixture of solution is denoted M and may be expressed as

$$M = \frac{\rho}{c} = \sum x_i M_i \tag{4.7.8a}$$

or

$$\frac{1}{M} = \sum \frac{m_i}{M_i} \tag{4.7.8b}$$

There are summation rules

$$\sum m_i = 1 \tag{4.7.9a}$$

$$\sum x_i = 1 \tag{4.7.9b}$$

It is often necessary to have the mass fraction of species *i* expressed explicitly in terms of mole fractions and molecular weights; this relation is

$$m_i = \frac{x_i M_i}{\sum x_j M_j} = x_i \frac{M_i}{M} \tag{4.7.10a}$$

and the corresponding relation for the mole fraction is

$$x_i = \frac{m_i / M_i}{\sum m_j / M_j} = m_i \frac{M}{M_i} \tag{4.7.10b}$$

Dalton's law of partial pressures for an ideal gas mixture states that

$$P = \sum P_i, \qquad \text{where} \quad P_i = \rho_i R_i T \tag{4.7.11}$$

Dividing partial pressure by total pressure and substituting $R_i = \mathcal{R}/M_i$ gives

$$\frac{P_i}{P} = \frac{\rho_i}{M_i} \frac{\mathcal{R} T}{P} = c_i \frac{\mathcal{R} T}{P} = x_i \frac{c \mathcal{R} T}{P} = x_i \tag{4.7.12}$$

Thus, for an ideal gas mixture, the mole fraction and partial pressure are equivalent measures of concentration (as also is the number fraction).

A commonly used specification of the composition of dry air is 78.1% N_2, 20.9% O_2, and 0.9% Ar, by volume. (The next largest component is CO_2, at 0.3%.) Since equal volumes of gases contain the same number of moles, specifying composition on a volume basis is equivalent to specifying mole fractions, namely,

$$x_{N_2} = 0.781; \qquad x_{O_2} = 0.209; \qquad x_{Ar} = 0.009$$

The corresponding mass fractions are calculated to be

$$m_{N_2} = 0.755; \qquad m_{O_2} = 0.231; \qquad m_{Ar} = 0.014$$

Concentrations at Interfaces

Although temperature is continuous across a phase interface, concentrations are usually discontinuous. In order to define clearly concentrations at interfaces, we introduce imaginary surfaces, denoted *u* and *s*, on both sides of the real interface, each indefinitely close to the interface, as shown in Figure 4.7.1 for water evaporating into an airstream. Thus, the liquid-phase quantities at the interface are subscripted *u*, and gas-phase quantities are subscripted *s*. If we ignore the small amount of air dissolved in the water, $x_{H_2O,u} = 1$. Notice that the subscript preceding the comma denotes the chemical species, and the subscript following the comma denotes location. To determine $x_{H_2O,s}$ we make use of the fact that, except in extreme circumstances, the water vapor and air mixture at the *s*-surface must be in thermodynamic equilibrium with water at the *u*-surface. Equilibrium data for this system are found in conventional steam tables: the saturation vapor pressure of steam at the water temperature, T_s, ($T_s = T_u$), is the required partial pressure $P_{H_2O,s}$. With the total pressure *P* known, $x_{H_2O,s}$ is calculated as $P_{H_2O,s}/P$. If $m_{H_2O,s}$ is required, Equation (4.7.10a) is used.

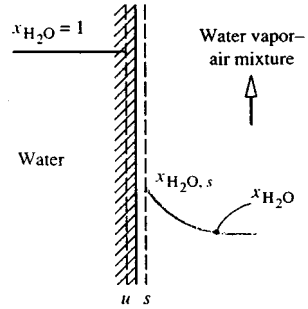


FIGURE 4.7.1 Concentrations at a water-air interface.

For example, at T_s = 320 K, the saturation vapor pressure is obtained from steam tables as 0.10535 $\times 10^5$ Pa. If the total pressure is 1 atm = 1.0133×10^5,

$$x_{H_2O,s} = \frac{0.10535 \times 10^5}{1.0133 \times 10^5} = 0.1040$$

$$m_{H_2O,s} = \frac{(0.1040)(18)}{(0.1040)(18) + (1 - 0.1040)(29)} = 0.06720$$

For a gas or solid dissolving in a liquid, equilibrium data are often referred to simply as solubility data, found in chemistry handbooks. Many gases are only sparingly soluble, and for such dilute solutions solubility data are conveniently represented by *Henry's law*, which states that the mole fraction of the gas at the *s*-surface is proportional to its mole fraction in solution at the *u*-surface, the constant of proportionality being the *Henry number*, He. For species *i*,

$$x_{i,s} = \mathrm{He}_i x_{i,u} \tag{4.7.13}$$

The Henry number is inversely proportional to total pressure and is also a function of temperature. The product of Henry number and total pressure is the *Henry constant*, C_{He}, and for a given species is a function of temperature only:

$$\mathrm{He}_i P = C_{\mathrm{He}_i}(T) \tag{4.7.14}$$

Solubility data are given in Table 4.7.1.

TABLE 4.7.1 Henry Constants C_{He} for Dilute Aqueous Solutions at Moderate Pressures ($P_{i,s}/x_{i,u}$ in atm, or in bar = 10^5 Pa, within the accuracy of the data).

Solute	290 K	300 K	310 K	320 K	330 K	340 K
H_2S	440	560	700	830	980	1,140
CO_2	1,280	1,710	2,170	2,720	3,220	—
O_2	38,000	45,000	52,000	57,000	61,000	65,000
H_2	67,000	72,000	75,000	76,000	77,000	76,000
CO	51,000	60,000	67,000	74,000	80,000	84,000
Air	62,000	74,000	84,000	92,000	99,000	104,000
N_2	16,000	89,000	101,000	110,000	118,000	124,000

For example, consider absorption of carbon dioxide from a stream of pure CO_2 at 2 bar pressure into water at 310 K. From Table 4.7.1, $C_{He} = 2170$ bar; thus

$$He_{CO_2} = \frac{2170}{2} = 1085; \qquad x_{CO_2,u} = \frac{1}{1085} = 9.22 \times 10^{-4}$$

Dissolution of gases into metals is characterized by varied and rather complex interface conditions. Provided temperatures are sufficiently high, hydrogen dissolution is reversible (similar to CO_2 absorption into water); hence, for example, titanium-hydrogen solutions can exist only in contact with a gaseous hydrogen atmosphere. As a result of hydrogen going into solution in atomic form, there is a characteristic square root relation

$$m_{H_2,u} \alpha P_{H_2,s}^{1/2}$$

The constant of proportionality is strongly dependent on temperature, as well as on the particular titanium alloy: for Ti-6Al-4V alloy it is twice that for pure titanium. In contrast to hydrogen, oxygen dissolution in titanium is irreversible and is complicated by the simultaneous formation of a rutile (TiO_2) scale on the surface. Provided some oxygen is present in the gas phase, the titanium-oxygen *phase diagram* (found in a metallurgy handbook) shows that $m_{O_2,u}$ in alpha-titanium is 0.143, a value essentially independent of temperature and O_2 partial pressure. Dissolution of oxygen in zirconium alloys has similar characteristics to those discussed above for titanium.

All the preceding examples of interface concentrations are situations where thermodynamic equilibrium can be assumed to exist at the interface. Sometimes thermodynamic equilibrium does not exist at an interface: a very common example is when a chemical reaction occurs at the interface, and temperatures are not high enough for equilibrium to be attained. Then the concentrations of the reactants and products at the *s*-surface are dependent both on the rate at which the reaction proceeds — that is, the *chemical kinetics* — as well as on mass transfer considerations.

Definitions of Fluxes and Velocities

The mass (or molar) flux of species *i* is a vector quantity giving the mass (or moles) of species *i* that pass per unit time through a unit area perpendicular to the vector (Figure 4.7.2). We denote the absolute mass and molar fluxes of species *i*, that is, relative to stationary coordinate axes, as $\boldsymbol{n}_i$ (kg/m^{2}sec) and $\boldsymbol{N}_i$ (kmol/m^{2}sec), respectively. The absolute mass flux of the mixture (mass velocity) is

$$\boldsymbol{n} = \sum \boldsymbol{n}_i \tag{4.7.15}$$

and the local mass-average velocity is

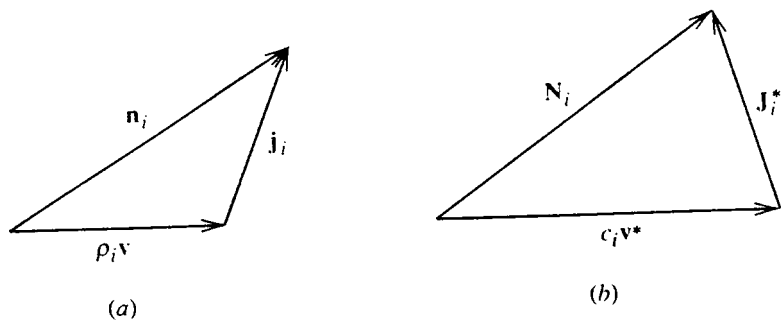


FIGURE 4.7.2 Flux vectors: (a) mass basis, (b) molar basis.

$$\boldsymbol{v} = \frac{\boldsymbol{n}}{\rho}\,\mathrm{m/sec} \tag{4.7.16}$$

The velocity v is the velocity that would be measure by a Pitot tube and corresponds to the velocity used in considering pure fluids. On a molar basis, the absolute molar flux of the mixture is

$$\boldsymbol{N} = \sum \boldsymbol{N}_i \tag{4.7.17}$$

and the local molar-average velocity is

$$\boldsymbol{v}^* = \frac{\boldsymbol{N}}{c}\,\mathrm{m/sec} \tag{4.7.18}$$

The absolute fluxes at species *i* have two components. On a mass basis we write

$$\boldsymbol{n}_i = \rho_i \boldsymbol{v} + \boldsymbol{j}_i \tag{4.7.19}$$

where $\rho_i \boldsymbol{n}$ is transport of species *i* by bulk motion of the fluid at velocity v and is the *convective* component. Thus, $\boldsymbol{j}_i$ is transport of species *i* relative to the mass average velocity; it is called the *diffusive* component because most commonly it is due to ordinary (concentration) diffusion of the species. On a molar basis the corresponding relation is

$$\boldsymbol{N}_i = c_i \boldsymbol{v}^* + \boldsymbol{J}_i^* \tag{4.7.20}$$

Some important relations follow from these definitions:

$$\sum \boldsymbol{j}_i = \sum \boldsymbol{J}_i^* = 0 \tag{4.7.21}$$

$$\boldsymbol{N}_i = \frac{\boldsymbol{n}_i}{M_i} \tag{4.7.22}$$

$$\boldsymbol{n}_i = \rho_i \boldsymbol{v} + \boldsymbol{j}_i = m_i \sum \boldsymbol{n}_i + \boldsymbol{j}_i \tag{4.7.23a}$$

$$\boldsymbol{N}_i = c_i \boldsymbol{v}^* + \boldsymbol{J}_i^* = x_i \sum \boldsymbol{N}_i + \boldsymbol{J}_i^* \tag{4.7.23b}$$

Mechanisms of Diffusion

Ordinary Diffusion

Fick's law of ordinary diffusion is a linear relation between the rate of diffusion of a chemical species and the local concentration gradient of that species. It is exact for a binary gas mixture, for which the kinetic theory of gases gives

$$\boldsymbol{j}_1 = -\rho \mathcal{D}_{12} \nabla m_1\,\mathrm{kg/m^2\ sec} \tag{4.7.24a}$$

on a mass basis, and

$$\boldsymbol{J}_1^* = -c\mathcal{D}_{12}\nabla x_1 \text{ kg/m}^2 \text{ sec} \tag{4.7.24b}$$

on a molar basis; $\mathcal{D}_{12}$ (m^2/sec) is the binary diffusion coefficient (or mass diffusivity), and $\mathcal{D}_{21} = \mathcal{D}_{21}$. Equations (4.7.24a) and (4.7.24b) are mathematically equivalent; however, notice that it is incorrect to write

$$\boldsymbol{j}_i = -\mathcal{D}_{12}\nabla\rho_1 \tag{4.7.25}$$

since $\nabla\rho_1 \neq \rho \nabla m_1$ in general. Fick's law in the form of Equations (4.7.24a) and (4.7.24b) is also valid for dilute liquid and solid solutions, for which it is often possible to assume ρ (or c) constant, and then Equation (4.7.25) or its molar equivalent are good approximations.

Ordinary diffusion in multicomponent systems is described by the Stefan–Maxwell equations (Hirschfelder et al., 1954). These equations are difficult to use for engineering analysis. In gas mixtures containing species that do not have widely varying molecular weights, it is possible to model approximately the diffusion process by using an effective binary diffusion coefficient in Fick's law. This coefficient is a suitable average over the species in the mixture, and may be calculated from

$$\mathcal{D}_{1m} = \frac{\left(1 - x_1\right)}{\sum_{i=2}^{n}\left(x_i/\mathcal{D}_{1i}\right)}; \qquad x_1 \ll 1 \tag{4.7.26}$$

This equation works well for most mixtures of combustion gases (except those containing appreciable concentrations of H or H_2).

Binary diffusion coefficients at 300 K are of the order of 10^{-5} m^2/sec in gases at 1 atm, 10^{-9} m^2/sec in aqueous solutions, and 10^{-10} to 10^{-13} m^2/sec in solids. However, the product $\rho\mathcal{D}$ or $(c\mathcal{D})$ is, at most, one order of magnitude different for gases and liquids. Data for diffusion coefficients may be found in Tables 4.7.2 through 4.7.5.

Molecules in a gas mixture, and in a liquid or solid solution, can diffuse by mechanisms other than ordinary diffusion governed by Fick's law. *Thermal diffusion* is diffusion due to a temperature gradient and is often called the *Soret effect*. Thermal diffusion is usually negligible compared with ordinary diffusion, unless the temperature gradient is very large. However, there are some important processes that depend on thermal diffusion, the most well known being the large-scale separation of uranium isotopes. *Pressure diffusion* is diffusion due to a pressure gradient and is also usually negligible unless the pressure gradient is very large. Pressure diffusion is the principle underlying the operation of a centrifuge. Centrifuges are used to separate liquid solutions and are increasingly being used to separate gaseous isotopes as well. *Forced diffusion* results from an external force field acting on a molecule. Gravitational force fields do not cause separation since the force per unit mass of a molecule is constant. Forced diffusion occurs when an electrical field is imposed on an electrolyte (for example, in charging an automobile battery), on a semiconductor, or on an ionized gas (for example, in a neon tube or metal-ion laser). Depending on the strength of the electric field, rates of forced diffusion can be very large.

Some interesting diffusion phenomena occur in porous solids. When a gas mixture is in a porous solid, such as a catalyst pellet or silica–gel particle, the pores can be smaller than the mean free path of the molecules. Then, the molecules collide with the wall more often than with other molecules. In the limit of negligible molecule collisions we have *Knudsen diffusion*, also called *free molecule flow* in the fluid mechanics literature. If the pore size approaches the size of a molecule, then Knudsen diffusion becomes negligible and *surface diffusion*, in which adsorbed molecules move along the pore walls, becomes the dominant diffusion mechanism.

TABLE 4.7.2 Diffusion Coefficients in Air at 1 atm (1.013×10^5 Pa)[a]

	Binary Diffusion Coefficient ($m^2/sec \times 10^4$)							
***T*(K)**	**O_2**	**CO_2**	**CO**	**C_7H_6**	**H_2**	**NO**	**SO_2**	**He**
200	0.095	0.074	0.098	0.036	0.375	0.088	0.058	0.363
300	0.188	0.157	0.202	0.075	0.777	0.180	0.126	0.713
400	0.325	0.263	0.332	0.128	1.25	0.303	0.214	1.14
500	0.475	0.385	0.485	0.194	1.71	0.443	0.326	1.66
600	0.646	0.537	0.659	0.270	2.44	0.603	0.440	2.26
700	0.838	0.684	0.854	0.364	3.17	0.782	0.576	2.91
800	1.05	0.857	1.06	0.442	3.93	0.978	0.724	3.64
900	1.26	1.05	1.28	0.538	4.77	1.18	0.887	4.42
1000	1.52	1.24	1.54	0.641	5.69	1.41	1.060	5.26
1200	2.06	1.69	2.09	0.881	7.77	1.92	1.440	7.12
1400	2.66	2.17	2.70	1.13	9.90	2.45	1.870	9.20
1600	3.32	2.75	3.37	1.41	12.5	3.04	2.340	11.5
1800	4.03	3.28	4.10	1.72	15.2	3.70	2.850	13.9
2000	4.80	3.94	4.87	2.06	18.0	4.48	3.360	16.6

[a] Owing to the practical importance of water vapor-air mixtures, engineers have used convenient empirical formulas for $\mathcal{D}_{H_2O\ air}$. A formula that has been widely used for many years is

$$\mathcal{D}_{H_2O\ air} = 1.97 \times 10^{-5} \left(\frac{P_0}{P}\right)\left(\frac{T}{T_0}\right)^{1.685} \text{ m}^2/\text{sec}; \quad 273\ \text{K} < T < 373\ \text{K}$$

where $P_0 = 1$ atm; $T_0 = 256$ K. More recently, the following formula has found increasing use. (Marrero, T.R. and Mason, E.A. 1992. Gaseous diffusion coefficients, *J. Phys. Chem. Ref. Data*, 1, 3–118):

$$\mathcal{D}_{H_2O\ air} = 1.87 \times 10^{-10} \frac{T^{2.072}}{P}; \quad 280\ \text{K} < T < 450\ \text{K}$$

$$= 2.75 \times 10^{-9} \frac{T^{1.632}}{P}; \quad 450\ \text{K} < T < 1070\ \text{K}$$

for P in atmospheres and T in kelvins. Over the temperature range 290 to 330 K, the discrepancy between the two formulas is less than 2.5%. For small concentrations of water vapor in air, the older formula gives a constant value of $Sc_{H_2O\ air} = 0.61$ over the temperature range 273 to 373 K. On the other hand, the Marrero and Mason formula gives values of $Sc_{H_2O\ air}$ that vary from 0.63 at 280 K to 0.57 at 373 K.

Very small particles of 10^{-3} to 10^{-1} μm size — for example, smoke, soot, and mist — behave much like large molecules. Ordinary diffusion of such particles is called *Brownian motion* and is described in most elementary physics texts. Diffusion of particles due to a temperature gradient is called *thermophoresis* and plays an important role for larger particles, typically in the size range 10^{-1} to 1 μm. Diffusion of particles in a gas mixture due to concentration gradients of molecular species is called *diffusiophoresis.* *Forced diffusion* of a charged particle in an electrical field is similar to that for an ionized molecular species. Thermal and electrostatic precipitators are used to remove particles from power plant and incinerator stack gases, and depend on thermophoresis and forced diffusion, respectively, for their operation. Diffusion phenomena are unimportant for particles of size greater than about 1 μm in air at 1 atm; the motion of such particles is governed by the laws of Newtonian mechanics. Transport of particles is dealt with in the *aerosol science* literature.

Species Conservation Equation

The principle of conservation of a chemical species is used to derive the *species conservation equation.* On a mass basis this equation is

TABLE 4.7.3 Schmidt Number for Vapors in Dilute Mixture in Air at Normal Temperature, Enthalpy of Vaporization, and Boiling Point at 1 atm[a]

Vapor	Chemical Formula	Sc[b]	h_{fg}, J/kg × 10^{-6}	T_{BP}, K
Acetone	CH_3COCH_3	1.42	0.527	329
Ammonia	NH_3	0.61	1.370	240
Benzene	C_6H_6	1.79	0.395	354
Carbon dioxide	CO_2	1.00	0.398	194
Carbon monoxide	CO	0.77	0.217	81
Chlorine	Cl_2	1.42	0.288	238
Ethanol	CH_3CH_2OH	1.32	0.854	352
Helium	He	0.22	—	4.3
Heptane	C_7H_{16}	2.0	0.340	372
Hydrogen	H_2	0.20	0.454	20.3
Hydrogen sulfide	H_2S	0.94	0.548	213
Methanol	CH_3OH	0.98	1.110	338
Naphthalene	$C_{10}H_8$	2.35[c]	—	491
Nitric oxide	NO	0.87	0.465	121
Octane	C_8H_{18}	2.66	0.303	399
Oxygen	O_2	0.83	0.214	90.6
Pentane	C_5H_{12}	1.49	0.357	309
Sulfur dioxide	SO_2	1.24	0.398	263
Water vapor	H_2O	0.61	2.257	373

[a] With the Clausius-Clapeyron relation, one may estimate vapor pressure as

$$P_{\text{sat}} \simeq \exp\left\{-\frac{Mh_{fg}}{\mathcal{R}}\left(\frac{1}{T}-\frac{1}{T_{BP}}\right)\right\} \text{ atm} \quad \text{for } T \sim T_{BP}$$

[b] The Schmidt number is defined as $\text{Sc} = \mu/\rho\mathcal{D} = \nu/\mathcal{D}$. Since the vapors are in small concentrations, values for μ, ρ, and ν can be taken as pure air values.

[c] From a recent study by Cho, C., Irvine, T.F., Jr., and Karni, J. 1992. Measurement of the diffusion coefficient of naphthalene into air, *Int. J. Heat Mass Transfer*, 35, 957–966. Also, $h_{sg} = 0.567 \times 10^6$ J/kg at 300 K.

$$\frac{\partial \rho_i}{\partial t} + \nabla \cdot \boldsymbol{n}_i = \dot{r}_i''' \tag{4.7.27}$$

and on a molar basis

$$\frac{\partial c_i}{\partial t} + \nabla \cdot \boldsymbol{N}_i = \dot{R}_i''' \tag{4.7.28}$$

where $\dot{r}_i'''$ and $\dot{R}_i'''$ are the mass and molar rates of production of species *i* due to chemical reactions. Summing Equation 4.7.27 over all species gives the mass conservation or continuity equation,

$$\frac{\partial \rho}{\partial t} + \nabla \cdot \rho \boldsymbol{v} = 0 \tag{4.7.29}$$

The molar form is

$$\frac{\partial c}{\partial t} + \nabla \cdot c\boldsymbol{v}^* = \sum_i \dot{R}_i''' \tag{4.7.30}$$

since, in general, moles are not conserved in chemical reactions. A useful alternative form to Equation 4.7.27 can be obtained using Equations (4.7.23a) and (4.7.29) and is

TABLE 4.7.4 Schmidt Numbers for Dilute Solution in Water at 300 K[a]

Solute	Sc	*M*
Helium	120	4.003
Hydrogen	190	2.016
Nitrogen	280	28.02
Water	340	18.016
Nitric oxide	350	30.01
Carbon monoxide	360	28.01
Oxygen	400	32.00
Ammonia	410	17.03
Carbon dioxide	420	44.01
Hydrogen sulfide	430	34.08
Ethylene	450	28.05
Methane	490	16.04
Nitrous oxide	490	44.02
Sulfur dioxide	520	64.06
Sodium chloride	540	58.45
Sodium hydroxide	490	40.00
Acetic acid	620	60.05
Acetone	630	58.08
Methanol	640	32.04
Ethanol	640	46.07
Chlorine	670	70.90
Benzene	720	78.11
Ethylene glycol	720	62.07
n-Propanol	730	60.09
i-Propanol	730	60.09
Propane	750	44.09
Aniline	800	93.13
Benzoic acid	830	122.12
Glycerol	1040	92.09
Sucrose	1670	342.3

[a] Schmidt number $Sc = \mu/\rho\mathcal{D}$; since the solutions are dilute, μ and ρ can be taken as pure water values. For other temperatures use $Sc/Sc_{300\ K} \simeq (\mu^2/\rho T)/(\mu^2\rho T)_{300\ K}$, where μ and ρ are for water, and T is absolute temperature. For chemically similar solutes of different molecular weights use $Sc_2/Sc_1 \simeq (M_2/M_1)^{0.4}$. A table of $(\mu^2/\rho T)/(\mu^2\rho T)_{300\ K}$ for water follows.

T(K)	$(\mu^2/\rho T)/(\mu^2/\rho T)_{300\ K}$
290	1.66
300	1.00
310	0.623
320	0.429
330	0.296
340	0.221
350	0.167
360	0.123
370	0.097

From Spalding, D.B. 1963. *Convective Mass Transfer*, McGraw-Hill, New York. With permission.

$$\rho\frac{Dm_i}{Dt} = \nabla\cdot\boldsymbol{j}_i + \dot{r}_i''' \tag{4.7.31}$$

where D/Dt is the substantial derivative operator.

If we consider a binary system of species 1 and 2 and introduce Fick's law, Equation (4.7.24a) into Equation (4.7.31), then

TABLE 4.7.5 Diffusion Coefficients in Solids, $\mathcal{D} = \mathcal{D}_o \exp(-E_a/\mathcal{R}T)$

System	$\mathcal{D}_o$, m²/sec	E_a, kJ/kmol
Oxygen-Pyrex glass	6.19×10^{-8}	4.69×10^{4}
Oxygen-fused silica glass	2.61×10^{-9}	3.77×10^{4}
Oxygen-titanium	5.0×10^{-3}	2.13×10^{5}
Oxygen-titanium alloy (Ti-6Al-4V)	5.82×10^{-2}	2.59×10^{5}
Oxygen-zirconium	4.68×10^{-5}	7.06×10^{5}
Hydrogen-iron	7.60×10^{-8}	5.60×10^{3}
Hydrogen-α-titanium	1.80×10^{-6}	5.18×10^{4}
Hydrogen-β-titanium	1.95×10^{-7}	2.78×10^{4}
Hydrogen-zirconium	1.09×10^{-7}	4.81×10^{4}
Hydrogen-Zircaloy[4]	1.27×10^{-5}	6.05×10^{5}
Deuterium-Pyrex glass	6.19×10^{-8}	4.69×10^{4}
Deuterium-fused silica glass	2.61×10^{-9}	3.77×10^{4}
Helium-Pyrex glass	4.76×10^{-8}	2.72×10^{4}
Helium-fused silica glass	5.29×10^{-8}	2.55×10^{4}
Helium-borosilicate glass	1.94×10^{-9}	2.34×10^{4}
Neon-borosilicate glass	1.02×10^{-10}	3.77×10^{4}
Carbon-FCC iron	2.3×10^{-5}	1.378×10^{5}
Carbon-BCC iron	1.1×10^{-6}	8.75×10^{4}

Various sources.

$$\rho\frac{Dm_i}{Dt} = \nabla\cdot\left(\rho\mathcal{D}_{12}\nabla m_1\right) + \dot{r}_1''' \tag{4.7.32}$$

When working on a mass basis we define a stationary medium as one in which the mass average velocity v is zero everywhere. Substituting in Equation (4.7.32) with no chemical reactions and assuming constant properties,

$$\frac{\partial m_1}{\partial t} = \mathcal{D}_{12}\nabla^2 m_1 \tag{4.7.33}$$

which is the *diffusion* equation, and is the mass transfer analog to Fourier's equation for heat conduction. For steady diffusion, Equation (4.7.33) reduces to Laplace's equation

$$\nabla^2 m_1 = 0 \tag{4.7.34}$$

Notice that since properties have been assumed constant, any measure of concentration can be used in Equations (4.7.33) and (4.7.34), for example ρ_1, c_1, and x_1.

Diffusion in a Stationary Medium

Many problems involving diffusion in a stationary medium are governed by the diffusion equation (Equation 4.7.33). Often solutions may be obtained from their heat conduction analogs. Some important cases follow.

Steady Diffusion through a Plane Wall

The mass flow of species 1 across a plane wall of thickness L and cross-sectional area A is

$$\dot{m}_1 = \frac{\rho\mathcal{D}_{12}A}{L}\left(m_{1,u} - m_{1,u'}\right)\ \text{kg/m}^2\,\text{sec} \tag{4.7.35}$$

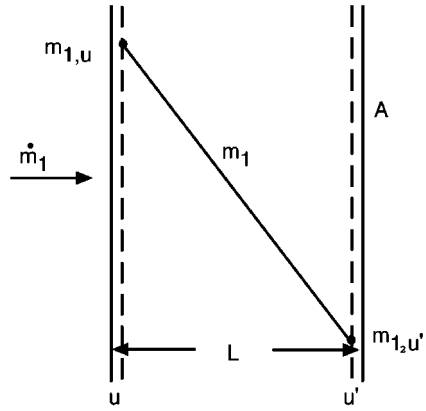


FIGURE 4.7.3 Steady diffusion across a plane wall.

where the *u*- and *u*′-surfaces are shown in Figure 4.7.3. Solubility data are required to relate the *u*- and *u*′-surface concentrations to *s*- and *s*′-surface concentrations. Alternatively for systems that obey Henry's law, a solubility $\mathcal{S}$ can be defined as the volume of solute gas (at STP of 0°C and 1 atm) dissolved in unit volume when the gas is at a partial pressure of 1 atm. Then, defining permeability $\mathcal{P}_{12}$ as the product $\mathcal{D}_{12}\mathcal{S}$, the volume flow of species 1 is

$$\dot{V}_1 = \frac{\mathcal{P}_{12}A}{L}\left(P_{1,s} - P_{1,s'}\right) \text{ m}^3 \text{ (STP)/sec} \tag{4.7.36}$$

where the partial pressures P_1 are in atmospheres. The SI units for permeability are m^3 (STP)/m^2sec(atm/m). Permeability and solubility data are given in Table 4.7.6. For example, consider helium at 10^5 Pa contained in a 7056-glass vessel with a 1-mm-thick wall at 680 K. For a surface area of 0.01 m^2, the leakage rate into ambient air is

$$\dot{V} = \frac{\left(1.0\times10^{-12}\right)(0.01)}{(0.001)}\left(10^5 - 0\right) = 1.0\times10^{-6} \text{ m}^3 \text{ (STP)/sec}$$

where the value $\mathcal{P}_{12}$ was obtained from Table 4.7.6.

In general, mass fractions are discontinuous across phase interfaces. Hence, Equation (4.7.35) cannot be generalized to a number of walls in series by simply adding diffusion resistances. However, equilibrium partial pressures P_1 are continuous, and for two walls *A* and *B*, Equation 4.7.36 becomes

$$\dot{V}_1 = \frac{P_{1,s} - P_{1,s'}}{\dfrac{L_A}{\mathcal{P}_{1A}A} + \dfrac{L_B}{\mathcal{P}_{1B}A}} \text{ m}^3 \text{ (STP)/sec} \tag{4.7.37}$$

Transient Diffusion in a Semi-Infinite Solid

The typically low diffusion coefficients characterizing solids result in many situations where concentration changes are limited to a thin region near the surface (of thickness $\delta_c \sim (\mathcal{D}_{12}t)^{1/2}$). Examples include case-hardening of mild steel and coloring of clear sapphires. Details of the geometry are then unimportant

TABLE 4.7.6 Solubility and Permeability of Gases in Solids

Gas	Solid	Temperature, K	$\mathcal{S}$ (m^3(STP)/m^3 atm) or $\mathcal{S}'$[a]	Permeability[b] m^3(STP)/m^2 sec (atm/m)
H_2	Vulcanized rubber	300	$\mathcal{S} = 0.040$	0.34×10^{-10}
	Vulcanized neoprene	290	$\mathcal{S} = 0.051$	0.053×10^{-10}
	Silicone rubber	300		4.2×10^{-10}
	Natural rubber	300		0.37×10^{-10}
	Polyethylene	300		0.065×10^{-10}
	Polycarbonate	300		0.091×10^{-10}
	Fused silica	400	$\mathcal{S}' \simeq 0\ 035$	
		800	$\mathcal{S}' \simeq 0.030$	
	Nickel	360	$\mathcal{S}' = 0.202$	
		440	$\mathcal{S}' = 0.192$	
He	Silicone rubber	300		2.3×10^{-10}
	Natural rubber	300		0.24×10^{-10}
	Polycarbonate	300		0.11×10^{-10}
	Nylon 66	300		0.0076×10^{-10}
	Teflon	300		0.047×10^{-10}
	Fused silica	300	$\mathcal{S}' \simeq 0.018$	
		800	$\mathcal{S}' \simeq 0.026$	
	Pyrex glass	300	$\mathcal{S}' \simeq 0.006$	
		800	$\mathcal{S}' \simeq 0.024$	
	7740 glass	470	$\mathcal{S} = 0.0084$	4.6×10^{-13}
	(94% SiO_2 + B_2O_3 + P_2O_5,	580	$\mathcal{S} = 0.0038$	1.6×10^{-12}
	5% Na_2O + Li_2 + K_2O,	720	$\mathcal{S} = 0.0046$	6.4×10^{-12}
	1% other oxides)			
	7056 glass	390	$\mathcal{S}' = 0.0039$	1.2×10^{-14}
	(90% SiO_2 + B_2O_3 + P_2O_5,	680	$\mathcal{S}' = 0.0059$	1.0×10^{-12}
	8% Na_2O + Li_2 + K_2O,			
	1% PbO, 5% other oxides)			
O_2	Vulcanized rubber	300	$\mathcal{S} = 0.070$	0.15×10^{-10}
	Silicone rubber	300		3.8×10^{-10}
	Natural rubber	300		0.18×10^{-10}
	Polyethylene	300		4.2×10^{-12}
	Polycarbonate	300		0.011×10^{-10}
	Silicone-polycarbonate copolymer (57% silicone)	300		1.2×10^{-10}
	Ethyl cellulose	300		0.09×10^{-10}
N_2	Vulcanized rubber	300	$\mathcal{S} = 0.035$	0.054×10^{-10}
	Silicone rubber	300		1.9×10^{-12}
	Natural rubber	300		0.062×10^{-10}
	Silicone-polycarbonate copolymer (57% silicone)	300		0.53×10^{-10}
	Teflon	300		0.019×10^{-10}
CO_2	Vulcanized rubber	300	$\mathcal{S} = 0.090$	1.0×10^{-10}
	Silicone rubber	290		21×10^{-10}
	Natural rubber	300		1.0×10^{-10}
	Silicone-polycarbonate copolymer (57% silicone)	300		7.4×10^{-10}
	Nylon 66	300		0.0013×10^{-10}
H_2O	Silicone rubber	310		$0.91–1.8 \times 10^{-10}$
Ne	Fused silica	300–1200	$\mathcal{S} \simeq 0.002$	
Ar	Fused silica	900–1200	$\mathcal{S} \simeq 0.01$	

TABLE 4.7.6 Solubility and Permeability of Gases in Solids

Gas	Solid	Temperature, K	$\mathcal{S}$ (m^3(STP)/m^3 atm) or $\mathcal{S}'$[a]	Permeability[b] m^3(STP)/m^2 sec (atm/m)

[a] Solubility $\mathcal{S}$ = volume of solute gas (0°C, 1 atm) dissolved in unit volume of solid when the gas is at 1 atm partial pressure. Solubility coefficient $\mathcal{S}' = c_{1,u}/c_{1,s}$.

[b] Permeability $\mathcal{P}_{12} = \mathcal{D}_{12}\mathcal{S}$.

From various sources, including Geankoplis, C.J. 1993. *Transport Processes and Unit Operations*, 3rd ed., Prentice-Hall; Englewood Cliffs, N.J.; Doremus, R.H. 1973. *Glass Science*, Wiley, New York; Altemose, V.O. 1961. Helium diffusion through glass, *J. Appl. Phys.*, 32, 1309–1316. With permission.

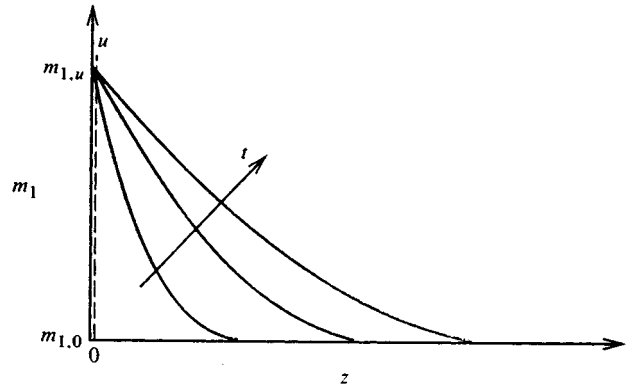


FIGURE 4.7.4 Transient diffusion in a plane slab.

and semi-infinite solid model can be used (Figure 4.7.4). For an initial concentration $m_{1,0}$ and a u-surface concentration suddenly changed to $m_{1,u}$ at time $t = 0$, the concentration distribution $m_1(z,t)$ is

$$\frac{m_1 - m_{1,0}}{m_{1,u} - m_{1,0}} = erfc\frac{z}{\left(4\mathcal{D}_{12}t\right)^{1/2}} \tag{4.7.38}$$

and the dissolution rate is

$$\dot{m}_1 = j_{1,u}A = \rho A\left(\frac{\mathcal{D}_{12}}{\pi t}\right)^{1/2}\left(m_{1,u} - m_{1,0}\right) \text{ kg/sec} \tag{4.7.39}$$

For example, consider a Pyrex glass slab at 800 K suddenly exposed to helium at 10^4 Pa. The molar equivalent to Equation (4.7.39) for an assumed constant solid phase molar concentration c is

$$\frac{\dot{M}_1}{A} = \left(\frac{\mathcal{D}_{12}}{\pi t}\right)^{1/2}\left(c_{1,u} - c_{1,0}\right)$$

From Table 4.7.6, $\mathcal{S}' = c_{1,u}/c_{1,s} \cong 0.024$; hence, $c_{1,u} = (0.024)(10^4)/(8314)(800) = 3.61 \times 10^{-5}$ kmol/m^3. From Table 4.7.4, $\mathcal{D}_{12} = 4.76 \times 10^{-8} \exp[-(2.72 \times 10^4)(10^3)/(8314)(800)] = 7.97 \times 10^{-10}$ m^2/sec. Hence,

$$\frac{\dot{M}_1}{A} = \left(\frac{7.97\times10^{-10}}{\pi t}\right)^{1/2}\left(3.61\times10^{-5} - 0\right) = 5.75\times10^{-10}/t \text{ kmol/sec}$$

Transient Diffusion in Slabs, Cylinders, and Spheres

Transient heat conduction in slates, cylinders, and spheres with surface convection is dealt with in Section 4.1. The analogous mass diffusion problem for the slab $-L < z < L$ is now considered. On a molar basis the governing differential equation is

$$\frac{\partial x_1}{\partial t} = \mathcal{D}_{12} \frac{\partial^2 x_1}{\partial z^2} \tag{4.7.40}$$

with initial condition $x_1 = x_{1,0}$ at $t = 0$. Boundary conditions are $\partial x_1/\partial z = 0$ at $z = 0$, and at the surface $z = L$,

$$-c\mathcal{D}_{12} \left.\frac{\partial x_1}{\partial z}\right|_{z=L} = \mathcal{G}_{m1}\left(y_{1,s} - y_{1,e}\right) \tag{4.7.41}$$

The convective boundary condition is of the same form as Newton's law of cooling, and defines the mole transfer conductance $\mathcal{G}_{m1}$ (kmol/m²sec) (see also the section on mass and mole transfer conductances). Also, we have followed chemical engineering practice and denoted mole fraction x in the solid (or liquid) phase and y in the liquid (or gas) phase, to emphasize that generally mole fraction is not continuous across a phase interface. For example, consider absorption of a sparingly soluble gas into a liquid for which Henry's law, Equation (4.7.13), applies: then $y_{1,s} = \text{He}x_{1,u}$.

In using heat conduction charts for mass diffusion problems, particular care must be taken with the evaluation of the Biot number. For heat conduction $\text{Bi} = h_c L/k$, where k is the solid conductivity. For mass diffusion the Biot number accounts for the discontinuity in concentration across the phase interface. Using gas absorption into a plane layer of liquid, for example, when Equation (4.7.41) is put into an appropriate dimensionless form, the mass transfer Biot number is seen to be

$$\text{Bi}_m = \frac{\mathcal{G}_{m1}\text{He}L}{c\mathcal{D}_{12}} \tag{4.7.42}$$

For sparingly soluble gases, e.g., O_2 or CO_2 in water, He, and hence Bi_m, are very large, and the absorption process is liquid-side controlled; that is, a uniform gas-phase composition can be assumed. Often interface equilibrium data are in graphical or tabular form; then an effective Biot number at the concentration of concern must be used.

For example, consider a 2-mm-diameter droplet of water at 300 K entrained in an air flow at 1 atm pressure containing 1% by volume CO_2. From Table 4.7.5, He = C_{He}= 1710. The liquid phase molar density can be approximated by the pure water value of $c = \rho/M = 996/18 = 55.3$ kmol/m³. The liquid phase diffusion coefficient is obtained from Table 4.7.4 as $\mathcal{D}_{12} = \nu_{H_2O}/Sc_{12} = 0.87 \times 10^{-6}/420 = 2.07 \times 10^{-9}$ m²/sec. For negligible relative motion between the droplet and gas, the Sherwood number (see the section on dimensionless groups) is approximately 2.0, and hence the gas phase mole transfer conductance is $\mathcal{G}_{m1} = 2c\mathcal{D}_{12}/\mathcal{D}$. For the gas phase, the molar density $c = \mathcal{P}/\mathcal{R}T = (1.0133 \times 10^5)/(8314)(300) = 0.0406$ kmol/m³ and $\mathcal{D}_{12} = 0.157 \times 10^{-4}$ m²/sec from Table 4.7.2. Thus,

$$\mathcal{G}_{m1} = \frac{(2)(0.0406)\left(0.157 \times 10^{-4}\right)}{(0.002)} = 6.37 \times 10^{-4} \text{ kmol/m}^2\,\text{sec}$$

From Equation 4.7.42 with $L = R$ the droplet radius, the mass transfer Biot number is

$$\text{Bi}_m = \frac{\left(6.37 \times 10^{-4}\right)(1710)(0.001)}{(55.3)\left(2.07 \times 10^{-9}\right)} = 9520$$

Thus, even for a small droplet with a relatively large gas-side mole transfer conductance, the absorption process is liquid-side controlled.

Diffusion in a Porous Catalyst

Porous catalysts are used to give a large surface area per unit volume of catalyst surface. Current practice for automobile catalytic convertors is to use a ceramic matrix as a support for a thin porous alumina layer that is impregnated with the catalyst (called a *washcoat*). A typical matrix has passages of hydraulic diameter 1 mm, and the washcoat may be about 20 μm thick. Pore sizes are of the order of 1 μm for which ordinary and Knudsen diffusion resistances are important. A simple model for diffusion in a porous catalyst is

$$\boldsymbol{J}_1 = -c\mathcal{D}_{1,\text{eff}}\nabla x_1 \ \text{kmol}/\text{m}^2\,\text{sec} \tag{4.7.43}$$

where the subscript eff denotes an effective diffusivity that accounts for the presence of the solid material. Assuming additive resistances,

$$\frac{1}{\mathcal{D}_{1,\text{eff}}} = \frac{1}{\mathcal{D}_{12,\text{eff}}} + \frac{1}{\mathcal{D}_{K1,\text{eff}}} \tag{4.7.44}$$

and

$$\mathcal{D}_{12,\text{eff}} = \frac{\varepsilon_v}{\tau}\mathcal{D}_{12}; \qquad \mathcal{D}_{K1,\text{eff}} = \frac{\varepsilon_v}{\tau}\mathcal{D}_{K1,\text{eff}} \tag{4.7.45}$$

where ε_v is the volume void fraction and τ is the tortuosity factor (usually between 4 and 8). From the kinetic theory of gases the Knudsen diffusion coefficient is

$$\mathcal{D}_{K1} = 97 r_e \left(T/M_1\right)^{1/2} \ \text{m}^2/\text{sec} \tag{4.7.46}$$

for effective pore radius r_e in meters and T in kelvins.

When a chemical reaction takes place within a porous layer, a concentration gradient is set up, and surfaces on pores deep within the pellet are exposed to lower reactant concentrations than surfaces near the pore openings. For a first-order reaction it is straightforward to obtain the concentration distribution. The results of such an analysis are conveniently given in the form of an effectiveness η_p, which is defined as the actual consumption rate of the reactant divided by that for an infinite diffusion coefficient. For a layer of thickness L exposed to reactants on one side, as shown in Figure 4.7.5.

$$\eta_p = \frac{\tanh bL}{bL}; \qquad b = \left(\frac{k'' \text{a}_P}{\mathcal{D}_{1,\text{eff}}}\right)^{1/2} \tag{4.7.47}$$

where k'' (m/sec) is the rate constant for a first-order reaction and a_p (m^{-1}) is the catalyst area per unit volume. Notice that this effectiveness is analogous to the efficiency of a heat transfer fin.

For example, consider a 30-μm-thick porous alumina washcoat with a volume void fraction $\varepsilon_v = 0.8$, a tortuosity factor $\tau = 4.0$, average pore radius $r_e = 1$ μm, and catalytic surface area per unit volume $a_p = 7.1 \times 10^5$ cm^2/cm^3. For carbon monoxide oxidation by copper oxide at 800 K, 1 atm, the rate constant is approximately 4.2×10^{-4} m^2/sec. To calculate the effectiveness of the washcoat, we first need to calculate the effective diffusion coefficient $\mathcal{D}_{1,\text{eff}}$:

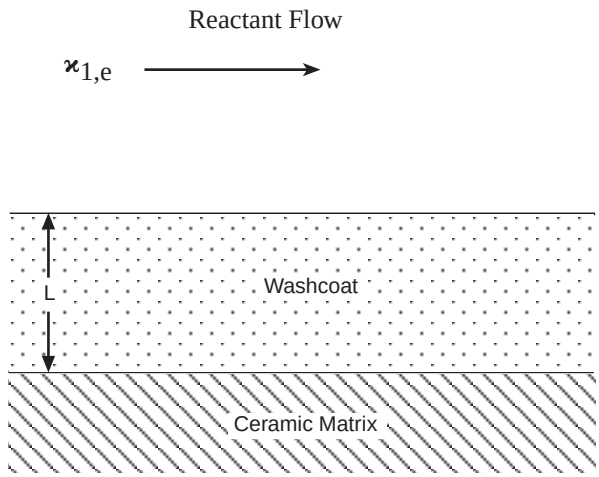


FIGURE 4.7.5 A catalyst layer.

$$\mathcal{D}_{12,\text{eff}} = \frac{\varepsilon_v}{\tau}\mathcal{D}_{12} = \frac{0.8}{4.0}\left(1.06\times10^{-4}\right) = 2.12\times10^{-5}\ \text{m}^2/\text{sec}$$

where $\mathcal{D}_{12}$ is approximated as the CO-air value from Table 4.7.2.

$$\mathcal{D}_{K1,\text{eff}} = \frac{\varepsilon_v}{\tau}\mathcal{D}_{12} = \frac{0.8}{4.0}(97)\left(1\times10^{-6}\right)(800/28)^{1/2} = 1.04\times10^{-4}\ \text{m}^2/\text{sec}$$

$$\frac{1}{\mathcal{D}_{1,\text{eff}}} = \frac{1}{2.12\times10^{-5}} + \frac{1}{1.04\times10^{-4}}; \qquad \mathcal{D}_{1,\text{eff}} = 1.76\times10^{-5}\ \text{m}^2/\text{sec}$$

$$b = \left[\frac{\left(4.2\times10^{-4}\right)\left(7.1\times10^{5}\right)\left(10^{2}\right)}{1.76\times10^{-5}}\right]^{1/2} = 4.2\times10^{4}\ \text{m}^{-1}; \qquad bL = \left(4.2\times10^{4}\right)\left(30\times10^{-6}\right) = 1.236$$

$$\eta_P = \frac{\tanh 1.236}{1.236} = 68.3\%$$

In an automobile catalytic convertor, Equation 4.7.47 applies to the catalyst washcoat. However, the mass transfer problem also involves a convective process for transport of reactants from the bulk flow. Referring to Figure 4.7.6 there are two mass transfer resistances in series, and the consumption rate of species 1 per unit surface area of the washcoat is

$$J_{1,s} = \frac{-x_{1,e}}{\dfrac{1}{L\eta_p k'' c} + \dfrac{1}{\mathcal{G}_{m1}}}\ \text{kmol}/\text{m}^2\ \text{sec} \tag{4.7.48}$$

where $\mathcal{G}_{m1}$ is the mole transfer conductance describing convective transport to the washcoat surface (see the section on mass and mole transfer conductances). Notice that when $\mathcal{G}_{m1} \ll L\eta_p k'' c$ the reaction rate is controlled by mass transfer from the gas stream to the washcoat surface; when $L\eta_p k'' c \ll \mathcal{G}_{m1}$, the reaction rate is controlled by diffusion within the washcoat and the kinetics of the reaction.

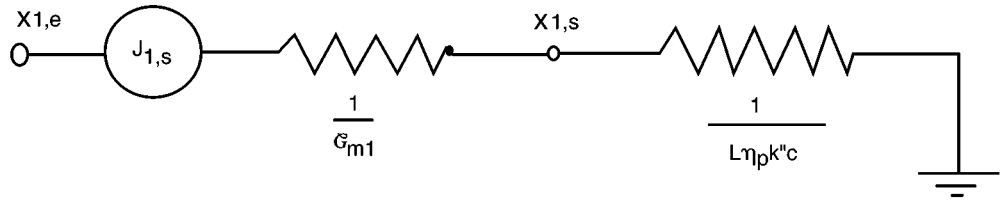


FIGURE 4.7.6 Equivalent circuit for mass transfer in an automobile catalytic convertor.

Diffusion in a Moving Medium

Net mass transfer across a surface results in a velocity component normal to the surface, and an associated convective flux in the direction of mass transfer. This convective flow is called a *Stefan flow*. The solutions of a number of mass transfer problems, involving a Stefan flow induced by the mass transfer process itself, follow. When necessary to obtain an analytical result, properties are assumed constant. Thus, use of these results requires evaluation of properties at a suitable reference state.

Diffusion with One Component Stationary

As an example, consider the simple heat pipe shown in Figure 4.7.7 with the evaporator and condenser located at the ends only (a bad design!). Then, if the working fluid is species 1, and a noncondensable gas is species 2, the concentration distribution is

$$\left(\frac{1-x_1}{1-x_{1,s}}\right)=\left(\frac{1-x_{1,e}}{1-x_{1,s}}\right)^{z/L} \tag{4.7.49}$$

and the vapor flux along the heat pipe is

$$N_1=\frac{c\mathcal{D}_{12}}{L}\ln\frac{1-x_{1,e}}{1-x_{1,s}}\ \text{kmol/m}^2\ \text{sec} \tag{4.7.50}$$

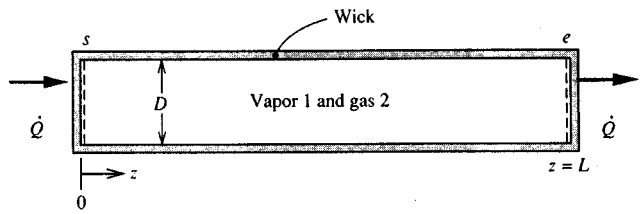


FIGURE 4.7.7 A simple heat pipe with the evaporator and condenser located at its ends.

Notice that $N_2 = 0$; that is, the gas is stationary. The rate of heat flow for a heat pipe of cross-sectional area of A_c is $\dot{Q} = N_1 M_1 h_{fg} A_c$. Evaluation of the $c\mathcal{D}$ product at a reference temperature $T_r = (1/2)(T_s + T_e)$ is adequate for most applications. Equation (4.7.50) applies to any situation where a one-dimensional model of mass transport is appropriate.

Heterogeneous Combustion

As an example, consider a small carbon particle entrained in a high-temperature airstream, as shown in Figure 4.7.8. The surface reaction is $2C + O_2 \rightarrow 2CO$ and there are no reactions in the gas phase. The stoichiometric ratio for the reaction is $r = 4/3$ kg oxygen/kg carbon. The reaction is diffusion controlled at the temperatures under consideration, that is, $m_{O_2,s} \simeq 0$. The mass transfer rate is n_s, which we give

the distinctive symbol $\dot{m}''$ since it is usually the desired result of an analysis; in this situation $\dot{m}'' = n_{C,u}$ is the combustion rate of carbon, and for a spherical particle of radius R is given by

$$\dot{m}'' = \frac{\rho \mathcal{D}_{O_2,m}}{R} \ln\left[1 + \frac{m_{O_2,e} - m_{O_2,s}}{m_{O_2,s} + 4/3}\right] = 0.160 \frac{\rho \mathcal{D}_{O_2,m}}{R} \text{ kg/m}^2 \text{ sec} \tag{4.7.51}$$

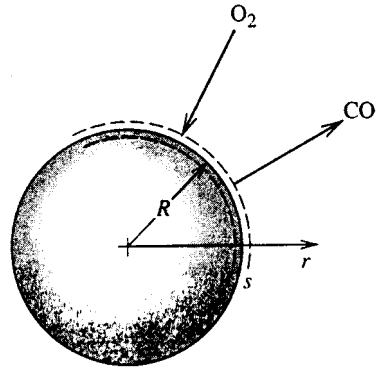


FIGURE 4.7.8 Combusion of a carbon particle in high-temperature air. The surface reaction is $2C + O_2 \rightarrow 2CO$.

The carbon particle temperature depends on its radius, and it is required to evaluate the property product $\rho\mathcal{D}$ at an appropriate reference temperature: an energy balance on the particle should be performed by this purpose. The resulting particle lifetime τ is

$$\tau = \frac{\rho_{\text{solid}} D_O^2}{1.28\left(\rho \mathcal{D}_{O_2,m}\right)_r} \text{ sec} \tag{4.7.52}$$

for an initial particle diameter of D_0. Air properties at an average mean film temperature can be used to evaluate $\rho\mathcal{D}_{O_2,m}$.

Consider a 10-μm-diameter carbon particle ignited in an airstream at 1500 K and 1 atm. An energy balance on the particle (including radiation to surroundings at 1500 K) shows that the average temperature of the particle is approximately 2550 K, and, thus, $T_r = (1/2)(1500 + 2550) = 2025$ K or $\rho \simeq \rho_{\text{air}} = 0.175$ kg/m^3 and $\mathcal{D}_{O_2,m} \simeq \mathcal{D}_{O_2,\text{air}} = 4.89 \times$ m^2/sec (from Table 4.7.1). Then

$$\tau = \frac{(1810)\left(10 \times 10^{-6}\right)^2}{(1.28)(0.175)\left(4.89 \times 10^{-4}\right)} = 1.65 \times 10^{-3} \text{ sec}$$

Droplet Evaporation

Consider a small droplet of species 1 entrained in a gas stream, species 2 (Figure 4.7.9). This is a simultaneous heat and mass transfer problem, and the mass transfer rate can be obtained by solving simultaneously

$$\dot{m}'' = \frac{\rho \mathcal{D}_{12}}{R} \ln\left(1 + \frac{m_{1,e} - m_{1,s}}{m_{1,s} - 1}\right) = \frac{k/c_{p1}}{R} \ln\left(1 + \frac{c_{p1}\left(T_e - T_s\right)}{h_{fg}}\right) \text{ kg/m}^2 \text{ sec} \tag{4.7.53a}$$

$$m_{1,s} = m_{1,s}(T, P) \quad \left(\text{from vapor–pressure data}\right) \tag{4.7.53b}$$

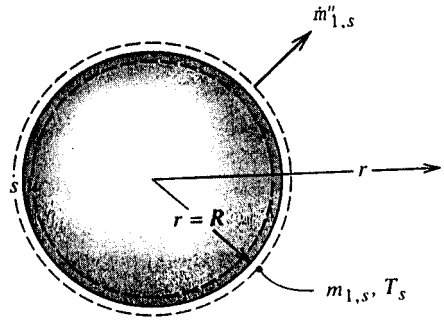


FIGURE 4.7.9 Evaporation of a droplet.

Temperature T_s is the adiabatic vaporization temperature and is essentially the psychrometric wet-bulb temperature. Properties can be evaluated at mean film temperature and composition; alternatively, c_{p1} can be set equal to the reference specific heat and all properties evaluated using Hubbard's $^1/_3$ rule, namely,

$$m_{1,r} = m_{1,s} + (1/3)\left(m_{1,e} - m_{1,s}\right) \tag{4.7.54a}$$

$$T_r = T_s + (1/3)\left(T_e - T_s\right) \tag{4.7.54b}$$

Droplet Combustion

Figure 4.7.10 shows a schematic of a volatile liquid hydrocarbon fuel droplet burning in air at zero gravity. The flame diameter is typically four to six times the droplet diameter. Heat is transferred from the flame to the droplet and serves to vaporize the fuel. In the flame the vapor reacts with oxygen to form gaseous products, primarily CO_2 and H_zO. When a fuel droplet ignites, there is a short initial transient during which the droplet heats up, until further conduction into the droplet is negligible and the droplet attains a steady temperature (approximately the wet-bulb temperature, which is very close to the boiling point for a typical hydrocarbon fuel). The reaction in the flame can be modeled as a single-step reaction with a constant stoichiometric ratio, r, and heat of combustion Δh_c J/kg of fuel.

The burning (mass transfer) rate of the droplet is given by the Godsave–Spalding formula,

$$\dot{m}'' = \frac{k/c_p}{R}\ln[1+\mathcal{B}]\ \text{kg/m}^2\ \text{sec} \tag{4.7.55}$$

where

$$\mathcal{B} = \frac{m_{ox,e}\,\Delta h_c/r + c_p\left(T_e - T_s\right)}{h_{fg}}$$

is the *mass transfer driving force* (or *transfer number*). The droplet lifetime is then

$$\tau = \frac{\rho_l D_O^2}{8\left(k/c_p\right)\ln(1+\mathcal{B})}\ \text{sec} \tag{4.7.56}$$

Based on experimental data for alkane droplets burning in air, Law and Williams (1972) recommend that properties be evaluated at a reference temperature $T_r = (1/2)(T_{BP} + T_{\text{flame}})$ where T_{flame} is the adiabatic flame temperature. The reference specific heat is $c_{pr} = c_{pfu}$, and the reference thermal conductivity is k_r

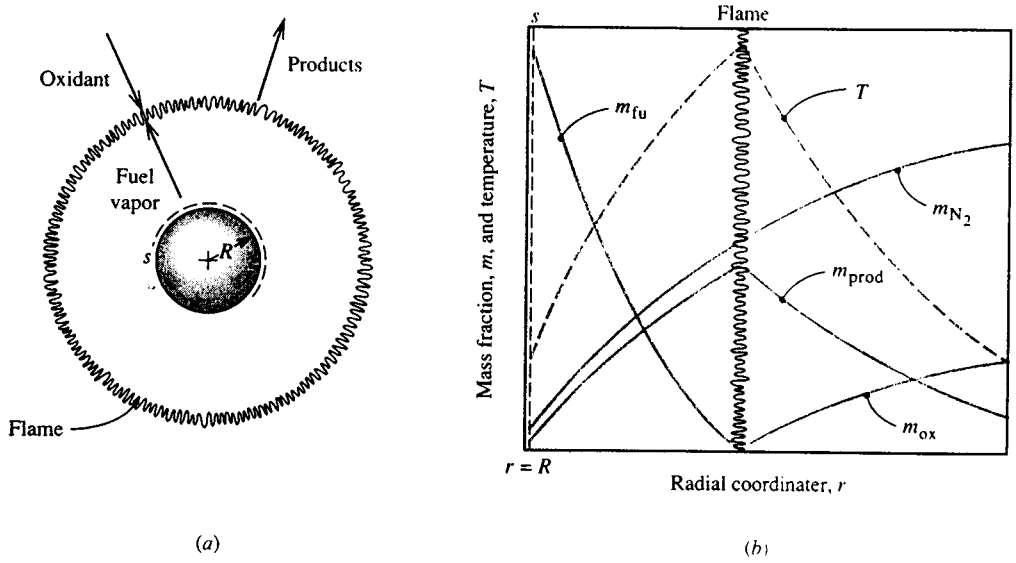


FIGURE 4.7.10 Combustion of a volatile fuel droplet burning in air: (a) schematic showing the flame, (b) concentration and temperature profiles.

$= 0.4k_{fu} + 0.6k_{\text{air}}$. Radiation has been ignored in the analysis leading to Equation (4.7.55) but is accounted for in using the Law and Williams reference-property scheme.

For example, consider a 1-mm-diameter *n*-octane droplet burning in air at 1 atm and 300 K, at near zero gravity. For *n*-octane (n-C_8H_{18}), $\rho_l = 611$ kg/m^3, $h_{fg} = 3.03 \times 10^5$ J/kg, $\Delta h_c = 4.44 \times 10^7$ J/kg, and $T_{BP} = 399$ K. The flame temperature is $T_{\text{flame}} = 2320$ K. At a reference temperature of (1/2) ($T_{\text{flame}} + T_{BP}$) = 1360 K, property values of *n*-octane vapor include $k = 0.113$ W/m K, $c_p = 4280$ J/kg K. The reaction is

$$C_8H_{18} + 12.5O_2 \rightarrow 8CO_2 + 9H_2O$$

Hence, the stoichiometric ratio $r = 400/114.2 = 3.50$. Also $m_{ox,e} = 0.231$ and $T_s \cong T_{BP} = 399$ K. Thus, the transfer number is

$$\mathcal{B} = \frac{(0.231)\left(4.44 \times 10^7\right)/(3.50) + 4280(300 - 399)}{3.03 \times 10^5} = 8.27$$

At $T_r = 1360$ K, $k_{\text{air}} = 0.085$ W/m K. Hence,

$$k_r = 0.4k_{fu} + 0.6k_{\text{air}} = (0.4)(0.113) + (0.6)(0.085) = 0.096 \text{ W/m K}$$

and the droplet lifetime is

$$\tau = \frac{(611)\left(1 \times 10^{-3}\right)^2}{(8)(0.096/4280)\ln(1 + 8.27)} = 1.53 \text{ sec}$$

Mass Convection

The terms *mass convection* or *convective mass transfer* are generally used to describe the process of mass transfer between a surface and a moving fluid, as shown in Figure 4.7.11. The surface may be that

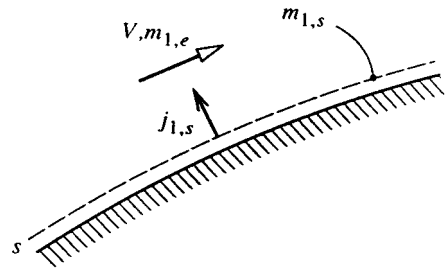


FIGURE 4.7.11 Notation for convective mass transfer into an external flow.

of a falling water film in an air humidifier, of a coke particle in a gasifier, or of a silica-phenolic heat shield protecting a reentry vehicle. As is the case for heat convection, the flow can be *forced* or *natural*, *internal* or *external*, and *laminar* or *turbulent*. In addition, the concept of whether the mass transfer rate is *low* or *high* plays an important role: when mass transfer rates are low, there is a simple analogy between heat transfer and mass transfer that can be efficiently exploited in the solution of engineering problems.

Mass and Mole Transfer Conductances

Analogous to convective heat transfer, the rate of mass transfer by convection is usually a complicated function of surface geometry and *s*-surface composition, the fluid composition and velocity, and fluid physical properties. For simplicity, we will restrict our attention to fluids that are either binary mixtures or solutions, or situations in which, although more than two species are present, diffusion can be adequately described using effective binary diffusion coefficients, as was discussed in the section on ordinary diffusion. Referring to Figure 4.7.11, we define the *mass transfer conductance* of species 1, g_{m1}, by the relation

$$j_{1,s} = g_{m1}\,\Delta m_1; \qquad \Delta m_1 = m_{1,s} - m_{1,e} \tag{4.7.57}$$

and the units of g_{m1} are seen to be the same as for mass flux (kg/m²sec). Equation (4.7.57) is of a similar form to Newton's law of cooling, which defines the heat transfer coefficient h_c. Why we should not use a similar name and notation (e.g., mass transfer coefficient and h_m) will become clear later. On a molar basis, we define the *mole transfer conductance* of species 1, G_{m1}, by a corresponding relation,

$$J_{1,s} = G_{m1}\,\Delta x_1; \qquad \Delta x_1 = x_{1,s} - x_{1,e} \tag{4.7.58}$$

where G_{m1} has units (kmol/m²sec).

Low Mass Transfer Rate Theory

Consider, as an example, the evaporation of water into air, as shown in Figure 4.7.12. The water–air interface might be the surface of a water reservoir, or the surface of a falling water film in a cooling tower or humidifier. In such situations the mass fraction of water vapor in the air is relatively small; the highest value is at the *s*-surface, but even if the water temperature is as high as 50°C, the corresponding value of $m_{H_2O,s}$ at 1 atm total pressure is only 0.077. From Equation 4.7.54 the driving potential for diffusion of water vapor away from the interface is $\Delta m_1 = m_{1,s} - m_{1,e}$, and is small compared to unity, even if the free-stream air is very dry such that $m_{1,e} \simeq 0$. We then say that the mass transfer rate is *low* and the rate of evaporation of the water can be approximated as $j_{1,s}$; for a surface area A,

$$\dot{m}_1 = \left(m_{1,s} n_s + j_{1,s}\right) A \simeq j_{1,s} A \text{ kg/sec} \tag{4.7.59}$$

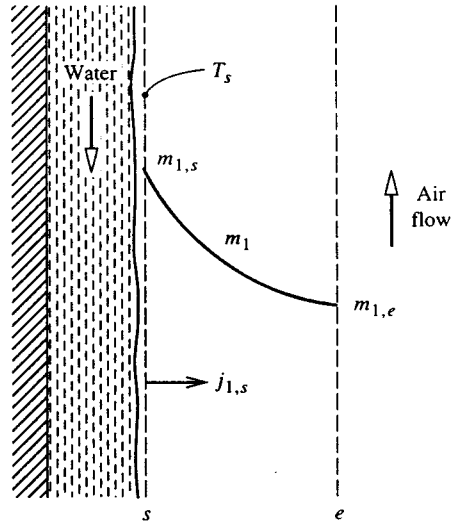


FIGURE 4.7.12 Evaporation of water into an air flow.

In contrast, if the water temperature approaches its boiling point, $m_{1,s}$ is no longer small, and of course, in the limit of $T_s = T_{BP}$, $m_{1,s} = 1$. The resulting driving potential for diffusion Δm_1 is then large, and we say that the mass transfer rate is *high*. Then, the evaporation rate cannot be calculated from Equation 4.7.59, as will be explained in the section on high mass transfer rate theory. For water evaporation into air, the error incurred in using low mass transfer rate theory is approximately ($^1/_2$) Δm_1, and a suitable criterion for application of the theory to engineering problems is $\Delta m_1 < 0.1$ or 0.2.

A large range of engineering problems can be adequately analyzed assuming low mass transfer rates. These problems include cooling towers and humidifiers as mentioned above, gas absorbers for sparingly soluble gases, and catalysis. In the case of catalysis, the *net* mass transfer rate is actually zero. Reactants diffuse toward the catalyst surface and the products diffuse away, but the catalyst only promotes the reaction and is not consumed. On the other hand, problems that are characterized by high mass transfer rates include condensation of steam containing a small amount of noncondensable gas, as occurs in most power plant condensers; combustion of volatile liquid hydrocarbon fuel droplets in diesel engines and oil-fired power plants, and ablation of phenolic-based heat shields on reentry vehicles.

Dimensionless Groups

Dimensional analysis of convective mass transfer yields a number of pertinent dimensionless groups that are, in general, analogous to dimensionless groups for convective heat transfer. The most important groups are as follows.

1. The Schmidt number, $Sc_{12} = \mu/\rho\mathcal{D}_{12}$, which is a properties group analogous to the Prandtl number. For gas mixtures, $Sc_{12} = O(1)$, and for liquid solutions, $Sc_{12} = O(100)$ to $O(1000)$. There are not fluids for which $Sc_{12} \ll 1$, as is the case of Prandtl number for liquid metals.
2. The Sherwood number (or mass transfer Nusselt number). $Sh = g_{m1}L/\rho\mathcal{D}_{12}$ ($= \mathcal{G}_{m1}L/c\mathcal{D}_{12}$) is a dimensionless conductance.
3. The mass transfer Stanton number $St_m = g_{m1}/\rho V$ ($= \mathcal{G}_{m1}/cV$) is an alternative dimensionless conductance.

As for convective heat transfer, forced convection flows are characterized by a Reynolds number, and natural convection flows are characterized by a Grashof or Rayleigh number. In the case of Gr or Ra it is not possible to replace $\Delta\rho$ by $\beta\Delta T$ since density differences can result from concentration differences (and both concentration and temperature differences for simultaneous heat and mass transfer problems).

Analogy between Convective Heat and Mass Transfer

A close analogy exists between convective heat and convective mass transfer owing to the fact that conduction and diffusion in a fluid are governed by physical laws of identical form, that is, Fourier's and Fick's laws, respectively. As a result, in many circumstances the Sherwood or mass transfer Stanton number can be obtained in a simple manner from the Nusselt number or heat transfer Stanton number for the same flow conditions. Indeed, in most gas mixtures Sh and St_m are nearly equal to their heat transfer counterparts. For dilute mixtures and solutions and low mass transfer rates, the rule for exploiting the analogy is simple: The *Sherwood or Stanton number is obtained by replacing the Prandtl number by the Schmidt number in the appropriate heat transfer correlation*. For example, in the case of fully developed turbulent flow in a smooth pipe

$$\mathrm{Nu}_D = 0.023\mathrm{Re}_D^{0.8}\mathrm{Pr}^{0.4}; \qquad \mathrm{Pr} > 0.5 \tag{4.7.60a}$$

which for mass transfer becomes

$$\mathrm{Sh}_D = 0.023\mathrm{Re}_D^{0.8}\mathrm{Sc}^{0.4}; \qquad \mathrm{Sc} > 0.5 \tag{4.7.60b}$$

Also, for natural convection from a heated horizontal surface facing upward,

$$\overline{\mathrm{Nu}} = 0.54\left(\mathrm{Gr}_L\mathrm{Pr}\right)^{1/4}; \qquad 10^5 < \mathrm{Gr}_L\mathrm{Pr} < 2\times10^7 \quad \text{(laminar)} \tag{4.7.61a}$$

$$\overline{\mathrm{Nu}} = 0.14\left(\mathrm{Gr}_L\mathrm{Pr}\right)^{1/3}; \qquad 2\times10^7 < \mathrm{Gr}_L\mathrm{Pr} < 3\times10^{10} \quad \text{(turbulent)} \tag{4.7.61b}$$

which for isothermal mass transfer with $\rho_s < \rho_e$ become

$$\overline{\mathrm{Sh}} = 0.54\left(\mathrm{Gr}_L\mathrm{Sc}\right)^{1/4}; \qquad 10^5 < \mathrm{Gr}_L\mathrm{Sc} < 2\times10^7 \quad \text{(laminar)} \tag{4.7.62a}$$

$$\overline{\mathrm{Sh}} = 0.14\left(\mathrm{Gr}_L\mathrm{Sc}\right)^{1/3}; \qquad 2\times10^7 < \mathrm{Gr}_L\mathrm{Sc} < 3\times10^{10} \quad \text{(turbulent)} \tag{4.7.62b}$$

With evaporation, the condition, $\rho_s < \rho_e$ will be met when the evaporating species has a smaller molecular weight than the ambient species, for example, water evaporating into air. Mass transfer correlations can be written down in a similar manner for almost all the heat transfer correlations given in Section 4.2. There are some exceptions: for example, there are no fluids with a Schmidt number much less than unity, and thus there are no mass transfer correlations corresponding to those given for heat transfer to liquid metals with $\mathrm{Pr} \ll 1$. In most cases it is important for the wall boundary conditions to be of analogous form, for example, laminar flow in ducts. A uniform wall temperature corresponds to a uniform concentration $m_{1,s}$ along the *s*-surface, whereas a uniform heat flux corresponds to a uniform diffusive flux $j_{1,s}$. In chemical engineering practice, the analogy between convective heat and mass transfer is widely used in a form recommended by Chilton and Colburn in 1934, namely, $\mathrm{St}_m/\mathrm{St} = (\mathrm{Sc}/\mathrm{Pr})^{-2/3}$. The Chilton-Colburn form is of adequate accuracy for most external forced flows but is inappropriate for fully developed laminar duct flows.

For example, air at 1 atm and 300 K flows inside a 3-cm-inside-diameter tube at 10 m/sec. Using pure-air properties the Reynolds number is $VD/\nu = (10)(0.03)/15.7 \times 10^{-6} = 1.911 \times 10^4$. The flow is turbulent. Using Equation (4.7.60b) with $\mathrm{Sc}_{12} = 0.61$ for small concentrations of H_2O in air,

$$\mathrm{Sh}_D = (0.023)\left(1.911\times10^4\right)^{0.8}(0.61)^{0.4} = 50.2$$

$$g_{m1} = \rho \mathcal{D}_{12} \mathrm{Sh}/D = \rho \nu \mathrm{Sh}/\mathrm{Sc}_{12} D = \frac{(1.177)\left(15.7 \times 10^{-6}\right)(50.2)}{(0.61)(0.03)} = 5.07 \times 10^{-2} \text{ kg/m}^2 \text{ sec}$$

Further insight into this analogy between convective heat and mass transfer can be seen by writing out Equations (4.7.60a) and (4.7.60b) as, respectively,

$$\frac{\left(h_c/c_p\right)D}{k/c_p} = 0.023 \mathrm{Re}_D^{0.8} \left(\frac{\mu}{k/c_p} \right)^{0.4} \tag{4.7.63a}$$

$$\frac{g_m D}{\rho \mathcal{D}_{12}} = 0.023 \mathrm{Re}_D^{0.8} \left(\frac{\mu}{\rho \mathcal{D}_{12}} \right)^{0.4} \tag{4.7.63b}$$

When cast in this form, the correlations show that the property combinations k/c_p and $\rho\mathcal{D}_{12}$ play analogous roles; these are *exchange coefficients* for heat and mass, respectively, both having units kg/m sec, which are the same as those for dynamic viscosity μ. Also, it is seen that the ratio of heat transfer coefficient to specific heat plays an analogous role to the mass transfer conductance, and has the same units (kg/m^2 sec). Thus, it is appropriate to refer to the ratio h_c/c_p as the *heat transfer conductance*, g_h, and for this reason we should not refer to g_m as the mass transfer *coefficient*.

Simultaneous Heat and Mass Transfer

Often problems involve simultaneous convective heat and mass transfer, for which the surface energy balance must be carefully formulated. Consider, for example, evaporation of water into air, as shown in Figure 4.7.13. With H_2O denoted as species 1, the steady-flow energy equation applied to a control volume located between the *u*- and *s*-surfaces requires that

$$\dot{m}\left(h_{1,s} - h_{1,u}\right) = A\left(q''_{\text{cond}} - q''_{\text{conv}} - q''_{\text{rad}}\right) \text{ W} \tag{4.7.64}$$

where it has been recognized that only species 1 crosses the *u*- and *s*-surfaces. Also, the water has been assumed to be perfectly opaque so that all radiation is emitted or absorbed between the *u*-surface and the interface.

If we restrict our attention to conditions for which low mass transfer rate theory is valid, we can write $\dot{m}/A \simeq j_{1,s} = g_{m1}\left(m_{1,s} - m_{1,e}\right)$. Also, we can then calculate the convective heat transfer as if there were no mass transfer, and write $q_{\text{conv}} = h_c(T_s - T_e)$. Substituting in Equation (4.7.64) with $q_{\text{conv}} = -k\partial T/\partial y|_u$, $h_{1,s} - h_{1,u} = h_{fg}$, and rearranging, gives

$$-k\left.\frac{\partial T}{\partial y}\right|_u = h_c\left(T_s - T_e\right) + g_{m1}\left(m_{1,s} - m_{1,e}\right)h_{fg} + q''_{\text{rad}} \text{ W/m}^2 \tag{4.7.65}$$

It is common practice to refer to the convective heat flux $h_c(T_s - T_e)$ as the *sensible* heat flux, whereas the term $g_{m1}\left(m_{1,s} - m_{1,e}\right)h_{fg}$ is called the *evaporative* or *latent* heat flux. Each of the terms in Equation 4.7.65 can be positive or negative, depending on the particular situation. Also, the evaluation of the conduction heat flux at the *u*-surface, $-k\partial T/\partial y|_u$, depends on the particular situation. Four examples are shown in Figure 4.7.13. For a water film flowing down a packing in a cooling tower (Figure 4.7.13b), this heat flux can be expressed in terms of convective heat transfer from the bulk water at temperature T_L to the surface of the film, $-k\partial T/\partial y|_u = h_{cL}(T_L - T_s)$. If the liquid-side heat transfer coefficient h_{cL} is large enough, we can simply set $T_s \simeq T_L$, which eliminates the need to estimate h_{cL}. The evaporation process is then *gas-side controlled*. Figure 4.7.13c shows film condensation from a steam-air mixture on the outside of a vertical tube. In this case we can write $k\partial T/\partial y|_u = U(T_s - T_c)$, where T_c is the coolant

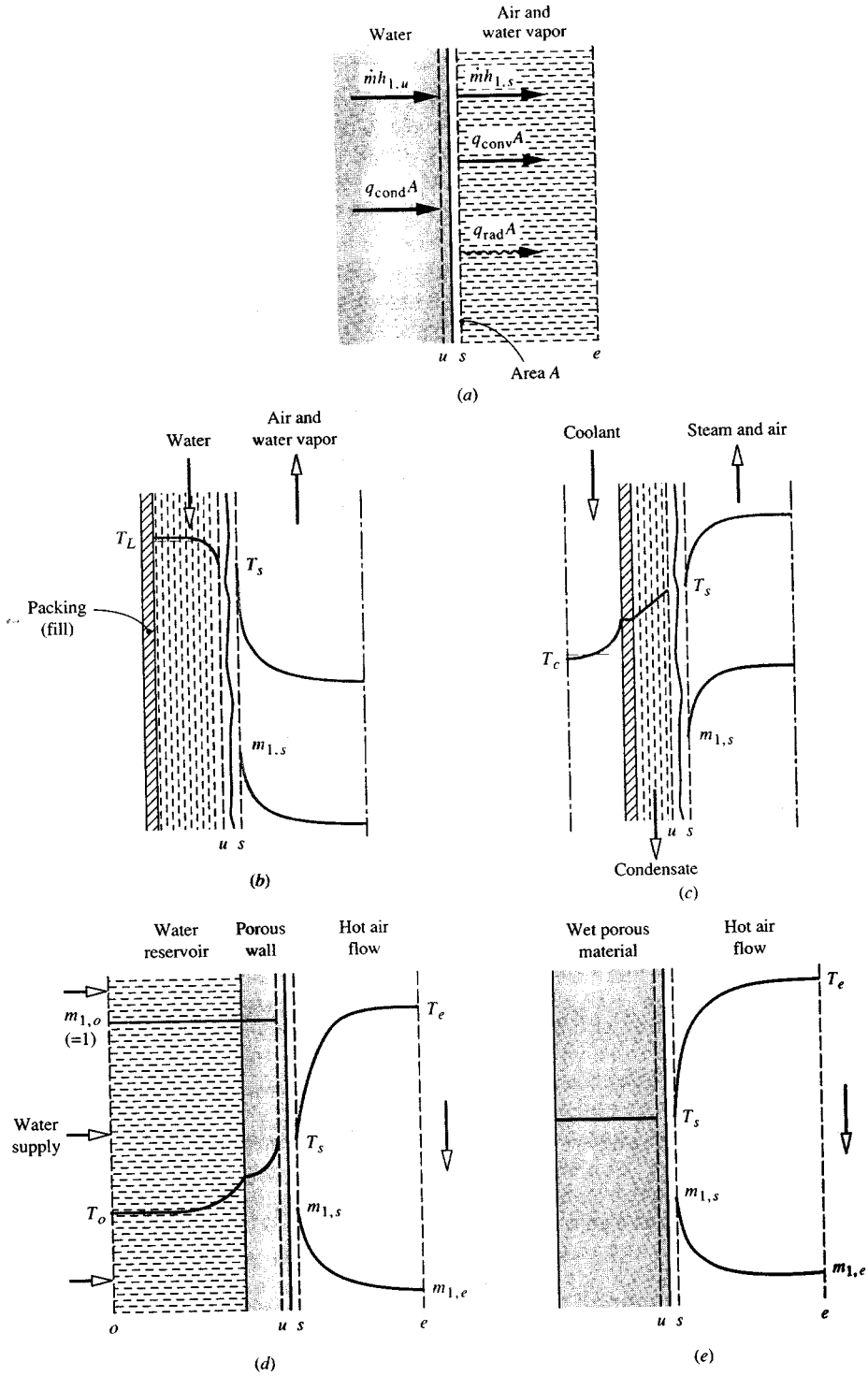


FIGURE 4.7.13 The surface energy balance for evaporation of water into an air stream.

bulk temperature. The overall heat transfer coefficient U includes the resistances of the condensate film, the tube wall, and the coolant. Sweat cooling is shown in Figure 4.7.13d, with water from a reservoir (or *plenum chamber*) injected through a porous wall at a rate just sufficient to keep the wall surface wet. In this case, the conduction across the u-surface can be related to the reservoir conditions by application of the steady-flow energy equation to a control volume located between the o- and u-surfaces. Finally, Figure 4.7.13e shows drying of a wet porous material (e.g., a textile or wood). During the constant-rate period of the process, evaporation takes place from the surface with negligible heat conduction into the solid; then $-k\partial T/\partial y|_u \simeq 0$. The term *adiabatic vaporization* is used to describe evaporation when $q_{\text{cond}} = 0$; constant-rate drying is one example, and the wet-bulb psychrometer is another.

Consider a 1-m-square wet towel on a washline on a day when there is a low overcast and no wind. The ambient air is at 21°C, 1 atm, and 50.5% RH. In the constant-rate drying period the towel temperature is constant, and $q_{\text{cond}} = 0$. An iterative calculation is required to obtain the towel temperature using correlations for natural convection on a vertical surface to obtain h_c and g_{m1}; q_{rad} is obtained as $q_{\text{rad}} = \sigma\varepsilon(T_s^4 - T_e^4)$ with $\varepsilon = 0.90$. The results are $T_s = 17.8$°C, $h_c = 1.69$ W/m^{2}K, $g_{m1} = 1.82 \times 10^{-3}$ kg/m^{2}sec, and the energy balance is

$$q_{\text{cond}} = h_c(T_s - T_e) + g_{m1}(m_{1,s} - m_{1,e})h_{fg} + q_{\text{rad}}$$

$$0 = -5.4 + 21.7 - 16.3 \text{ W/m}^2$$

Evaluation of composition-dependent properties, in particular the mixture specific heat and Prandtl number, poses a problem. In general, low mass transfer rates imply small composition variations across a boundary layer, and properties can be evaluated for a mixture of the free-stream composition at the mean film temperature. In fact, when dealing with evaporation of water into air, use of the properties of dry air at the mean film temperature gives results of adequate engineering accuracy. If there are large composition variations across the boundary layer, as can occur in some catalysis problems, properties should be evaluated at the mean film composition and temperature.

The Wet- and Dry-Bulb Psychrometer

The wet- and dry-bulb psychrometer is used to measure the moisture content of air. In its simplest form, the air is made to flow over a pair of thermometers, one of which has its bulb covered by a wick whose other end is immersed in a small water reservoir. Evaporation of water from the wick causes the wet bulb to cool and its steady-state temperature is a function of the air temperature measured by the dry bulb and the air humidity. The wet bulb is shown in Figure 4.7.14. In order to determine the water vapor mass fraction $m_{1,e}$, the surface energy balance Equation (4.7.66) is used with conduction into the wick and q''_{rad} set equal to zero. The result is

$$m_{1,e} = m_{1,s} - \frac{c_p}{h_{fg}}\left(\frac{\text{Pr}}{\text{Sc}_{12}}\right)^{-2/3}(T_e - T_s) \tag{4.7.66}$$

Usually $m_{1,e}$ is small and we can approximate $c_p = c_{p\ \text{air}}$ and $(\text{Pr}/\text{Sc}_{12})^{-2/3} = 1/1.08$. Temperatures T_s and T_e are the known measured wet- and dry-bulb temperatures. With T_s known, $m_{1,s}$ can be obtained using steam tables in the usual way. For example, consider an air flow at 1000 mbar with measured wet- and dry-bulb temperatures of 305.0 and 310.0 K, respectively. Then $P_{1,s} = P_{\text{sat}}(T_s) = P_{\text{sat}}(305.0\text{ K}) = 4714$ Pa from steam tables. Hence, $x_{1,s} = P_{1,s}/P = 4714/10^5 = 0.04714$, and

$$m_{1,s} = \frac{0.04714}{0.04714 + (29/18)(1 - 0.04714)} = 0.02979$$

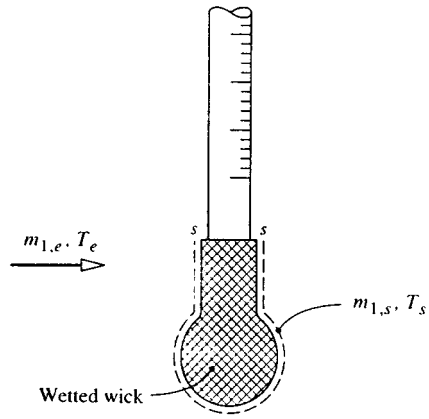


FIGURE 4.7.14 Wet bulb of a wet- and dry-bulb psychrometer.

Also, h_{fg} (305 K) = 2.425×10^6 J/kg, and $c_{p\ \text{air}}$ = 1005 J/kg K; thus

$$m_{1,e} = 0.02979 - \frac{1005}{(1.08)\left(2.425\times10^6\right)}(310-305) = 0.02787$$

$$x_{1,e} = \frac{0.02787}{0.02787 + (18/29)(1-0.02787)} = 0.04415$$

$$P_{1,e} = x_{1,e}P = (0.04415)\left(10^5\right) = 4412 \text{ Pa}$$

By definition, the relative humidity is RH = $P_{1,e}/P_{\text{sat}}(T_e)$; RH = 4415/6224 = 70.9%.

In the case of other adiabatic vaporization processes, such as constant-rate drying or evaporation of a water droplet, $m_{1,e}$ and T_e are usually known and Equation (4.7.66) must be solved for T_s. However, the thermodynamic wet-bulb temperature obtained from psychrometric charts or software is accurate enough for engineering purposes.

High Mass Transfer Rate Theory

When there is net mass transfer across a phase interface, there is a convective component of the absolute flux of a species across the *s*-surface. From Equation (4.7.23a) for species 1,

$$n_{1,s} = \rho_{1,s}v_s + j_{1,s} \text{ kg/m}^2 \text{ sec} \tag{4.7.67}$$

During evaporation the convection is directed in the gas phase, with a velocity normal to the surface v_s. When the convective component cannot be neglected, we say that the mass transfer rate is *high*. There are two issues to consider when mass transfer rates are high. First, the rate at which species 1 is transferred across the *s*-surface is not simply the diffusive component $j_{1,s}$ as assumed in low mass transfer rate theory, but is the sum of the convective and diffusive components shown in Equation 4.7.67. Second, the normal velocity component v_s has a *blowing* effect on the concentration profiles, and hence on the Sherwood number. The Sherwood number is no longer analogous to the Nusselt number of conventional heat transfer correlations, because those Nusselt numbers are for situations involving impermeable surfaces, e.g., a metal wall, for which v_s = 0.

Substituting for $j_{1,s}$ from Equation (4.7.57) into Equation (4.7.67) gives

$$\dot{m}'' = g_{m1}\frac{m_{1,e} - m_{1,s}}{m_{1,s} - n_{1,s}/\dot{m}''} = g_{m1}\mathcal{B}_{m1} \tag{4.7.68}$$

where $\dot{m}'' = n_s$ is the mass transfer rate introduced in the section on heterogeneous combustion and $\mathcal{B}_{m1}$ is the *mass transfer driving force*. In the special case where only species 1 is transferred, $n_{1,s}/\dot{m}'' = 1$, for example, when water evaporates into air, and dissolution of air in the water is neglected. It is convenient to rewrite Equation (4.7.68) as

$$\dot{m}'' = g^*_{m1}\left(g_{m1}/g^*_{m1}\right)\mathcal{B}_{m1} \text{ kg/m}^2 \text{ sec} \tag{4.7.69a}$$

where

$$g^*_{m1} = \lim_{\mathcal{B}_m \to 0} g_{m1} \tag{4.7.69b}$$

Now g^*_{m1} is the limit value of g_{m1} for zero mass transfer (i.e., $v_s = 0$), and Sh* can be obtained from conventional heat transfer Nusselt number correlations for impermeable surfaces. The ratio (g_{m1}/g^*_{m1}) is termed a *blowing factor* and accounts for the effect of v_s on the concentration profiles. Use of Equation (4.7.69) requires appropriate data for the blowing factor. For the constant-property laminar boundary layer on a flat plate, Figure 4.7.15 shows the effect of the Schmidt number on the blowing factor. The abscissa is a *blowing parameter* $B_m = \dot{m}''/g^*_m$.

The blowing velocity also affects the velocity and temperature profiles, and hence the wall shear stress and heat transfer. The curve for Sc = 1 in Figure 4.7.15 also gives the effect of blowing on shear stress as τ_s/τ^*_s, and the curve for Sc = 0.7 gives the effect of blowing on heat transfer for air injection into air as h_c/h^*_c (since Pr = 0.7 for air).

Variable Property Effects of High Mass Transfer Rates

High mass transfer rate situations are usually characterized by large property variations across the flow, and hence property evaluation for calculating g_m and h_c is not straightforward. An often-encountered situation is transfer of a single species into an inert laminar or turbulent boundary layer flow. The effect of variable properties can be very large as shown in Figure 4.7.16 for laminar boundary layers, and Figure 4.7.17 for turbulent boundary layers.

A simple procedure for correlating the effects of flow type and variable properties is to use weighting factors in the exponential functions suggested by a constant-property Couette-flow model (Mills, 1995). Denoting the injected species as species *i*, we have

$$\frac{g_{m1}}{g^*_{mi}} = \frac{a_{mi}B_{mi}}{\exp\left(a_{mi}B_{mi}\right) - 1}; \qquad B_{mi} = \frac{\dot{m}''}{g^*_{mi}} \tag{4.7.70a}$$

or

$$\frac{g_{m1}}{g^*_{mi}} = \frac{\ln\left(1 + a_{mi}\mathcal{B}_{mi}\right)}{a_{mi}\mathcal{B}_{mi}}; \qquad \mathcal{B}_{mi} = \frac{\dot{m}''}{g_{mi}} = \frac{m_{i,e} - m_{i,e}}{m_{i,s} - 1}$$

$$\frac{\tau_s}{\tau^*_s} = \frac{a_{fi}B_f}{\exp\left(a_{fi}B_f\right) - 1}; \qquad B_f = \frac{\dot{m}''u_e}{\tau^*_s} \tag{4.7.70b}$$

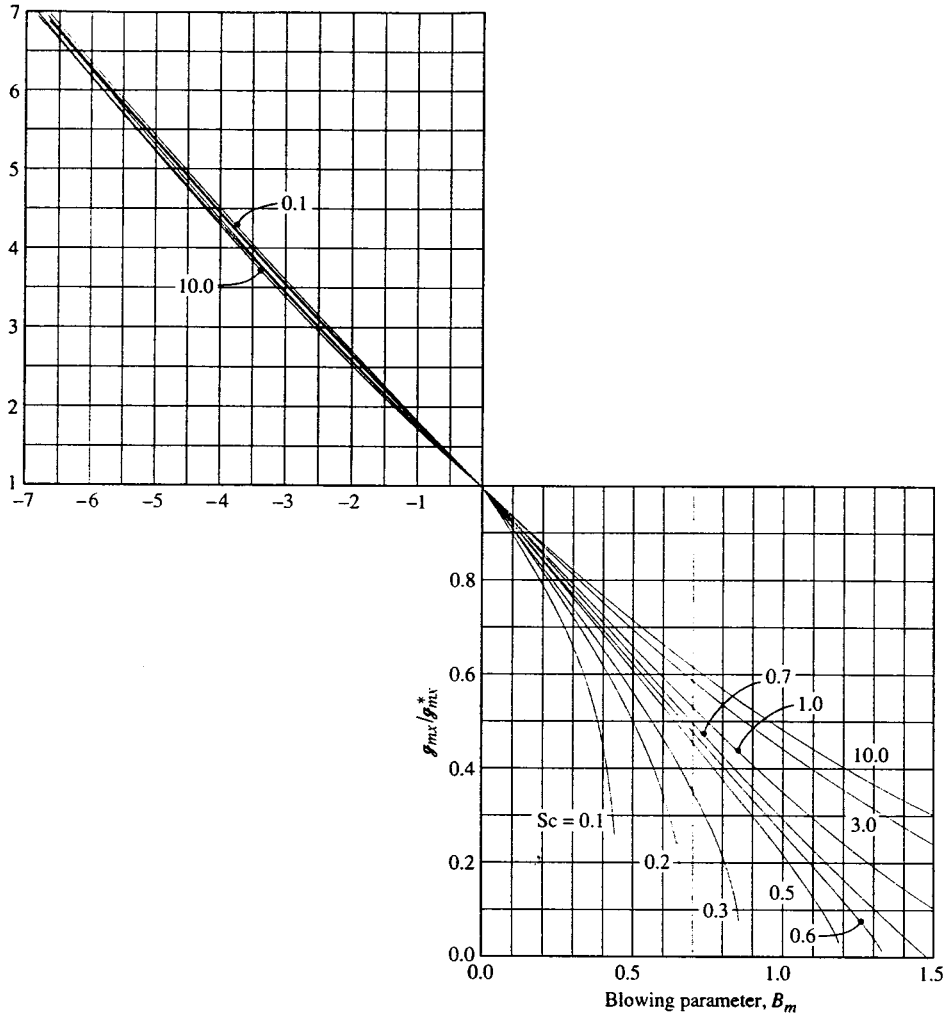


FIGURE 4.7.15 Effect of mass transfer on the mass transfer conductance for a laminar boundary layer on a flat plate: g_m / g^*_m vs. blowing parameter $B_m = \dot{m}'' / g^*_m$.

$$\frac{h_c}{h_c^*} = \frac{a_{hi} B_h}{\exp(a_{hi} B_h) - 1}; \qquad B_h = \frac{\dot{m}'' c_{pe}}{h_c^*} \tag{4.7.70c}$$

Notice that g^*_{mi}, τ_s^*, h_c^*, and c_{pe} are evaluated using properties of the free-stream gas at the mean film temperature. The weighting factor a may be found from exact numerical solutions of boundary layer equations or from experimental data. Some results for laminar and turbulent boundary layers follow.

1. *Laminar Boundary Layers.* We will restrict our attention to low-speed air flows, for which viscous dissipation and compressibility effects are negligible, and use exact numerical solutions of the self-similar laminar boundary layer equations (Wortman, 1969). Least-squares curve fits of the numerical data were obtained using Equations (4.7.70a) to (4.7.70c). Then, the weighting factors for axisymmetric stagnation-point flow with a cold wall (T_s/T_e = 0.1) were correlated as

$$a_{mi} = 1.65 \left(M_{\text{air}} / M_i \right)^{10/12} \tag{4.7.71a}$$

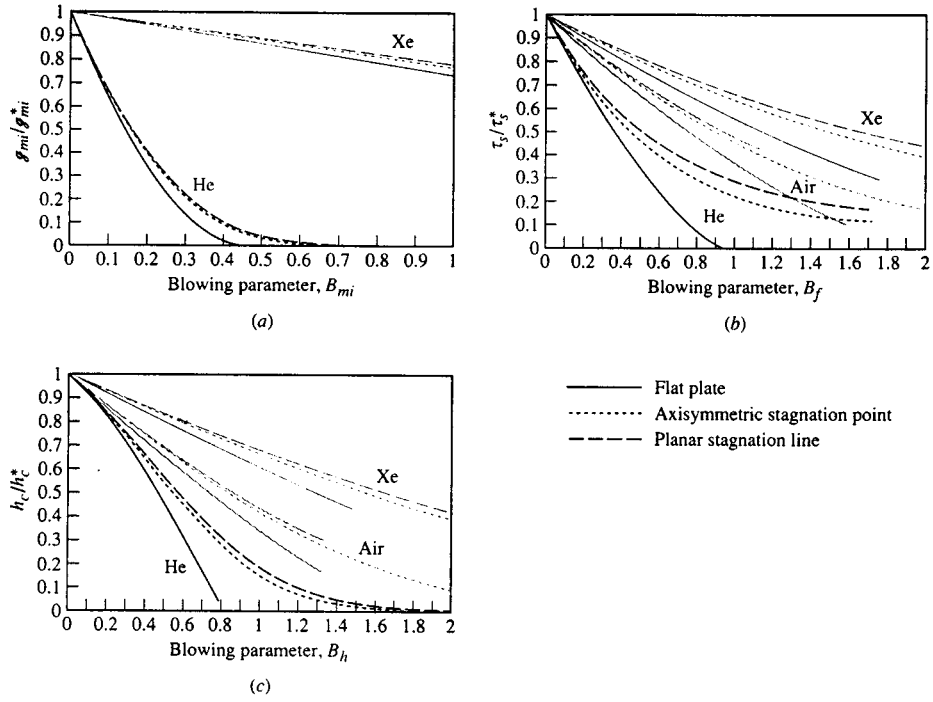


FIGURE 4.7.16 Numerical results for the effect of pressure gradient and variable properties on blowing factors for laminar boundary layers: low-speed air flow over a cold wall ($T_s/T_e = 0.1$) with foreign gas injection: (a) mass transfer conductance, (b) wall shear stress, (c) heat transfer coefficient. (From Wortman, A., Ph.D. dissertation, University of California, Los Angeles, 1969. With permission.)

$$a_{fi} = 1.38\left(M_{\text{air}}/M_i\right)^{5/12} \tag{4.7.71b}$$

$$a_{hi} = 1.30\left(M_{\text{air}}/M_i\right)^{3/12}\left[c_{pi}/\left(2.5\mathcal{R}/M_i\right)\right] \tag{4.7.71c}$$

Notice that $c_{pi}/2.5\mathcal{R}/M_i)$ is unity for a monatomic species. For the planar stagnation line and the flat plate, and other values of the temperature ratio T_s/T_e, the values of the species weighting factors are divided by the values given by Equations (4.7.71a,b,c) to give correction factors G_{mi}, G_{fi}, and G_{hi}, respectively. The correction factors are listed in Table 4.7.7.

The exponential relation blowing factors cannot accurately represent some of the more anomalous effects of blowing. For example, when a light gas such as H_2 is injected, Equation (4.7.70c) indicates that the effect of blowing is always to reduce heat transfer, due to both the low density and high specific heat of hydrogen. However, at very low injection rates, the heat transfer is actually increased, as a result of the high thermal conductivity of H_2. For a mixture, $k \approx \sum x_i k_i$ whereas $c_p = \sum m_i c_{pi}$. At low rates of injection, the mole fraction of H_2 near the wall is much larger than its mass fraction; thus, there is a substantial increase in the mixture conductivity near the wall, but only a small change in the mixture specific heat. An increase in heat transfer results. At higher injection rates, the mass fraction of H_2 is also large, and the effect of high mixture specific heat dominates to cause a decrease in heat transfer.

2. *Turbulent Boundary Layers.* Here we restrict our attention to air flow along a flat plate for Mach numbers up to 6, and use numerical solutions of boundary layer equations with a mixing length turbulence model (Landis, 1971). Appropriate species weighting factors for $0.2 < T_s/T_e < 2$ are

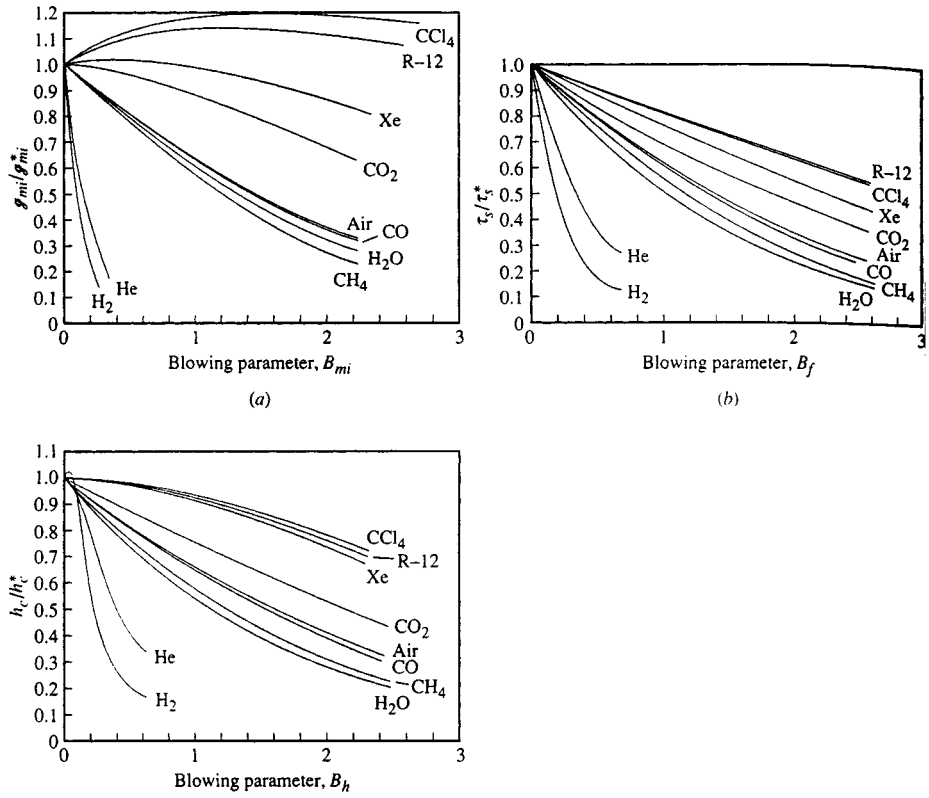


FIGURE 4.7.17 Numerical results for the effect of variable properties on blowing factors for a low-speed turbulent air boundary layer on a cold flat plate (T_s/T_e = 0.2) with foreign gas injection: (a) mass transfer conductance, (b) wall shear stress, (c) heat transfer coefficient. (From Landis, R.B., Ph.D. dissertation, University of California, Los Angeles, 1971. With permission.)

$$a_{mi} = 0.79\left(M_{\text{air}}/M_i\right)^{1.33} \tag{4.7.72a}$$

$$a_{fi} = 0.91\left(M_{\text{air}}/M_i\right)^{0.76} \tag{4.7.72b}$$

$$a_{hi} = 0.86\left(M_{\text{air}}/M_i\right)^{0.73} \tag{4.7.72c}$$

In using Equation (4.7.70), the limit values for $\dot{m}'' = 0$ are elevated at the same location along the plate. Whether the injection rate is constant along the plate or varies as $x^{-0.2}$ to give a self-similar boundary layer has little effect on the blowing factors. Thus, Equation (4.7.72) has quite general applicability. Notice that the effects of injectant molecular weight are greater for turbulent boundary layers than for laminar ones, which is due to the effect of fluid density on turbulent transport. Also, the injectant specific heat does not appear in a_{hi} as it did for laminar flows. In general, c_{pi} decreases with increasing M_i and is adequately accounted for in the molecular weight ratio.

Reference State Schemes. The reference state approach, in which constant-property data are used with properties evaluated at some reference state, is an alternative method for handling variable-property effects. In principle, the reference state is independent of the precise property data used and of the

TABLE 4.7.7 Correction Factors for Foreign Gas Injection into Laminar Air Boundary Layers

Geometry	Species	G_{mi} T_s/T_e			G_{fi} T_s/T_e			G_{hi} T_s/T_e		
		0.1	**0.5**	**0.9**	**0.1**	**0.5**	**0.9**	**0.1**	**0.5**	**0.9**
	H	1.14	1.36	1.47	1.30	1.64	1.79	1.15	1.32	—
	H_2	1.03	1.25	1.36	1.19	1.44	1.49	1.56	1.17	1.32
	He	1.05	1.18	1.25	1.34	1.49	1.56	1.18	1.32	—
	Air	—	—	—	1.21	1.27	1.27	1.17	1.21	—
	Xe	1.21	1.13	1.15	1.38	1.34	1.34	1.19	1.18	—
	CCl_4	1.03	0.95	1.00	1.00	1.03	1.03	1.04	1.04	—
Axisymmetric	H	1.00	1.04	1.09	1.00	0.62	0.45	1.00	0.94	0.54
stagnation point	H_2	1.00	1.06	1.06	1.00	0.70	0.62	1.00	1.00	1.01
	He	1.00	1.04	1.03	1.00	0.66	0.56	1.00	1.00	0.95
	C	1.00	1.01	1.00	1.00	0.79	0.69	1.00	0.99	0.87
	CH_4	1.00	1.01	1.00	1.00	0.88	0.84	1.00	1.00	1.00
	O	1.00	0.98	0.97	1.00	0.79	0.70	1.00	0.98	0.95
	H_2O	1.00	1.01	1.00	1.00	0.82	0.73	1.00	1.00	0.99
	Ne	1.00	1.00	0.98	1.00	0.83	0.75	1.00	0.97	0.95
	Air	—	—	—	1.00	0.87	0.82	1.00	0.99	0.97
	A	1.00	0.97	0.94	1.00	0.93	0.91	1.00	0.96	0.95
	CO_2	1.00	0.97	0.95	1.00	0.96	0.94	1.00	0.99	0.97
	Xe	1.00	0.98	0.96	1.00	0.96	1.05	1.00	1.06	0.99
	CCl_4	1.00	0.90	0.83	1.00	1.03	1.07	1.00	0.96	0.93
	I_2	1.00	0.91	0.85	1.00	1.02	1.05	1.00	0.97	0.94
Planar stagnation line	He	0.96	0.98	0.98	0.85	0.53	0.47	0.93	0.91	0.92
	Air	—	—	—	0.94	0.84	0.81	0.94	0.94	—
	Xe	0.92	0.87	0.83	0.90	0.93	0.95	0.93	0.93	—

Based on numerical data of Wortman (1969). Correlations developed by Dr. D.W. Hatfield.

combination of injectant and free-stream species. A reference state for a boundary layer on a flat plate that can be used in conjunction with Figure 4.7.14 is (Knuth, 1963)

$$m_{1,r} = 1 - \frac{M_2}{M_2 - M_1} \frac{\ln\left(M_e/M_s\right)}{\ln\left(m_{2,e}M_e/m_{2,s}M_s\right)} \tag{4.7.73}$$

$$T_r = 0.5\left(T_e + T_s\right) + 0.2r^*\left(u_e^2/2c_{pr}\right) + 0.1\left[B_{hr} + \left(B_{hr} + B_{mr}\right)\frac{c_{p1} - c_{pr}}{c_{pr}}\right]\left(T_s - T_e\right) \tag{4.7.74}$$

where species 1 is injected into species 2 and r^* is the recovery factor for an impermeable wall. Use of the reference state method is impractical for hand calculations: a computer program should be used to evaluate the required mixture properties.

References

Hirschfelder, J.O., Curtiss, C.F., and Bird, R.B. 1954. *Molecular Theory of Gases and Liquids*, John Wiley & Sons, New York.

Knuth, E.L. 1963. Use of reference states and constant property solutions in predicting mass-, momentum-, and energy-transfer rates in high speed laminar flows, *Int. J. Heat Mass Transfer*, 6, 1–22.

Landis, R.B. 1972. Numerical solution of variable property turbulent boundary layers with foreign gas injection, Ph.D. dissertation, School of Engineering and Applied Science, University of California, Los Angeles.

Law, C.K. and Williams, F.A. 1972. Kinetics and convection in the combustion of alkane droplets, *Combustion and Flame*, 19, 393–405.

Mills, A.F. 1995. *Heat and Mass Transfer*, Richard D. Irwin, Chicago.
Wortman, A. 1969. Mass transfer in self-similar boundary-layer flows, Ph.D. dissertation, School of Engineering and Applied Science, University of California, Los Angeles.

Further Information

Geankoplis, C.J. 1993. *Transport Processes and Unit Operations*, 3rd ed., Prentice-Hall, Englewood Cliffs, NJ. This text gives a chemical engineering perspective on mass transfer.
Mills, A.F. 1995. *Heat and Mass Transfer*, Richard D. Irwin, Chicago. Chapter 11 treats mass transfer equipment relevant to mechanical engineering.
Strumillo, C. and Kudra, T. 1986. *Drying: Principles, Applications and Design*, Gordon and Breach, New York.
Mujamdar, A.S.. Ed. 1987. *Handbook of Industrial Drying*, Marcel Dekker, New York.

4.8 Applications

Enhancement

Arthur E. Bergles

Introduction

Energy- and materials-saving considerations, as well as economic incentives, have led to efforts to produce more efficient heat exchange equipment. Common thermal-hydraulic goals are to reduce the size of a heat exchanger required for a specified heat duty, to upgrade the capacity of an existing heat exchanger, to reduce the approach temperature difference for the process streams, or to reduce the pumping power.

The study of improved heat transfer performance is referred to as heat transfer *enhancement*, *augmentation*, or *intensification*. In general, this means an increase in heat transfer coefficient. Attempts to increase "normal" heat transfer coefficients have been recorded for more than a century, and there is a large store of information. A survey (Bergles et al., 1991) cites 4345 technical publications, excluding patents and manufacturers' literature. The literature has expanded rapidly since 1955.

Enhancement techniques can be classified either as passive methods, which require no direct application of external power (Figure 4.8.1), or as active methods, which require external power. The effectiveness of both types of techniques is strongly dependent on the mode of heat transfer, which may range from single-phase free convection to dispersed-flow film boiling. Brief descriptions of these methods follow.

Treated surfaces involve fine-scale alternation of the surface finish or coating (continuous or discontinuous). They are used for boiling and condensing; the roughness height is below that which affects single-phase heat transfer.

Rough surfaces are produced in many configurations ranging from random sand-grain-type roughness to discrete protuberances. See Figure 4.8.1a. The configuration is generally chosen to disturb the viscous sublayer rather than to increase the heat transfer surface area. Application of rough surfaces is directed primarily toward single-phase flow.

Extended surfaces are routinely employed in many heat exchangers. See Figure 4.8.1a to d. Work of special interest to enhancement is directed toward improvement of heat transfer coefficients on extended surfaces by shaping or perforating the surfaces.

Displaced enhancement devices are inserted into the flow channel so as indirectly to improve energy transport at the heated surface. They are used with forced flow. See Figure 4.8.1e and f.

Swirl-flow devices include a number of geometric arrangements or tube inserts for forced flow that create rotating and/or secondary flow: coiled tubes, inlet vortex generators, twisted-tape inserts, and axial-core inserts with a screw-type winding.

Surface-tension devices consist of wicking or grooved surfaces to direct the flow of liquid in boiling and condensing.

Additives for liquids include solid particles and gas bubbles in single-phase flows and liquid trace additives for boiling systems.

Additives for gases are liquid droplets or solid particles, either dilute-phase (gas-solid suspensions) or dense-phase (fluidized beds).

Mechanical aids involve stirring the fluid by mechanical means or by rotating the surface. Surface "scraping," widely used for batch processing of viscous liquids in the chemical process industry, is applied to the flow of such diverse fluids as high-viscosity plastics and air. Equipment with rotating heat exchanger ducts is found in commercial practice.

Surface vibration at either low or high frequency has been used primarily to improve single-phase heat transfer.

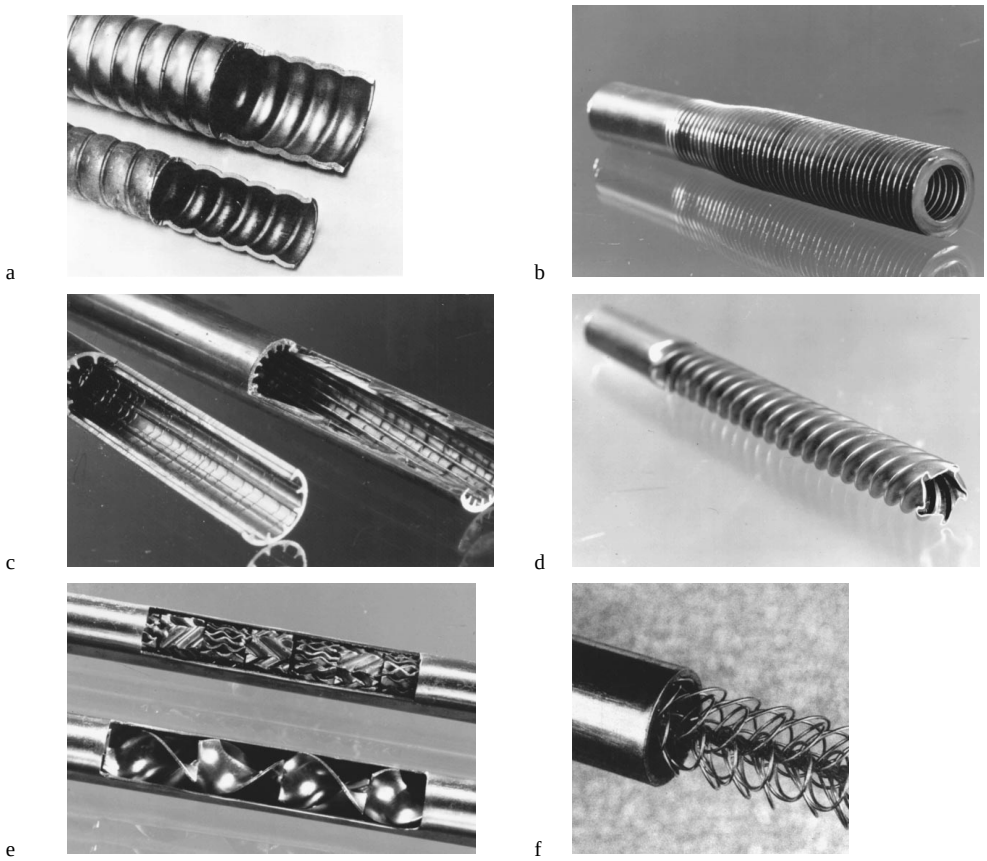


FIGURE 4.8.1 Enhanced tubes for augmentation of single-phase heat transfer. (a) Corrugated or spirally indented tube with internal protuberances. (b) Integral external fins. (c) Integral internal fins. (d) Deep spirally fluted tube. (e) Static mixer inserts. (f) Wire-wound insert.

Fluid vibration is the practical type of vibration enhancement because of the mass of most heat exchangers. The vibrations range from pulsations of about 1 Hz to ultrasound. Single-phase fluids are of primary concern.

Electrostatic fields (DC or AC) are applied in many different ways to dielectric fluids. Generally speaking, electrostatic fields can be directed to cause greater bulk mixing or fluid or disruption of fluid flow in the vicinity of the heat transfer surface, which enhances heat transfer.

Injection is utilized by supplying gas to a stagnant or flowing liquid through a porous heat transfer surface or by injecting similar fluid upstream of the heat transfer section. Surface degassing of liquids can produce enhancement similar to gas injection. Only single-phase flow is of interest.

Suction involves vapor removal, in nucleate or film boiling, or fluid withdrawal, in single-phase flow, through a porous heated surface.

Two or more of the above techniques may be utilized simultaneously to produce an enhancement that is larger than either of the techniques operating separately. This is termed *compound enhancement*.

It should be emphasized that one of the motivations for studying enhanced heat transfer is to assess the effect of an inherent condition on heat transfer. Some practical examples include roughness produced by standard manufacturing, degassing of liquids with high gas content, surface vibration resulting from rotating machinery or flow oscillations, fluid vibration resulting from pumping pulsation, and electrical fields present in electrical equipment.

The surfaces in Figure 4.8.1 have been used for both single-phase and two-phase heat transfer enhancement. The emphasis is on effective and cost-competitive (proved or potential) techniques that have made the transition from the laboratory to commercial heat exchangers.

Single-Phase Free Convection

With the exception of the familiar technique of providing extended surfaces, the passive techniques have little to offer in the way of enhanced heat transfer for free convection. This is because the velocities are usually too low to cause flow separation or secondary flow.

The restarting of thermal boundary layers in interrupted extended surfaces increases heat transfer so as to more than compensate for the lost area.

Mechanically aided heat transfer is a standard technique in the chemical and food industries when viscous liquids are involved. The predominant geometry for surface vibration has been the horizontal cylinder, vibrated either horizontally or vertically. Heat transfer coefficients can be increased tenfold for both low-frequency/high-amplitude and high-frequency/low-amplitude situations. It is, of course, equally effective and more practical to provide steady forced flow. Furthermore, the mechanical designer is concerned that such intense vibrations could result in equipment failures.

Since it is usually difficult to apply surface vibrations to practical equipment, an alternative technique is utilized whereby vibrations are applied to the fluid and focused toward the heated surface. With proper transducer design, it is also possible to improve heat transfer to simple heaters immersed in gases or liquids by several hundred percent.

Electric fields are particularly effective in increasing heat transfer coefficients in free convection. Dielectrophoretic or electrophoretic (especially with ionization of gases) forces cause greater bulk mixing in the vicinity of the heat transfer surface. Heat transfer coefficients may be improved by as much as a factor of 40 with electrostatic fields up to 100,000 V. Again, the equivalent effect could be produced at lower capital cost and without the voltage hazard by simply providing forced convection with a blower or fan.

Single-Phase Forced Convection

The present discussion emphasizes enhancement of heat transfer *inside* ducts that are primarily of circular cross section. Typical data for turbulence promoters inserted inside tubes are shown in Figure 4.8.2. As shown in Figure 4.8.2a, the promoters produce a sizable elevation in the Nusselt number, or heat transfer coefficient, at constant Reynolds number, or velocity. However, as shown in Figure 4.8.2b, there is an accompanying large increase in the friction factor.

Surface roughness has been used extensively to enhance forced convection heat transfer. Integral roughness may be produced by the traditional manufacturing processes of machining, forming, casting, or welding. Various inserts can also provide surface protuberances. In view of the infinite number of possible geometric variations, it is not surprising that, even after more than 300 studies, no completely satisfactory unified treatment is available.

In general, the maximum enhancement of laminar flow with many of the techniques is the same order of magnitude, and seems to be independent of the wall boundary condition. The enhancement with some rough tubes, corrugated tubes, inner-fin tubes, various static mixers, and twisted-type inserts is about 200%. The improvements in heat transfer coefficient with turbulent flow in rough tubes (based on nominal surface area) are as much as 250%. Analogy solutions for sand-grain-type roughness and for square-repeated-rib roughness have been proposed. A statistical correlation is also available for heat transfer coefficient and friction factor.

The following correlations are recommended for tubes with transverse or helical repeated ribs (Figure 4.8.1a) with turbulent flow (Ravigururajan and Bergles, 1985):

$$\mathrm{Nu}_{D_i,a}/\mathrm{Nu}_{D_i,s} = \left\{1+\left[2.64\mathrm{Re}^{0.036}\left(e/D_i\right)^{0.212}\left(\left(p/D_i\right)^{-0.21}\right)\left(\alpha/90\right)^{0.29}\left(\mathrm{Pr}\right)^{-0.024}\right]^7\right\}^{1/7} \qquad (4.8.1)$$

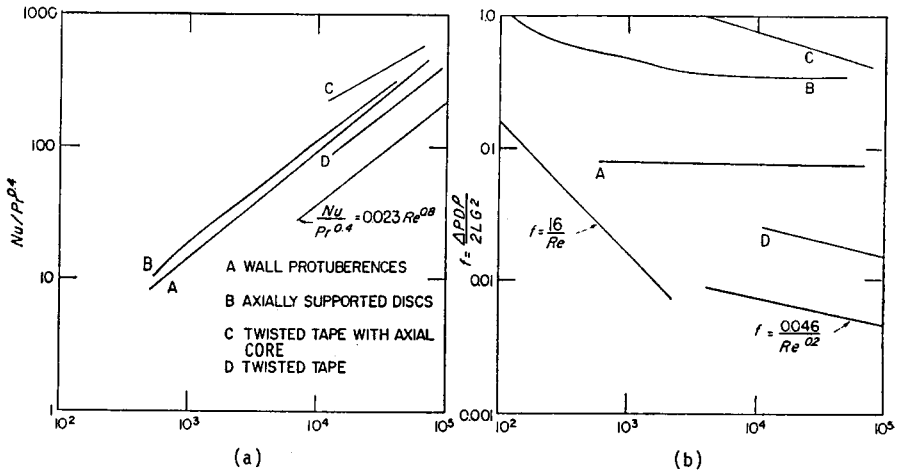


FIGURE 4.8.2 Typical data for turbulence promoters inserted inside tubes: (a) heat transfer data, (b) friction data. (From Bergles, 1969. With permission.)

$$f_a/f_s = \left\{1+\left[29.1\mathrm{Re}_{D_i}^{(0.67-0.06p/D_i-0.49a/90)}\times\left(e/D_i\right)^{(0.37-0.157p/D_i)}\times\left(p/D_i\right)^{\left(-1.66\times10^{-6}\mathrm{Re}_{D_i}-0.33\alpha/90\right)}\right.\right.$$
$$\left.\left.\times\left(\alpha/90\right)^{\left(4.59+4.11\times10^{-6}\mathrm{Re}_{D_i}-0.15p/D_i\right)}\times\left(1+\frac{2.94}{n}\right)\sin\beta\right]^{15/16}\right\}^{16/15} \tag{4.8.2}$$

where the subscript *a* refers to the enhanced tube and the subscript *s* refers to the smooth tube. The special symbols are given as follows: e = protuberance height; p = repeated-rib pitch; α = spiral angle for helical ribs, °; n = number of sharp corners facing the flow; and β = contact angle of rib profile, °.

Also,

$$\mathrm{Nu}_s = 0.125 f\,\mathrm{Re}_{D_i}\mathrm{Pr}\Big/\left(1+12.7\left(0.125f\right)^{0.5}\mathrm{Pr}^{0.667}-1\right)$$

and

$$f_s = \left(1.82\log_{10}\mathrm{Re}_{D_i}-1.64\right)^{-2^*\,*}$$

Much work has been done to obtain the enhanced heat transfer of parallel angled ribs in short rectangular channels, simulating the interior of gas turbine blades. Jets are frequently used for heating, cooling, and drying in a variety of industrial applications. A number of studies have reported that roughness elements of the transverse-repeated-rib type mitigate the deterioration in heat transfer downstream of stagnation.

Extended surfaces can be considered "old technology" as far as most applications are concerned. The real interest now is in increasing heat transfer coefficients on the extended surface. Compact heat exchangers of the plate-fin or tube-and-center variety use several enhancement techniques: offset strip fins, louvered fins, perforated fins, or corrugated fins. Coefficients are several hundred percent above the

* The Fanning friction factor is used throughout this section.

smooth-tube values; however, the pressure drop is also substantially increased, and there may be vibration and noise problems.

For the case of offset strip fins the following correlations are recommended for calculating the *j* and *f* characteristics (Manglik and Bergles, 1990)

$$j_h = 0.6522\mathrm{Re}_h^{-0.5403}\,\alpha^{-0.1541}\,\delta^{0.1499}\,\gamma^{-0.0678}\times\left[1+5.269\times10^{-5}\mathrm{Re}_h^{1.340}\,\alpha^{0.504}\,\delta^{0.456}\,\gamma^{-1.055}\right]^{0.1} \tag{4.8.3}$$

$$f_h = 9.6243\mathrm{Re}_h^{-0.7422}\,\alpha^{-0.1856}\,\delta^{0.3053}\,\gamma^{-0.2659}\times\left[1+7.669\times10^{-8}\mathrm{Re}_h^{4.429}\,\alpha^{0.920}\,\delta^{3.767}\,\gamma^{0.236}\right]^{0.1} \tag{4.8.4}$$

where j_H (the heat transfer *j*-factor $\mathrm{Nu}_H/\mathrm{Re}_H\mathrm{Pr}^{1/3}$), and f_h, and Re_h are based on the hydraulic diameter given by

$$D_h = 4shl/\left[2(sl+hl+th)+ts\right] \tag{4.8.5}$$

Special symbols are α = aspect ratio *s/h*, δ = ratio *t/l*, γ = ratio *t/s*, *s* = lateral spacing of strip fin, *h* = strip fin height, *l* = length of one offset module of strip fins, and *t* = fin thickness.

These equations are based on experimental data for 18 different offset strip-fin geometries, and they represent the data continuously in the laminar, transition, and turbulent flow regions.

Internally finned circular tubes are available in aluminum and copper (or copper alloys). Correlations (for heat transfer coefficient and friction factor) are available for laminar flow, for both straight and spiral continuous fins.

Turbulent flow in tubes with straight or helical fins (Figure 4.8.1c) was correlated by (Carnavos, 1979)

$$\mathrm{Nu}_h = 0.023\mathrm{Pr}^{0.4}\mathrm{Re}_h^{0.8}\left[\frac{A_c}{A_{ci}}\right]^{0.1}\left[\frac{A_{s,i}}{A_s}\right]^{0.5}(\sec\alpha)^3 \tag{4.8.6}$$

$$f_h = 0.046\mathrm{Re}_h^{-0.2}\left[\frac{A_c}{A_{c,i}}\right]^{0.5}(\sec\alpha)^{0.75} \tag{4.8.7}$$

where $A_{c,i}$ is based on the maximum inside (envelope) flow area, $A_{s,i}$ is based on the maximum inside (envelope) surface area, and α the spiral angle for helical fins, °.

A numerical analysis of turbulent flow in tubes with idealized straight fins was reported. The necessary constant for the turbulence model was obtained from experimental data for air. Further improvements in numerical techniques are expected, so that a wider range of geometries and fluids can be handled without resort to extensive experimental programs.

Many proprietary surface configurations have been produced by deforming the basic tube. The "convoluted," "corrugated," "spiral," or "spirally fluted" tubes (Figure 4.8.1a) have multiple-start spiral corrugations, which add area, along the tube length. A systematic survey of the single-tube performance of condenser tubes indicates up to 400% increase in the nominal inside heat transfer coefficient (based on diameter of a smooth tube of the same maximum inside diameter); however, pressure drops on the water side are about 20 times higher.

Displaced enhancement devices are typically in the form of inserts, within elements arranged to promote transverse mixing (static mixers, Figure 4.8.1e). They are used primarily for viscous liquids, to promote either heat transfer or mass transfer. Displaced promoters are also used to enhance the radiant heat transfer in high-temperature applications. In the flue-tube of a hot-gas-fired hot water heater, there is a trade-off between radiation and convection. Another type of displaced insert generates vortices, which enhance the downstream flow. Delta-wing and rectangular wing promoters, both co-rotating and

counterrotating, have been studied. Wire-loop inserts (Figure 4.8.1f) have also been used for enhancement of laminar and turbulent flow.

Twisted-tape inserts have been widely used to improve heat transfer in both laminar and turbulent flow. Correlations are available for laminar flow, for both uniform heat flux and uniform wall temperature conditions. Turbulent flow in tubes with twisted-tape inserts has also been correlated. Several studies have considered the heat transfer enhancement of a decaying swirl flow, generated, say, by a short twisted-tape insert.

Performance Evaluation Criteria for Single-Phase Forced Convection in Tubes

Numerous, and sometimes conflicting, factors enter into the ultimate decision to use an enhancement technique: heat duty increase or area reduction that can be obtained, initial cost, pumping power or operating cost, maintenance cost (especially cleaning), safety, and reliability, among others. These factors are difficult to quantitize, and a generally acceptable selection criterion may not exist. It is possible, however, to suggest some performance criteria for preliminary design guidance. As an example, consider the basic geometry and the pumping power fixed, with the objective of increasing the heat transfer. The following ratio is then of interest

$$R_3 = \left(\frac{h_a}{h_s}\right)_{D_i,L,N,P,T_{in},\Delta T} = \frac{\left(\mathrm{Nu}/\mathrm{Pr}^{0.4}\right)_a}{\left(\mathrm{Nu}/\mathrm{Pr}^{0.4}\right)_s} = \frac{q_a}{q_s} \tag{4.8.8}$$

where P = pumping power, T_{in} = inlet bulk temperature of fluid, and ΔT = average wall-fluid temperature difference.

With the pumping power (neglecting entrance and exit losses) given as

$$P = NVA_c 4f(L/D)\rho V^2/2 \tag{4.8.9}$$

and

$$f_s = 0.046/\mathrm{Re}_s^{0.2} \tag{4.8.10}$$

$$A_{c,a} f_a \mathrm{Re}_a^3 = {}^{0.046}A_{c,s}\mathrm{Re}_s^{2.8} \tag{4.8.11}$$

The calculation best proceeds by picking $\mathrm{Re}_{D_i,a}$, and reading $\mathrm{Nu}_{D_i,a}/\mathrm{Pr}^{0.4}$ and f_a. $\mathrm{Re}_{D_i,s}$ is then obtained from Equation (4.8.11) and $\mathrm{Nu}_{D_i,s}/\mathrm{Pr}^{0.4}$ obtained from a conventional, empty-tube correlation. The desired ratio of Equation (4.8.8) is then obtained. Typical results are presented in Figure 4.8.3 for a repeated-rib roughness (Bergles et al., 1974).

Active and Compound Techniques for Single-Phase Forced Convection

Under active techniques, mechanically aided heat transfer in the form of surface scraping can increase forced convection heat transfer. Surface vibration has been demonstrated to improve heat transfer to both laminar and turbulent duct flow of liquids. Fluid vibration has been extensively studied for both air (loudspeakers and sirens) and liquids (flow interrupters, pulsators, and ultrasonic transducers). Pulsations are relatively simple to apply to low-velocity liquid flows, and improvements of several hundred percent can be realized.

Some very impressive enhancements have been recorded with electrical fields, particularly in the laminar-flow region. Improvements of at least 100% were obtained when voltages in the 10-kV range were applied to transformer oil. It is found that even with intense electrostatic fields, the heat transfer enhancement disappears as turbulent flow is approached in a circular tube with a concentric inner electrode.

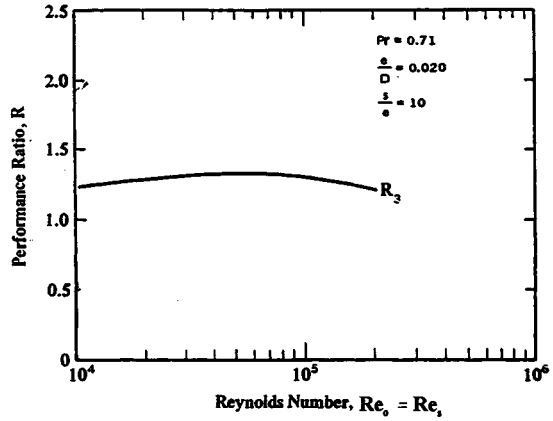


FIGURE 4.8.3 Constant pumping power performance criterion applied to repeated rib roughness.

Compound techniques are a slowly emerging area of enhancement that holds promise for practical applications, since heat transfer coefficients can usually be increased above each of the several techniques acting along. Some examples that have been studied are as follows: rough tube wall with twisted-tape inserts, rough cylinder with acoustic vibrations, internally finned tube with twisted-tape inserts, finned tubes in fluidized beds, externally finned tubes subjected to vibrations, rib-roughened passage being rotated, gas-solid suspension with an electrical field, fluidized bed with pulsations of air, and a rib-roughened channel with longitudinal vortex generation.

Pool Boiling

Selected passive and active enhancement techniques have been shown to be effective for pool boiling and flow boiling/evaporation. Most techniques apply to nucleate boiling; however, some techniques are applicable to transition and film boiling.

It should be noted that phase-change heat transfer coefficients are relatively high. The main thermal resistance in a two-fluid heat exchanger often lies on the non-phase-change side. (Fouling of either side can, of course, represent the dominant thermal resistance.) For this reason, the emphasis is often on enhancement of single-phase flow. On the other hand, the overall thermal resistance may then be reduced to the point where significant improvement in the overall performance can be achieved by enhancing the two-phase flow. Two-phase enhancement would also be important in double-phase-change (boiling/condensing) heat exchangers.

As discussed elsewhere, surface material and finish have a strong effect on nucleate and transition pool boiling. However, reliable control of nucleation on plain surfaces is not easily accomplished. Accordingly, since the earliest days of boiling research, there have been attempts to relocate the boiling curve through use of relatively gross modification of the surface. For many years, this was accomplished simply by area increase in the form of low helical fins. The subsequent tendency was to structure surfaces to improve the nucleate boiling characteristics by a fundamental change in the boiling process. Many of these advanced surfaces are being used in commercial shell-and-tube boilers.

Several manufacturing processes have been employed: machining, forming, layering, and coating. In Figure 4.8.4a standard low-fin tubing is shown. Figure 4.8.4c depicts a tunnel-and-pore arrangement produced by rolling, upsetting, and brushing. An alternative modification of the low fins is shown in Figure 4.8.4d, where the rolled fins have been split and rolled to a T shape. Further modification of the internal, Figure 4.8.4e, or external, Figure 4.8.4f, surface is possible. Knurling and rolling are involved in producing the surface shown in Figure 4.8.4g. The earliest example of a commercial structured surface, shown in Figure 4.8.4b is the porous metallic matrix produced by sintering or brazing small particles. Wall superheat reductions of up to a factor of ten are common with these surfaces. The advantage is not

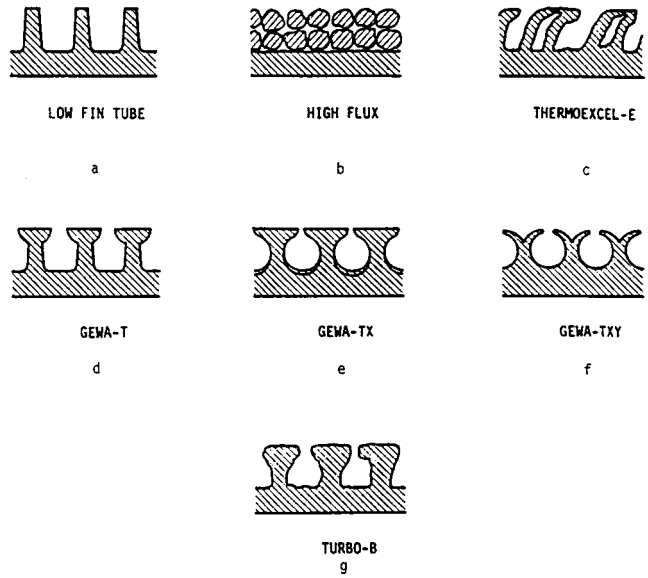


FIGURE 4.8.4 Examples of commercial structured boiling surfaces. (From Pate, M.B. et al., in *Compact Heat Exchangers*, Hemisphere Publishing, New York, 1990. With permission.)

only a high nucleate boiling heat transfer coefficient, but the fact that boiling can take place at very low temperature differences.

These structured boiling surfaces, developed for refrigeration and process applications, have been used as "heat sinks" for immersion-cooled microelectronic chips.

The behavior of tube bundles is often different with structured-surface tubes. The enhanced nucleate boiling dominates, and the convective boiling enhancement, found in plain tube bundles, does not occur.

Active enhancement techniques include heated surface rotation, surface wiping, surface vibration, fluid vibration, electrostatic fields, and suction at the heated surface. Although active techniques are effective in reducing the wall superheat and/or increasing the critical heat flux, the practical applications are very limited, largely because of the difficulty of reliably providing the mechanical or electrical effect.

Compound enhancement, which involves two or more techniques applied simultaneously, has also been studied. Electrohydrodynamic enhancement was applied to a finned tube bundle, resulting in nearly a 200% increase in the average boiling heat transfer coefficient of the bundle, with a small power consumption for the field.

Convective Boiling/Evaporation

The structured surfaces described in the previous section are generally not used for in-tube vaporization, because of the difficulty of manufacture. One notable exception is the high-flux surface in a vertical thermosiphon reboiler. The considerable increase in the low-quality, nucleate boiling coefficient is desirable, but it is also important that more vapor is generated to promote circulation.

Helical repeated ribs and helically coiled wire inserts have been used to increase vaporization coefficients and the dry-out heat flux in once-through boilers.

Numerous tubes with internal fins, either integral or attached, are available for refrigerant evaporators. Original configurations were tightly packed, copper, offset strip fin inserts soldered to the copper tube or aluminum, star-shaped inserts secured by drawing the tube over the insert. Examples are shown in Figure 4.8.5. Average heat transfer coefficients (based on surface area of smooth tube of the same diameter) for typical evaporator conditions are increased by as much as 200%. A cross-sectional view of a typical "microfin" tube is included in Figure 4.8.5. The average evaporation boiling coefficient is

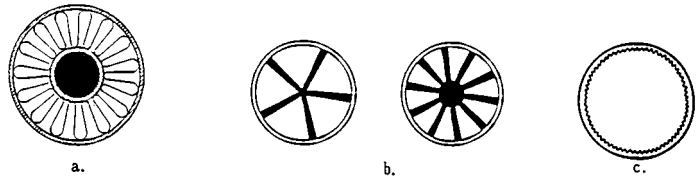


FIGURE 4.8.5 Inner-fin tubes for refrigerant evaporators: (a) Strip-fin inserts, (b) Star-shaped inserts, (c) Microfin.

increased 30 to 80%. The pressure drop penalties are less; that is, lower percentage increases in pressure drop are frequently observed.

Twisted-tape inserts are generally used to increase the burnout heat flux for subcooled boiling at high imposed heat fluxes $10^7 - 10^8$ W/m^2, as might be encountered in the cooling of fusion reactor components. Increases in burnout heat flux of up to 200% were obtained at near atmospheric pressure.

Vapor-Space Condensation

As discussed elsewhere, condensation can be either filmwise or dropwise. In a sense, dropwise condensation is enhancement of the normally occurring film condensation by surface treatment. The only real application is for steam condensers, because nonwetting coatings are not available for most other working fluids. Even after much study, little progress has been made in developing permanently hydrophobic coatings for practical steam condensers. The enhancement of dropwise condensation is pointless, because the heat transfer coefficients are already so high.

Surface extensions are widely employed for enhancement of condensation. The integral low fin tubing (Figure 4.8.4a), used for kettle boilers, is also used for horizontal tube condensers. With proper spacing of the fins to provide adequate condensate drainage, the average coefficients can be several times those of a plain tube with the same base diameter. These fins are normally used with refrigerants and other organic fluids that have low condensing coefficients, but which drain effectively, because of low surface tension.

The fin profile can be altered according to mathematical analysis to take full advantage of the Gregorig effect, whereby condensation occurs mainly at the tops of convex ridges. Surface tension forces then pull the condensate into concave grooves, where it runs off. The average heat transfer coefficient is greater than that for an axially uniform film thickness. The initial application was for condensation of steam on vertical tubes used for reboilers and in desalination. According to numerical solutions, the optimum geometry is characterized by a sharp fin tip, gradually changing curvature of the fin surface from tip to root, wide grooves between fins to collect condensate, and periodic condensate strippers. Figure 4.8.6 schematically presents the configuration.

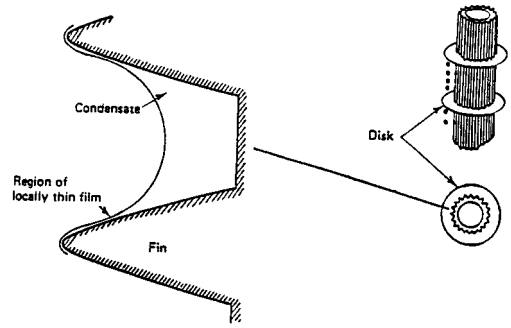


FIGURE 4.8.6 Recommended flute profile and schematic of condensate strippers.

Recent interest has centered on three-dimensional surfaces for horizontal-tube condensers. The considerable improvement relative to low fins or other two-dimensional profiles is apparently due to multidimensional drainage at the fin tips. Other three-dimensional shapes include circular pin fins, square pins, and small metal particles that are bonded randomly to the surface.

Convective Condensation

This final section on enhancement of the various modes of heat transfer focuses on in-tube condensation. The applications include horizontal kettle-type reboilers, moisture separator reheaters for nuclear power plants, and air-conditioner condensers.

Internally grooved or knurled tubes, deep spirally fluted tubes, random roughness, conventional inner-fin tubes have been shown to be effective for condensation of steam and other fluids.

The microfin tubes mentioned earlier have also been applied successfully to in-tube condensing. As in the case of evaporation, the substantial heat transfer improvement is achieved at the expense of a lesser percentage increase in pressure drop. By testing a wide variety of tubes, it has been possible to suggest some guidelines for the geometry, e.g., more fins, longer fins, and sharper tips; however, general correlations are not yet available. Fortunately for heat-pump operation, the tube that performs best for evaporation also performs best for condensation.

Twisted-tape inserts result in rather modest increases in heat transfer coefficient for complete condensation of either steam or refrigerant. The pressure drop increases are large because of the large wetted surface. Coiled tubular condensers provide a modest improvement in average heat transfer coefficient.

References

Bergles, A.E. 1969. Survey and evaluation of techniques to augment convective heat and mass transfer, in *Progress in Heat and Mass Transfer*, Vol. 1, Pergamon, Oxford, England.

Bergles, A.E. 1985. Techniques to augment heat transfer, in *Handbook of Heat Transfer Applications*, W.M. Rohsenow, J.P. Hartnett, and E.N. Ganic, Eds., McGraw-Hill, New York, 3-1–3-80.

Bergles, A.E. 1988. Some perspectives on enhanced heat transfer — second generation heat transfer technology, *J. Heat Transfer*, 110, 1082–1096.

Bergles, A.E. 1997. Heat transfer enhancement — the encouragement and accommodation of high heat fluxes. *J. Heat Transfer*, 119, 8–19.

Bergles, A.E., Blumenkrantz, A.R., and Taborek, J. 1974. Performance evaluation criteria for enhanced heat transfer surfaces, in *Heat Transfer 1974*, The Japan Society of Mechanical Engineers, Tokyo, Vol. II, 234–238.

Bergles, A.E., Jensen, M.K., Somerscales, E.F.C., and Manglik, R.M. 1991. Literature Review of Heat Transfer Enhancement Technology for Heat Exchangers in Gas-Fired Applications, Gas Research Institute Report, GR191-0146.

Carnavos, T.C. 1979. Heat transfer performance of internally finned tubes in turbulent flow, in *Advances in Advanced Heat Transfer*, ASME, New York, 61–67.

Manglik, R.M. and Bergles, A.E. 1990. The thermal-hydraulic design of the rectangular offset-strip-fin compact heat exchanger, in *Compact Heat Exchangers*, Hemisphere Publishing, New York, 123–149.

Pate, M.B., Ayub, Z.H., and Kohler, J. 1990. Heat exchangers for the air-conditioning and refrigeration industry: state-of-the-art design and technology, in *Compact Heat Exchangers*, Hemisphere Publishing, New York, 567–590.

Ravigururajan, S. and Bergles, A.E. 1985. General Correlations for Pressure Drop and Heat Transfer for Single-Phase Turbulent Flow in Internally Ribbed Tubes, in *Augmentation of Heat Transfer in Energy Systems*, HTD-Vol. 52, ASME, New York, 9–20.

Thome, J.R. 1990. *Enhanced Boiling Heat Transfer*, Hemisphere Publishing, New York.

Webb, R.L. 1994. *Principles of Enhanced Heat Transfer*, John Wiley & Sons, New York.

Further Information

This section gives some indication as to why heat transfer enhancement is one of the fastest growing areas of heat transfer. Many techniques are available for improvement of the various modes of heat transfer. Fundamental understanding of the transport mechanism is growing, but, more importantly, design correlations are being established. Many effective and cost-competitive enhancement techniques have made the transition from the laboratory to commercial heat exchangers.

Broad reviews of developments in enhanced heat transfer are available (Bergles, 1985; Bergles, 1988; Thome, 1990; Webb, 1994, Bergles, 1997). Also, several journals, especially *Heat Transfer Engineering*, *Enhanced Heat Transfer*, and *International Journal of Heating, Ventilating, Air-Conditioning and Refrigerating Research*, feature this technology.

Cooling Towers

Anthony F. Mills

Introduction

In a wet cooling tower, water is evaporated into air with the objective of cooling the water stream. Both natural- and mechanical-draft towers are popular, and examples are shown in Figure 4.8.7. Large natural-draft cooling towers are used in power plants for cooling the water supply to the condenser. Smaller mechanical-draft towers are preferred for oil refineries and other process industries, as well as for central air-conditioning systems and refrigeration plant. Figure 4.8.7a shows a natural draft *counterflow* unit in which the water flows as thin films down over a suitable packing, and air flows upward. In a natural-draft tower the air flows upward due to the buoyancy of the warm, moist air leaving the top of the packing. In a mechanical-draft tower, the flow is forced or induced by a fan. Since the air inlet temperature is usually lower than the water inlet temperature, the water is cooled both by evaporation and by sensible heat loss. For usual operating conditions the evaporative heat loss is considerably larger than the sensible heat loss. Figure 4.8.7b shows a mechanical draft cross-flow unit. Figure 4.8.8 shows a natural-draft cross-flow tower for a power plant.

Packing Thermal Performance

Counterflow units. Merkel's method (Merkel, 1925) for calculating the number of transfer units required to cool the water stream, for specified inlet and outlet water temperatures and inlet air condition is (Mills, 1995)

$$N_{tu} = \frac{\beta_m S}{\dot{m}_L} = \int_{h_{L,\text{in}}}^{h_{L,\text{out}}} \frac{dh_L}{h_s - h_G} \tag{4.8.12}$$

$$h_G = h_{G,\text{in}} + \left(\dot{m}_L / \dot{m}_G\right)\left(h_L - h_{L,\text{out}}\right) \tag{4.8.13}$$

$$h_s\left(P, T_s\right) = h_s\left(P, T_L\right) \tag{4.8.14}$$

It is imperative that the usual enthalpy datum states be used, namely, zero enthalpy for dry air and liquid water at 0°C. Table 4.8.1 gives enthalpy data for 1 atm pressure. The important assumptions required to obtain this result include

1. A Lewis number of unity;
2. Low mass transfer rate theory is valid;
3. The liquid-side heat transfer resistance is negligible, that is, $T_s \simeq T_L$;
4. The amount of water evaporated is small compared with the water and air flow rates.

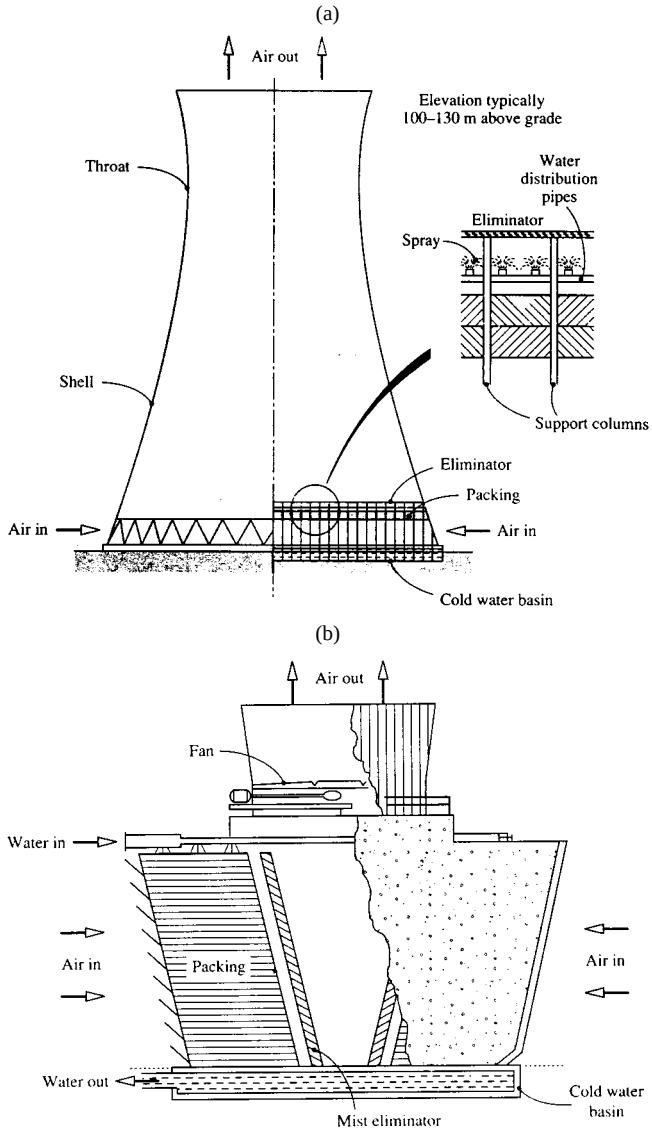


FIGURE 4.8.7 (a) A natural-draft counterflow cooling tower for a power plant. (b) A cross-flow cooling tower for an air-conditioning system.

The method is accurate up to temperatures of about 60°C; comparisons with more exact results are usually within 3 to 5%, and seldom show errors greater than 10%. Notice that the method does not give the outlet state of the air; however, in situations encountered in practice, the outlet air can be assumed to be saturated for the purposes of calculating its density. It is possible to extend Merkel's method to include a finite liquid-side heat transfer resistance, but such refinement is seldom warranted. For typical operating conditions the bulk liquid temperature is seldom more than 0.3 K above the interface temperature.

Cross-flow units. Figure 4.8.9 shows a schematic of a cross-flow packing. If we assume that both the liquid and gas streams are unidirectional, and that there is no mixing in either stream, then use of Merkel's assumptions leads to the following pair of differential equations (Mills, 1995):

FIGURE 4.8.8 A natural-draft cross-flow cooling tower for a power plant.

$$\frac{\partial h_G}{\partial x} = \frac{g_m a}{G}\left(h_s - h_G\right) \tag{4.8.15}$$

$$\frac{\partial h_L}{\partial y} = -\frac{g_m a}{L}\left(h_s - h_G\right) \tag{4.8.16}$$

Also $h_s = h_s(h_L)$ for a negligible liquid-side heat transfer resistance and the required boundary conditions are the inlet enthalpies of both streams. Equations (4.8.15) and (4.8.16) are solved numerically and the solution used to evaluate the average enthalpy of the outlet liquid,

$$\bar{h}_{L,\text{out}} = \frac{1}{X}\int_0^X h_{L,\text{out}}\, dx \tag{4.8.17}$$

Substituting in an exchanger energy balance on the liquid stream gives the heat transfer as

$$q = \dot{m}_L\left(h_{L,\text{in}} - h_{L,\text{out}}\right) \tag{4.8.18}$$

TABLE 4.8.1 Thermodynamic Properties of Water Vapor-Air Mixtures at 1 atm

		Specific Volume, m^3/kg		Enthalpy[a,b] kJ/kg		
Temp., °C	Saturation Mass Fraction	Dry Air	Saturated Air	Liquid Water	Dry Air	Saturated Air
10	0.007608	0.8018	0.8054	42.13	10.059	29.145
11	0.008136	0.8046	0.8086	46.32	11.065	31.481
12	0.008696	0.8075	0.8117	50.52	12.071	33.898
13	0.009289	0.8103	0.8148	54.71	13.077	36.401
14	0.009918	0.8131	0.8180	58.90	14.083	38.995
15	0.01058	0.8160	0.8212	63.08	15.089	41.684
16	0.01129	0.8188	0.8244	67.27	16.095	44.473
17	0 01204	0.8217	0.8276	71.45	17.101	47.367
18	0 01283	0.8245	0.8309	75.64	18.107	50.372
19	0.01366	0.8273	0.8341	79.82	19.113	53.493
20	0.01455	0.8302	0.8374	83.99	20.120	56.736
21	0.01548	0.8330	0.8408	88.17	21.128	60.107
22	0.01647	0.8359	0.8441	92.35	22.134	63.612
23	0.01751	0.8387	0.8475	96.53	23.140	67.259
24	0.01861	0.8415	0.8510	100.71	24.147	71.054
25	0.01978	0.8444	0.8544	104.89	25.153	75.004
26	0.02100	0.8472	0.8579	109.07	26.159	79.116
27	0.02229	0.8500	0.8615	113.25	27.166	83.400
28	0.02366	0.8529	0.8650	117.43	28.172	87.862
29	0.02509	0.8557	0.8686	121.61	29.178	92.511
30	0.02660	0.8586	0.8723	125.79	30.185	97.357
31	0.02820	0.8614	0.8760	129.97	31.191	102.408
32	0.02987	0.8642	0.8798	134.15	32.198	107.674
33	0.03164	0.8671	0.8836	138.32	33.204	113.166
34	0.03350	0.8699	0.8874	142.50	34.211	118.893
35	0.03545	0.8728	0.8914	146.68	35.218	124.868
36	0.03751	0.8756	0.8953	150.86	36.224	131.100
37	0.03967	0.8784	0.8994	155.04	37.231	137.604
38	0.04194	0.8813	0.9035	159.22	38.238	144.389
39	0.04432	0.8841	0.9077	163.40	39.245	151.471
40	0.04683	0.8870	0.9119	167.58	40.252	158.862
41	0.04946	0.8898	0.9162	171.76	41.259	166.577
42	0.05222	0.8926	0.9206	175.94	42.266	174.630
43	0.05512	0.8955	0.9251	180.12	43.273	183.037
44	0.05817	0.8983	0.9297	184.29	44.280	191.815
45	0.06137	0.9012	0.9343	188.47	45.287	200.980
46	0.06472	0.9040	0.9391	192.65	46.294	210.550
47	0.06842	0.9068	0.9439	196.83	47.301	220.543
48	0.07193	0.9097	0.9489	201.01	48.308	230.980
49	0.07580	0.9125	0.9539	205.19	49.316	241.881

[a] The enthalpies of dry air and liquid water are set equal to zero at a datum temperature of 0°C.

[b] The enthalpy of an unsaturated water vapor-air mixture can be calculated as $h = h_{\text{dry air}} + (m_1/m_{1,\text{sat}})(h_{\text{sat}} - h_{\text{dry air}})$.

Sample calculation. Consider a counterflow unit that is required to cool water from 40 to 26°C when the inlet air is at 10°C, 1 atm, and saturated. We will calculate the number of transfer units required for balanced flow, that is, $\dot{m}_G/\dot{m}_L = 1$. Equation (4.8.12) is to be integrated numerically, with h_G obtained from Equation 4.8.13. The required thermodynamic properties can be obtained from Table 4.8.1. Using Table 4.8.1, $h_{G,\text{in}} = h_{\text{sat}}$ (10°C) = 29.15 kJ/kg, $h_{L,\text{out}} = h_L$(26°C) = 109.07 kJ/kg. Substituting in Equation (4.8.13),

$$h_G = 29.15 + \left(h_L - 109.07\right)$$

T_L, °C	h_L, kJ/kg	h_G, kJ/kg	h_s, kJ/kg	$h_s - h_g$, kJ/kg	$\frac{1}{h_s - h_g}$
26	109.07	29.15	79.12	49.97	0.02001
28	117.43	37.51	87.86	50.35	0.01986
30	125.79	45.87	97.36	51.49	0.01942
32	134.15	54.23	107.67	53.44	0.01871
34	142.50	62.58	118.89	56.31	0.01776
36	150.86	70.94	131.10	60.16	0.01662
38	159.22	79.30	144.39	65.09	0.01536
40	167.58	87.66	158.86	71.20	0.01404

Choosing 2°C intervals for convenient numerical integration, the above table is constructed, with h_L and $h_s = h_s(T_L)$ also obtained from Table 4.8.1.
Using the trapezoidal rule,

$$\int_{h_{L,\text{out}}}^{h_{L,\text{in}}} \frac{dh_L}{h_s - h_G} = \frac{8.36}{2}\left[0.02001 + 2(0.01986 + 0.01942 + 0.01871 + 0.01776 + 0.01662 + 0.01536) + 0.01404\right]$$

$$= 1.043$$

From Equation (4.8.12), $N_{tu} = 1.043$. Also, by using Table 4.8.1, $T_{G,\text{out}} = 27.9°$ for saturated outlet air.

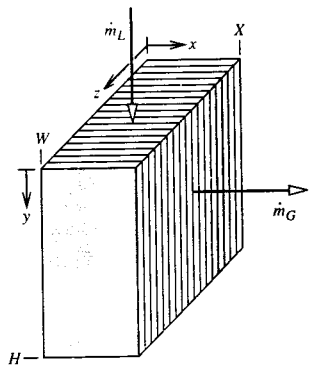


FIGURE 4.8.9 Schematic of a cross-flow cooling tower packing showing the coordinate system.

Thermal-Hydraulic Design of Cooling Towers

The thermal-hydraulic design of a mechanical-draft cooling tower is relatively straightforward. The flow rate ratio $\dot{m}_L/\dot{m}_G$ can be specified and varied parametrically to obtain an optimal design, for which the size and cost of the packing is balanced against fan power requirements and operating cost. Data are required for mass transfer conductances and friction for candidate packings. Tables 4.8.2a and b give correlations for a selection of packings. In Table 4.8.2b, the mass transfer conductance is correlated as $g_m a/L$, where a is the transfer area per unit volume and $L = \dot{m}_L/A_{fr}$ is the superficial mass velocity of the water flow (also called the *water loading* on the packing). Similarly, we define $G = \dot{m}_G/A_{fr}$. Typical water loadings are 1.8 to 2.7 kg/m² sec, and superficial air velocities fall in the range 1.5 to 4 m/sec. No attempt is made to correlate g_m and a separately. The number of transfer units of a packing of height H is then

$$N_{tu} = \frac{g_m S}{\dot{m}_L} = \frac{g_m aH}{L} \tag{4.8.19}$$

TABLE 4.8.2a Packings for Counterflow and Cross-Flow Cooling Towers: Designations and Descriptions

Counterflow Packings

1. Flat asbestos sheets, pitch 4.45 cm
2. Flat asbestos sheets, pitch 3.81 cm
3. Flat asbestos sheets, pitch 3.18 cm
4. Flat asbestos sheets, pitch 2.54 cm
5. 60° angle corrugated plastic, Munters M12060, pitch 1.17 in.
6. 60° angle corrugated plastic, Munters M19060, pitch 1.8 in.
7. Vertical corrugated plastic, American Tower Plastics Coolfilm, pitch 1.63 in.
8. Horizontal plastic screen, American Tower Plastics Cooldrop, pitch 8 in. 2 in. grid
9. Horizontal plastic grid, Ecodyne shape 10, pitch 12 in.
10. Angled corrugated plastic, Marley MC67, pitch 1.88 in.
11. Dimpled sheets, Toschi Asbestos-Free Cement, pitch 0.72 in.
12. Vertical plastic honeycomb, Brentwood Industries Accu-Pack, pitch 1.75 in.

Cross-Flow Packings

1. Doron V-bar, 4 × 8 in. spacing
2. Doron V-bar, 8 × 8 in. spacing
3. Ecodyne T-bar, 4 × 8 in. spacing
4. Ecodyne T-bar, 8 × 8 in. spacing
5. Wood lath, parallel to air flow, 4 × 4 in. spacing
6. Wood lath, perpendicular to air flow, 4 × 4 in. spacing
7. Marley α-bar, parallel to air flow, 16 × 4 in. spacing
8. Marley ladder, parallel to air flow, 8× 2 in. spacing

The correlations are in terms of dimensionless mass velocities L^+ and G^+, and a *hot water correction* T^+_{HW}. The hot water correction accounts for a number of factors, such as errors associated with Merkel's method, deviations from low mass transfer rate theory at higher values of T_s, and fluid property dependence on temperature. Frictional resistance to air flow through the packings is correlated as a *loss coefficient* $N = \Delta P/(\rho V^2/2)$ per unit height or depth of packing, as a function of L^+ and G^+. The velocity V is superficial gas velocity. No hot water correction is required.

In a natural-draft tower, the thermal and hydraulic performance of the tower are coupled, and the flow rate ratio $\dot{m}_L/\dot{m}_G$ cannot be specified *a priori*. The buoyancy force producing the air flow depends on the state of the air leaving the packing which in turn depends on $\dot{m}_L/\dot{m}_G$ and the inlet air and water states. An iterative solution is required to find the operating point of the tower. The buoyancy force available to overcome the shell and packing pressure drops is

$$\Delta P^B = g\left(\rho_a - \rho_{G,\mathrm{out}}\right)H \tag{4.8.20}$$

where ρ_a is the ambient air density and H is usually taken as the distance from the bottom of the packing to the top of the shell. The various pressure drops are conveniently expressed as

$$\Delta P_i = N_i \frac{\rho_{Gi} V_i^2}{2} \tag{4.8.21}$$

TABLE 4.8.2b Mass Transfer and Pressure Drop Correlations for Cooling Towers

Packing Number	C_1, m^{-1}	n_1	n_2	n_3	C_2, m^{-1}	n_4	n_5
Counterflow Packings: $L_0 = G_0 =$ 3.391 kg/m^2 sec							
1	0.289	–0.70	0.70	0.00	2.72	0.35	–0.35
2	0.361	–0.72	0.72	0.00	3.13	0.42	–0.42
3	0.394	–0.76	0.76	0.00	3.38	0.36	–0.36
4	0.459	–0.73	0.73	0.00	3.87	0.52	–0.36
5	2.723	–0.61	0.50	–0.34	19.22	0.34	0.19
6	1.575	–0.50	0.58	–0.40	9.55	0.31	0.05
7	1.378	–0.49	0.56	–0.35	10.10	0.23	–0.04
8	0.558	–0.38	0.48	–0.54	4.33	0.85	–0.60
9	0.525	–0.26	0.58	–0.45	2.36	1.10	–0.64
10	1.312	–0.60	0.62	–0.60	8.33	0.27	–0.14
11	0.755	–0.51	0.93	–0.52	1.51	0.99	0.04
12	1.476	–0.56	0.60	–0.38	6.27	0.31	0.10
Cross-Flow Packings: L_0 = 8.135 kg/m^2 sec, G_0 = 2.715 kg/m^2 sec							
1	0.161	–0.58	0.52	–0.44	1.44	0.66	–0.73
2	0.171	–0.34	0.32	–0.43	1.97	0.72	–0.82
3	0.184	–0.51	0.28	–0.31	1.38	1.30	0.22
4	0.167	–0.48	0.20	–0.29	1.25	0.89	0.07
5	0.171	–0.58	0.28	–0.29	3.18	0.76	–0.80
6	0.217	–0.51	0.47	–0.34	4.49	0.71	–0.59
7	0.213	–0.41	0.50	–0.42	3.44	0.71	–0.85
8	0.233	–0.45	0.45	–0.48	4.89	0.59	0.16

Correlations (SI units)

Mass transfer: $\dfrac{\mathcal{J}_m a}{L\left[\text{kg/m}^2\,\text{sec}\right]} = C_1\left(L^+\right)^{n_1}\left(G^+\right)^{n_2}\left(T_{\text{HW}}^+\right)^{n_3}$; Pressure drop: $\dfrac{N}{H \text{ or } X} = C_2\left(L^+\right)^{n} + \left(G^+\right)^{n_5}$

where $L^+ = \dfrac{L}{L_0}$, $G^+ = \dfrac{G}{G_0}$, $T_{\text{HW}}^+ = \dfrac{1.8T_{L,\text{in}}[°\text{C}] + 32}{110}$

Sources: Lowe, H.J. and Christie, D.G. 1961. "Heat transfer and pressure drop data on cooling tower packings, and model studies of the resistance of natural draft towers to airflow" Paper 113, *International Developments in Heat Transfer, Proc. of the International Heat Transfer Conference*, Boulder, CO, ASME, New York; Johnson, B.M., Ed. 1990. *Cooling Tower Performance Prediction and Improvement*, Vols. 1 and 2, EPRI GS-6370, Electric Power Research Institute, Palo Alto, CA. With permission.

Where N_i is the loss coefficient and V_i is the air velocity at the corresponding location. The pressure drops are associated with the shell, the packing, the mist eliminators, supports and pipes, and the water spray below the packing. Some sample correlations are given in Table 4.8.3.

Water loadings in counterflow natural-draft towers typically range from 0.8 to 2.4 kg/m^2 sec, and superficial air velocities range from 1 to 2 m/sec. The ratio of base diameter to height may be 0.75 to 0.85, and the ratio of throat to base diameter 0.55 to 0.65. The height of the air inlet is usually 0.10 to 0.12 times the base diameter to facilitate air flow into the tower. In practice the air flow distribution in natural-draft towers is not very uniform. However, the assumption of uniform air and water flows in our model of counterflow packing is adequate for most design purposes.

Cost-optimal design of cooling towers requires consideration of the complete power or refrigeration system. For refrigeration, the economics are determined mainly by the operating cost of the chiller (Kintner-Meyer and Emery, 1955).

TABLE 4.8.3 Pressure Drop Correlations for Cooling Tower Shells, Sprays, Supports, and Mist Eliminators

1. Shell (natural draft counterflow):

   $$N = 0.167\left(\frac{D_B}{b}\right)^2$$

   where D_B *is* the diameter of the shell base and b is the height of the air inlet.
2. Spray (natural-draft counterflow):

   $$N = 0.526(Z_p[\text{m}] + 1.22)\ (\dot{m}_L / \dot{m}_G)^{1.32}$$
3. Mist eliminators:

   N = 2–4
4. Support columns, pipes, etc. (natural-draft counterflow):

   N = 2–6
5. Fan exit losses for mechanical-draft towers (velocity based on fan exit area):

   $$N = 1.0, \text{ forced draft}$$

   $$\simeq 0.5, \text{ induced draft, depending on diffuser design}$$
6. Miscellaneous losses for mechanical-draft towers (velocity based on packing crosssectional area):

   $N \simeq 3$

Note: N is the loss coefficient defined by Equation 4.8.21, with velocity based on cross-sectional area for air flow underneath the packing in items 1 through 4.

Sources: Lowe, H.J. and Christie, D.G. 1961. Heat transfer and pressure drop data on cooling tower packings, and model studies of the resistance of natural draft towers to airflow. Paper 113, *International Developments in Heat Transfer Proc. of the International Heat Transfer Conference*, Boulder, CO, ASME, New York; Singham, J.R. 1990. Natural draft towers, in *Hemisphere Handbook of Heat Exchanger Design*, Sec. 3.12.3, Hewitt, G.E., Coord. Ed., Hemisphere, New York. With permission.

Cooling Tower Behavior

There are a number of computer programs available that use variations of Merkel's method to calculate the cooling tower performance, for example, TEFRI (Bourillot, 1983), VERA2D-84 (Mujamdar et al., 1985), CTOWER (Mills, 1995). These programs can be used to perform parametric studies to obtain the response of cooling towers to environmental, duty, and design changes. However, before using such programs, some thought should be given to the important characteristics of cooling tower behavior. For this purpose, it is useful to consider a graphical representation of Merkel's theory for a counterflow tower. Figure 4.8.10 shows a chart with moist air enthalpy plotted vs. water enthalpy (or, equivalently, water temperature) at 1 atm pressure. The *saturation curve* $h_s(T_s)$ is the enthalpy of saturated air. The *operating lines* $h_G(h_L)$ are given by Equation (4.8.13) and relate the air enthalpy to the water enthalpy at each location in the packing. The slope of an operating line is L/G. Since the assumption $T_s = T_L$ is made in Merkel's method, vertical lines on the chart connect h_s and h_G at each location in the packing. The driving force for enthalpy transfer, $(h_s - h_G)$, is the vertical distance between the saturation curve and the operating line. The integral in Equation (4.8.12) averages the reciprocal of this distance. By using this chart, a number of observations about cooling tower behavior can be made.

1. Figure 4.8.10 shows the effect of L/G for fixed water inlet and outlet temperatures, and fixed inlet air temperature and humidity. If we imagine L to be fixed as well, we see that as G decreases, the driving forces decrease, and so a larger NTU is required.
2. The minimum NTU required corresponds to $L/G = 0$, that is, an infinite air flow rate, for which the operating line is horizontal.
3. Due to the curvature of the operating line, it is possible for the operating line to be tangent to the saturation curve. The indicated NTU is then infinite, which tells us that the air flow rate must be increased in order to achieve the desired water cooling range.

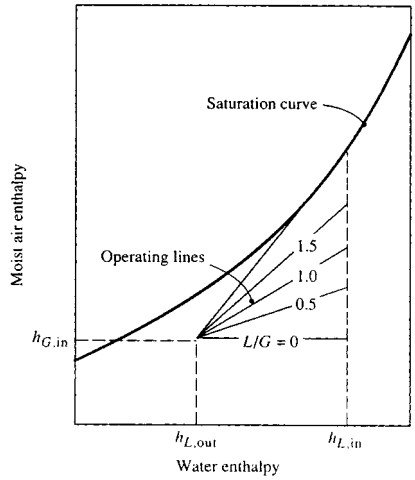


FIGURE 4.8.10 Counterflow cooling tower operating lines for various water-to-air flow-rate ratios shown on an enthalpy chart.

4. For a mechanical-draft tower, the optimal value of *L/G* lies between the two limits described in items 2 and 3 above. If *L/G* is large, the required height of packing is large, and the capital cost will be excessive. If *L/G* is small, the fan power will be excessive (since fan power is proportional to air volume flow rate times pressure drop).

Range and Approach

Cooling tower designers and utility engineers have traditionally used two temperature differences to characterize cooling tower operation. The *range* is the difference between the water inlet and outlet temperatures (also called simply the hot and cold water temperatures). The *approach* is the difference between the outlet water temperature and the wet-bulb temperature of the entering (ambient) air. The approach characterizes cooling tower performance; for a given inlet condition, a larger packing will produce a smaller approach to the ambient wet-bulb temperature, and hence a lower water outlet temperature. (The water cannot be cooled below the ambient wet-bulb temperature.) The approach concept is useful because the ambient dry-bulb temperature has little effect on performance at usual operating conditions (for a specified wet-bulb temperature).

Cooling Demand Curves

Electrical utility engineers have found it convenient to use charts of *cooling demand curves* to evaluate packing specifications. Figure 4.8.11 is an example of such a chart, on which the required NTU, for a given inlet air wet-bulb temperature and range, is plotted vs. *L/G* with the approach as a parameter. Such a plot is possible since the inlet air dry-bulb temperature has only a small effect under usual operating conditions. Now, if it is possible to correlate the mass transfer conductance as

$$\frac{g_m a}{L} = C\left(\frac{L}{G}\right)^{-n} \tag{4.8.22}$$

the NTU of a packing of height *H* is

$$\frac{g_m S}{\dot{m}_L} = \frac{g_m aH}{L} = C\left(\frac{L}{G}\right)^{-n} H \tag{4.8.23}$$

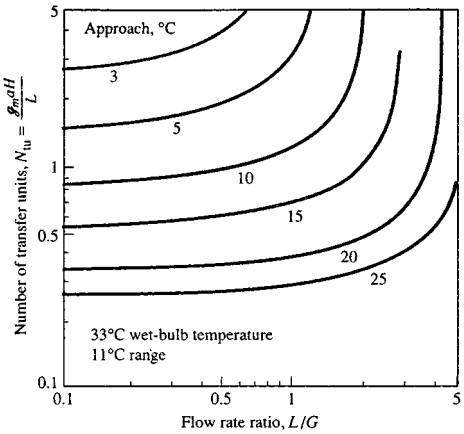


FIGURE 4.8.11 Example of cooling demand curves for a specified wet-bulb temperature and range: NTU vs. flow rate ratio for a fixed approach.

Equation (4.8.23) can also be plotted on the chart to give the *packing capability line*. For a required approach, the *operating point* of the tower is the intersection of the cooling demand curve and packing capability line. Charts of cooling demand curves are available (Cooling Tower Institute, 1967; Kelly, 1976). Correlations of the form of Equation (4.8.22) do not necessarily fit experimental data well. A dependence $g_m a \propto L^{1-n} G^n$ is implied and, in the past, experimental data were often forced to fit such a relation. If the $g_m a$ correlation does not have the form of Equation (4.8.22), the NTU cannot be plotted as a line on a cooling demand chart.

With the almost universal use of computers and the availability of suitable computer programs, one can expect less use of cooling demand charts in the future. The major sources of error in the predictions made by these programs are related to nonuniform air and water flow, and the correlations of packing mass transfer and pressure drop experimental data. The experimental data are obtained in small-scale test rigs, in which it is impossible to simulate many features of full-size towers — for example, nonuniform flow due to entrance configuration, nonuniform wetting of the packing, and, in the case of counterflow towers, the effect of spray above the packing and rain below the packing. Furthermore, since testing of packings in small-scale test rigs is itself not easy, considerable scatter is seen in such test data. Correlations of the data typically have root mean square errors of 10 to 20%.

Legionnaires' Disease

Legionnaires' disease is a form of pneumonia caused by a strain of legionnella bacteria (sero group I). Smokers and sick people are particularly vulnerable to the disease. Major outbreaks have occurred at conventions and in hospitals, for which the source of the bacteria has been traced to cooling towers of air-conditioning systems. The bacteria require nutrients such as algae or dead bacteria in sludge, and thrive if iron oxides are present. However, properly designed, installed, and maintained cooling towers have never been implicated in an outbreak of the disease. Key requirements to be met include the following:

1. Mist (drift) eliminators should be effective.
2. The tower should be located so as to minimize the possibility of mist entering a ventilation system.
3. Corrosion in the tower and water lines should be minimized by use of glass fiber, stainless steel, and coated steel.
4. The design should facilitate inspection and cleaning, to allow early detection and remedy of sludge buildup.
5. Water treatment and filtration procedures should meet recommended standards.

References

Bourillot, C. 1983. *TEFRI: Numerical Model for Calculating the Performance of an Evaporative Cooling Tower*, EPRI CS-3212-SR, Electric Power Research Institute, Palo Alto, CA.

Cooling Tower Institute, 1967. *Cooling Tower Performance Curves*, the Institute, Houston.

Kelly, N.W. 1976. *Kelly's Handbook of Cross-Flow Cooling Tower Performance*, Neil W. Kelly and Associates, Kansas City, MO.

Kintner-Meyer, M. and Emery, A.F. 1995. Cost-optimal design of cooling towers, *ASHRAE J.*, April, 46–55.

Merkel, F. 1925. Verdunstungskühlung, *Forschungsarb. Ing. Wes.*, no. 275.

Mills, A.F. 1995. *Heat and Mass Transfer*, Richard D. Irwin, Chicago.

Majumdar, A.K., Singhal, A.K., and Spalding, D.B. 1985. *VERA2D-84: A Computer Program for 2-D Analysis of Flow, Heat and Mass Transfer in Evaporative Cooling Towers*, EPRI CS-4073, Electric Power Research Institute, Palo Alto, CA.

Further Information

Baker, D. 1984. *Cooling Tower Performance*, Chemical Publishing Company, New York.

Johnson, B.M. Ed. 1990. *Cooling Tower Performance Prediction and Improvement*, Vols. 1 and 2, EPRI GS-6370, Electric Power Research Institute, Palo Alto, CA.

Singham, J.R. 1990. Natural draft towers, in *Hemisphere Handbook of Heat Exchanger Design*, Section 3.12.3, Hewitt, G.E., Coord Ed., Hemisphere Publishing, New York.

Stoeker, W.F. and Jones, J.W. 1982. *Refrigeration and Air Conditioning*, 2nd ed., McGraw-Hill, New York.

Webb, R.L. 1988. A critical review of cooling tower design methods, in *Heat Transfer Equipment Design*, Shah, R.K., Subba Rao, E.C., and Mashelkar, R.A., Eds., Hemisphere Publishing, Washington, D.C.

Heat Pipes

Larry W. Swanson

Introduction

The heat pipe is a vapor-liquid phase-change device that transfers heat from a hot reservoir to a cold reservoir using **capillary forces** generated by a **wick** or porous material and a working fluid. Originally conceived by Gaugler in 1944, the operational characteristics of heat pipes were not widely publicized until 1963 when Grover and his colleagues at Los Alamos Scientific Laboratory independently reinvented the concept. Since then many types of heat pipes have been developed and used by a wide variety of industries.

Figure 4.8.12 shows a schematic of a heat pipe aligned at angle ψ relative to the vertical axis (gravity vector). The heat pipe is composed of a container lined with a wick that is filled with liquid near its saturation temperature. The vapor-liquid interface, usually found near the inner edge of the wick, separates the liquid in the wick from an open vapor core. Heat flowing into the evaporator is transferred through the container to the liquid-filled wicking material, causing the liquid to evaporate and vapor to flow into the open core portion of the evaporator. The capillary forces generated by the evaporating interface increase the pressure difference between the vapor and liquid. The vapor in the open core flows out of the evaporator through the adiabatic region (insulated region) and into the condenser. The vapor then condenses, generating capillary forces similar, although much less in magnitude, to those in the evaporator. The heat released in the condenser passes through the wet wicking material and container out into the cold reservoir. The condensed liquid is then pumped, by the liquid pressure difference due to the net capillary force between the evaporator and condenser, out of the condenser back into the

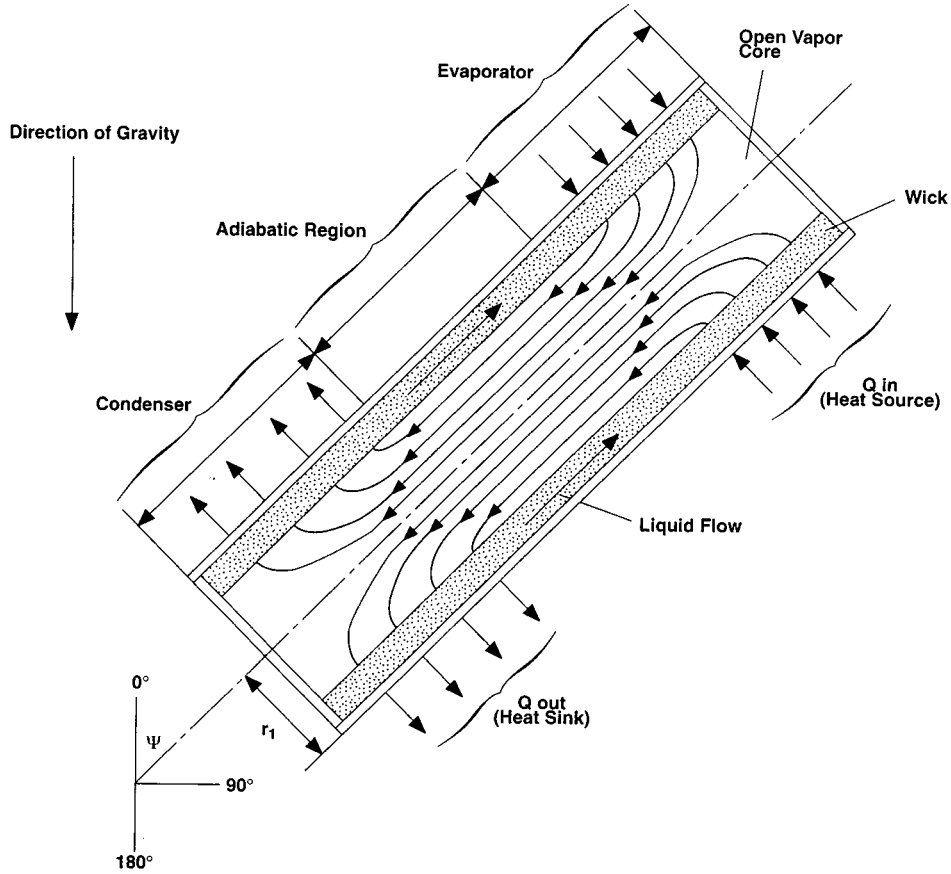


FIGURE 4.8.12 Schematic of a typical heat pipe.

evaporator. Proper selection and design of the pipe container, working fluid, and wick structure are essential to the successful operation of a heat pipe. The **heat transfer limitations**, **effective thermal conductivity**, and axial temperature difference define the operational characteristics of the heat pipe.

Heat Pipe Container, Working Fluid, and Wick Structures

The container, working fluid, and wick structure of a heat pipe determine its operational characteristics. One of the most important considerations in choosing the material for the heat pipe container and wick is its compatibility with the working fluid. Degradation of the container or wick and contamination of the working fluid due to chemical reaction can seriously impair heat pipe performance. For example, noncondensable gas created during a chemical reaction eventually can accumulate near the end of the condenser, decreasing the condensation surface area. This reduces the ability of the heat pipe to transfer heat to the external heat sink. The material and geometry of the heat pipe container also must have a high burst strength, low weight, high thermal conductivity, and low porosity.

Using the proper working fluid for a given application is another critical element of proper heat pipe operation. The working fluid must have good thermal stability properties at the specified operational temperature and pressure. The operational temperature range of the working fluid has to lie between its triple point and its critical point for liquid to exist in the wicking material. The **wettability** of the working fluid contributes to its capillary pumping and priming capability. High-surface-tension fluids are commonly used in heat pipes because they provide the capillary pumping and wetting characteristics necessary for proper operation. Other desirable thermophysical properties include a high liquid thermal

TABLE 4.8.4 Thermophysical Properties of Some Heat-Pipe Fluids

Temperature (°C)	Latent Heat (kJ /kg)	Liquid Density (kg/m³)	Vapor Density (kg/m³)	Liquid Thermal Conductivity (W/m°C)	Liquid Viscosity (cP)	Vapor Viscosity (cP, × 10²)	Vapor Pressure (bars)	Vapor Specific Heat (kJ/kg°C)	Liquid Surface Tension (N/m × 10²)
				Methanol					
–50	1194	843.5	0.01	0.210	1.700	0.72	0.01	1.20	3.26
–30	1187	833.5	0.01	0.208	1.300	0.78	0.02	1.27	2.95
–10	1182	818.7	0.04	0.206	0.945	0.85	0.04	1.34	2.63
10	1175	800.5	0.12	0.204	0.701	0.91	0.10	1.40	2.36
30	1155	782.0	0.31	0.203	0.521	0.98	0.25	1.47	2.18
50	1125	764.1	0.77	0.202	0.399	1.04	0.55	1.54	2.01
70	1085	746.2	1.47	0.201	0.314	1.11	1.31	1.61	1.85
90	1035	724.4	3.01	0.199	0.259	1.19	2.69	1.79	1.66
110	980	703.6	5.64	0.197	0.211	1.26	4.98	1.92	1.46
130	920	685.2	9.81	0.195	0.166	1.31	7.86	1.92	1.25
150	850	653.2	15.90	0.193	0.138	1.38	8.94	1.92	1.04
				Water					
20	2448	998.0	0.02	0.603	1.00	0.96	0.02	1.81	7.28
40	2402	992.1	0.05	0.630	0.65	1.04	0.07	1.89	7.00
60	2359	983.3	0.13	0.649	0.47	1.12	0.20	1.91	6.66
80	2309	972.0	0.29	0.668	0.36	1.19	0.47	1.95	6.26
100	2258	958.0	0.60	0.680	0.28	1.27	1.01	2.01	5.89
120	2200	945.0	1.12	0.682	0.23	1.34	2.02	2.09	5.50
140	2139	928.0	1.99	0.683	0.20	1.41	3.90	2.21	5.06
160	2074	909.0	3.27	0.679	0.17	1.49	6.44	2.38	4.66
180	2003	888.0	5.16	0.669	0.15	1.57	10.04	2.62	4.29
200	1967	865.0	7.87	0.659	0.14	1.65	16.19	2.91	3.89
				Potassium					
350	2093	763.1	0.002	51.08	0.21	0.15	0.01	5.32	9.50
400	2078	748.1	0.006	49.08	0.19	0.16	0.01	5.32	9.04
450	2060	735.4	0.015	47.08	0.18	0.16	0.02	5.32	8.69
500	2040	725.4	0.031	45.08	0.17	0.17	0.05	5.32	8.44
550	2020	715.4	0.062	43.31	0.15	0.17	0.10	5.32	8.16
600	2000	705.4	0.111	41.81	0.14	0.18	0.19	5.32	7.86
650	1980	695.4	0.193	40.08	0.13	0.19	0.35	5.32	7.51
700	1960	685.4	0.314	38.08	0.12	0.19	0.61	5.32	7.12
750	1938	675.4	0.486	36.31	0.12	0.20	0.99	5.32	6.72
800	1913	665.4	0.716	34.81	0.11	0.20	1.55	5.32	6.32
850	1883	653.1	1.054	33.31	0.10	0.21	2.34	5.32	5.92

conductivity, high latent heat of vaporization, low liquid viscosity, and a low vapor viscosity. Table 4.8.4 gives the thermophysical properties for three typical heat pipe working fluids that span a fairly wide operating temperature range. The thermophysical properties for other heat pipe working fluids can be obtained from Dunn and Reay (1982) and Peterson (1994).

The wick structure and working fluid generate the capillary forces required to (1) pump liquid from the condenser to the evaporator and (2) keep liquid evenly distributed in the wicking material. Heat pipe wicks can be classified as either homogeneous wicks or composite wicks. Homogeneous wicks are composed of a single material and configuration. The most common types of homogeneous wicks include wrapped screen, sintered metal, axial groove, annular, crescent, and arterial. Composite wicks are composed of two or more materials and configurations. The most common types of composite wicks include variable screen mesh, screen-covered groove, screen slab with grooves, and screen tunnel with

grooves. Regardless of the wick configuration, the desired material properties and structural characteristics of heat pipe wick structures are a high thermal conductivity, high wick porosity, small capillary radius, and high wick permeability. Table 4.8.2 gives the geometric properties of some commonly used homogeneous wicks. The properties of other wick structures, including nonhomogenous types, can be obtained from Peterson (1994). The container, wick structure, and working fluid are used to determine the heat transfer limitations of heat pipes.

Heat Transfer Limitations

Heat pipes undergo various heat transfer limitations depending on the working fluid, the wick structure, the dimensions of the heat pipe, and the heat pipe operational temperature. Figure 4.8.13 gives a qualitative description of the various heat transfer limitations, which include vapor-pressure, sonic, entrainment, capillary, and boiling limitations. The composite curve enclosing the shaded region in Figure 4.8.13 gives the maximum heat transfer rate of the heat pipe as a function of the operational temperature. The figure shows that as the operational temperature increases, the maximum heat transfer rate of the heat pipe is limited by different physical phenomena. As long as the operational heat transfer rate falls within the shaded region, the heat pipe will function properly.

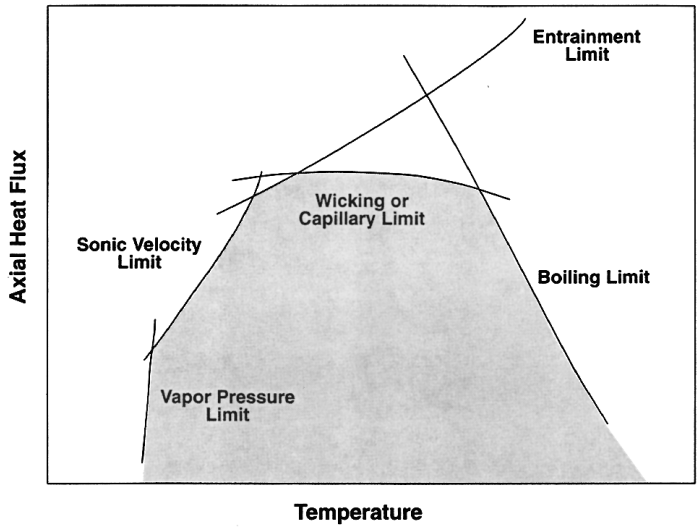


FIGURE 4.8.13 Heat transfer limitations in heat pipes.

The vapor-pressure limitation (or viscous limitation) in heat pipes develops when the pressure drop in the vapor core reaches the same order of magnitude as the vapor pressure in the evaporator. Under these conditions, the pressure drop due to flow through the vapor core creates an extremely low vapor pressure in the condenser preventing vapor from flowing in the condenser. A general expression for the vapor-pressure limitation is (Dunn and Reay, 1982)

$$Q_{vp,\max} = \frac{\pi r_v^4 h_{fg} \rho_{v,e} P_{v,e}}{12 \mu_{v,e} l_{\text{eff}}} \tag{4.8.24}$$

where r_v is the cross-sectional radius of the vapor core (m), h_{fg} is the latent heat of vaporization (J/kg), $\rho_{v,e}$ is the vapor density in the evaporator (kg/m^3), P_{ve} is the vapor pressure in the evaporator (Pa), and $\mu_{v,e}$ is the vapor viscosity in the evaporator (N sec/m^2). l_{eff} is the effective length of the heat pipe (m) equal to $l_{\text{eff}} = 0.5(l_e + 2l_a + l_c)$. The vapor-pressure limitation can occur during the start-up of heat pipes at the lower end of the working-fluid-temperature range.

The sonic limitation also can occur in heat pipes during start-up at low temperatures. The low temperature produces a low vapor density, thereby reducing the speed of sound in the vapor core. Thus, a sufficiently high mass flow rate in the vapor core can cause sonic flow conditions and generate a shock wave that chokes the flow and restricts the pipes ability to transfer heat to the condenser. Dunn and Reay (1982) give an expression for the sonic limitation that agrees very well with experimental data,

$$Q_{s,\max} = 0.474 A_v h_{fg} \left(\rho_v P_v\right)^{1/2} \tag{4.8.25}$$

where A_v is the cross-sectional area of the vapor core (m^2). The sonic limitation should be avoided because large temperature gradients occur in heat pipes under choked-flow conditions.

The entrainment limitation in heat pipes develops when the vapor mass flow rate is large enough to shear droplets of liquid off the wick surface causing dry-out in the evaporator. A conservative estimate of the maximum heat transfer rate due to entrainment of liquid droplets has been given by Dunn and Reay (1982) as

$$Q_{e,\max} = A_v h_{fg} \left[\frac{\rho_v \sigma_l}{2 r_{c,\text{ave}}}\right]^{1/2} \tag{4.8.26}$$

where σ_l is the surface tension (N/m) and $r_{c,\text{ave}}$ is the average capillary radius of the wick. Note that for many applications $r_{c,\text{ave}}$ is often approximated by $r_{c,e}$.

The capillary limitation in heat pipes occurs when the net capillary forces generated by the vapor-liquid interfaces in the evaporator and condenser are not large enough to overcome the frictional pressure losses due to fluid motion. This causes the heat pipe evaporator to dry out and shuts down the transfer of heat from the evaporator to the condenser. For most heat pipes, the maximum heat transfer rate due to the capillary limitation can be expressed as (Chi, 1976).

$$Q_{c,\max} = \left[\frac{\rho_l \sigma_l h_{fg}}{\mu_l}\right]\left[\frac{A_w K}{l_{\text{eff}}}\right]\left(\frac{2}{r_{c,e}} - \left[\frac{\rho_l}{\sigma_l}\right] g L_t \cos\Psi\right) \tag{4.8.27}$$

where K is the wick permeability (m^2), A_w is the wick cross-sectional area (m^2), ρ_l is the liquid density (m^3), μ_l is the liquid viscosity (N sec/m^2), $r_{c,e}$ is the wick capillary radius in the evaporator (m), g is the acceleration due to gravity (9.8 m/sec^2), and L_t is the total length of the pipe (m). For most practical operating conditions, this limitation can be used to determine maximum heat transfer rate in heat pipes.

The boiling limitation in heat pipes occurs when the degree of liquid superheat in the evaporator is large enough to cause the nucleation of vapor bubbles on the surface of the wick or the container. Boiling is usually undesirable in heat pipes because local hot spots can develop in the wick, obstructing the flow of liquid in the evaporator. An expression for the boiling limitation is (Chi, 1976)

$$Q_{b,\max} = \frac{4\pi l_{\text{eff}} k_{\text{eff}} T_v \sigma_v}{h_{fg} \rho_l \ln\left(r_i / r_v\right)}\left(\frac{1}{r_n} - \frac{1}{r_{c,e}}\right) \tag{4.8.28}$$

where k_{eff} is the effective thermal conductivity of the composite wick and working fluid (W/m K), T_v is the vapor saturation temperature (K), r_i is the inner container radius (m), r_n is the nucleation radius (equal to 2.00×10^{-6} m in the absence of noncondensable gas).

Effective Thermal Conductivity and Heat Pipe Temperature Difference

One key attribute of the heat pipe is that it can transfer a large amount of heat while maintaining nearly isothermal conditions. The temperature difference between the external surfaces of the evaporator and the condenser can be determined from the following expression

$$\Delta T = R_t Q \tag{4.8.29}$$

where R_t is the total thermal resistance (K/W) and Q is the heat transfer rate (W). Figure 4.8.14 shows the thermal resistance network for a typical heat pipe and the associated thermal resistances. In most cases, the total thermal resistance can be approximated by

$$R_t = R_1 + R_2 + R_3 + R_5 + R_7 + R_8 + R_9 \tag{4.8.30}$$

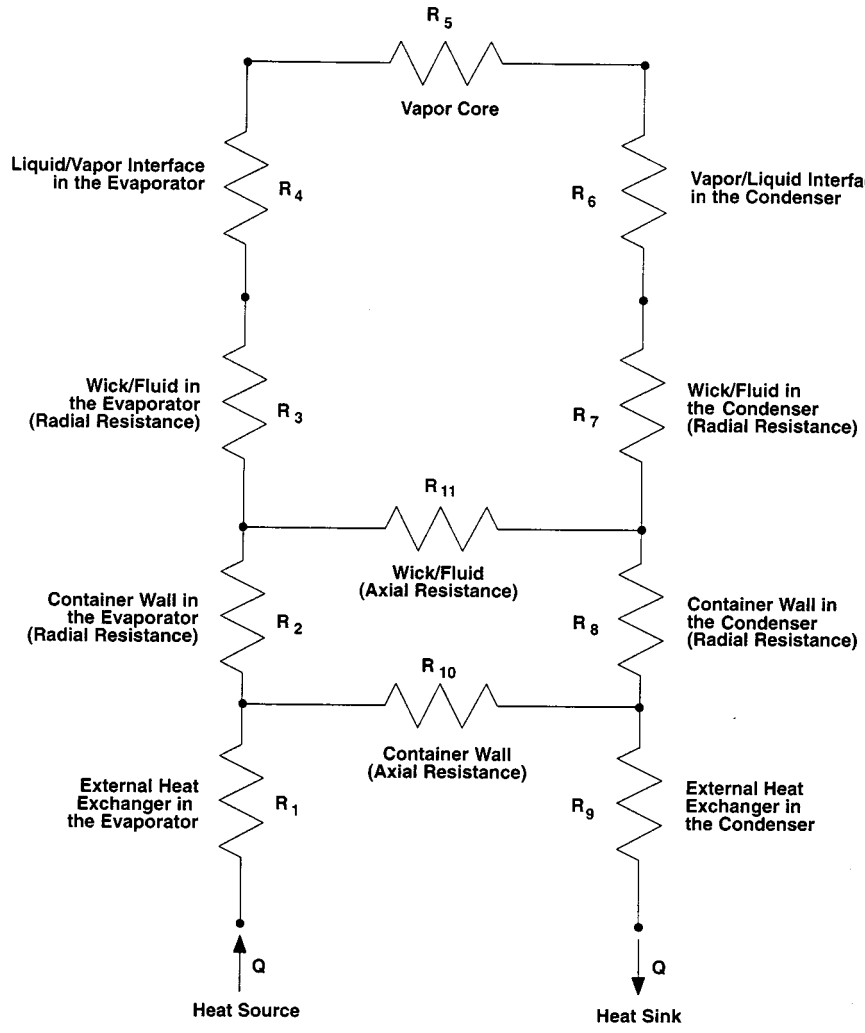


FIGURE 4.8.14 Thermal resistance network in a heat pipe.

The reader is referred to Peterson (1994) for the specific mathematical relationships used to calculate each thermal resistance. The effective thermal conductivity of the heat pipe is defined as the heat transfer rate divided by the temperature difference between the heat source and heat sink,

$$k_{\text{eff}} = \frac{L_t}{R_t A_t} \tag{4.8.31}$$

where A_t is the overall cross-sectional area of the pipe (m^2). Under normal operating conditions, the total thermal resistance is relatively small, making the external surface temperature in the evaporator approximately equal to that in the condenser. Thus, the effective thermal conductivity in a heat pipe can be very large (at least an order of magnitude larger than that of aluminum).

Design Example

Design a water heat pipe to transport 80 W of waste heat from an electronics package to cooling water. The heat pipe specifications are

1. Axial orientation — complete gravity-assisted operation (condenser above the evaporator; ψ = 180°)
2. Maximum heat transfer rate — 80 W
3. Nominal operating temperature — 40°C
4. Inner pipe diameter — 3 cm
5. Pipe length — 25 cm evaporator length, 50 cm adiabatic section, and 25 cm condenser length

The simplest type of wick structure to use is the single-layer wire mesh screen wick shown in Table 4.8.5. The geometric and thermophysical properties of the wick have been selected as (this takes some forethought)

$$d = 2.0 \times 10^{-5} \text{ m}$$

$$w = 6.0 \times 10^{-5} \text{ m}$$

$$\frac{1}{2N} = r_c = {}^1/_2(2.0 \times 10^{-5} + 6 \times 10^{-5}) = 4.0 \times 10^{-5} \text{ m}$$

$$\varepsilon = 1$$

$$k_{\text{eff}} = k_1 = 0.630 \ \frac{\text{W}}{\text{mK}}$$

$$t_w = 1.0 \times 10^{-3} \text{ m}$$

$$K = \frac{t_w^2}{12} = \frac{\left(1 \times 10^{-3}\right)^2}{12} = 8.33 \times 10^{-8} \text{ m}^2$$

The other heat pipe geometric properties are

$$r_v = r_i - t_w = 0.015 - 0.001 = 0.014 \text{ m}$$

$$l_{\text{eff}} = \frac{0.25 + 0.25}{2} + 0.5 = 0.75 \text{ m}$$

$$L_t = 0.25 + 0.50 + 0.25 + 1.0 \text{ m}$$

$$A_w = \pi(r_i^2 - r_v^2) = \pi[(0.015)^2 - (0.014)^2] = 9.11 \times 10^{-5} \text{ m}^2$$

$$A_v = \pi r_v^2 = \pi(0.014)^2 = 6.16 \times 10^{-4} \text{ m}^2$$

The thermophysical properties of water at 40°C are (see Table 4.8.4):

$$\rho_l = 992.1 \text{ kg/m}^3$$

$$\rho_v = 0.05 \text{ kg/m}^3$$

$\sigma_l = 2.402 \times 10^6$ J/kg
$\mu_l = 6.5 \times 10^{-3}$ kg/m sec
$\mu_v = 1.04 \times 10^{-4}$ kg/m sec
$P_v = 7000$ Pa

The various heat transfer limitations can now be determined to ensure the heat pipe meets the 80 W heat transfer rate specification. The vapor-pressure limitation is

$$Q_{vp,\max} = \frac{\pi(0.014)^4\left(2.402\times10^6\right)(0.05)(7000)}{12\left(1.04\times10^{-4}\right)(0.75)} = 1.08\times10^5 \text{ W} \tag{4.8.32}$$

The sonic limitation is

$$\begin{aligned} Q_{s,\max} &= 0.474\left(6.16\times10^{-4}\right)\left(2.402\times10^6\right)\left[(0.05)(7000)\right]^{1/2} \\ &= 1.31\times10^4 \text{ W} \end{aligned} \tag{4.8.33}$$

The entrainment limitation is

$$\begin{aligned} Q_{e,\max} &= \left(6.16\times10^{-4}\right)\left(2.402\times10^6\right)\left[\frac{(0.05)(0.07)}{2\left(4.0\times10^{-5}\right)}\right]^{1/2} \\ &= 9.79\times10^3 \text{ W} \end{aligned} \tag{4.8.34}$$

Noting that $\cos\psi = -1$, the capillary limitation is

$$\begin{aligned} Q_{c,\max} &= \left[\frac{(992.1)(0.07)\left(2.402\times10^6\right)}{6.5\times10^{-3}}\right]\left[\frac{\left(9.11\times10^{-5}\right)\left(8.33\times10^{-8}\right)}{0.75}\right]\left[\frac{2}{4.0\times10^{-5}}+\frac{992.1}{0.07}9.8(1.0)\right] \\ &= 4.90\times10^4 \text{ W} \end{aligned} \tag{4.8.35}$$

Finally, the boiling limitation is

$$\begin{aligned} Q_{b,\max} &= \frac{4\pi(0.75)(0.63)(313)(0.07)}{\left(2.402\times10^6\right)(992.1)\ln\left(\frac{0.015}{0.014}\right)}\left[\frac{1}{2.0\times10^{-6}}-\frac{1}{4.0\times10^{-5}}\right] \\ &= 0.376 \text{ W} \end{aligned} \tag{4.8.36}$$

All of the heat transfer limitations, with the exception of the boiling limitation, exceed the specified heat transfer rate of 80 W. The low value of 0.376 W for the boiling limitation strongly suggests that the liquid will boil in the evaporator and possibly cause local dry spots to develop. The reason the liquid boils is because the effective thermal conductivity of the wick is equal to the conductivity of the liquid, which is very low in this case. Because the liquid is saturated at the vapor-liquid interface, a low effective thermal conductivity requires a large amount of wall superheat which, in turn, causes the liquid to boil. This problem can be circumvented by using a high conductivity wire mesh or sintered metal wick, which greatly increases the effective conductivity. It should be noted, however, that because porous wicks have

TABLE 4.8.5 Physical Properties of Wick Structures

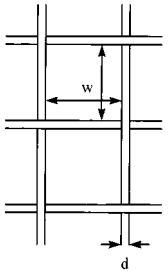
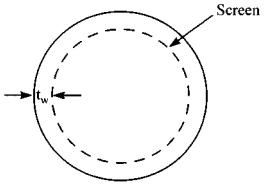
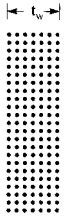
	Wick Type[a]	Thermal Conductivity	Porosity	Minimum Capillary Radius	Permeability
Single-layer wire mesh screens (heat-pipe axis in the plane of the paper in this sketch)  $1/N = d + w$ N = number of aperatures per unit length	 Annular	$k_{eff} = k_e$	$\varepsilon = 1$	$r_c = 1/(2N)$	$K = t_w^2/12$
	Multiple wire mesh screens,[b] plain or sintered (screen dimensions as for single layers illustrated above)	$k_{eff} = \dfrac{k_e\left[k_e + k_s - (1-\varepsilon)(k_e - k_s)\right]}{k_e + k_s + (1-\varepsilon)(k_e - k_s)}$	Estimated from $\varepsilon = 1 - (\pi N d)/4$	$r_c = 1/(2N)$	$k = \dfrac{d^2\varepsilon^2}{122(1-\varepsilon)^2}$

TABLE 4.8.5 Physical Properties of Wick Structures (continued)

	Wick Type[a]		Thermal Conductivity	Porosity	Minimum Capillary Radius	Permeability
	Unconsolidated packed spherical particles (d = average particle diameter)	Plain	$k_{\text{eff}} = \dfrac{k_e\left[2k_e + k_s - 2(1-\varepsilon)(k_e - k_s)\right]}{2k_e + k_s + (1-\varepsilon)(k_e - k_s)}$	Estimated from (assuming cubic packing) ε = 0.48	$r_c = 0.21d$	$k = \dfrac{d^2\varepsilon^2}{150(1-\varepsilon)^2}$
		Sintered	$k_{\text{eff}} = \dfrac{k_e\left[2k_s + k_e - 2\varepsilon(k_s - k_e)\right]}{2k_s + k_e + \varepsilon(k_s - k_e)}$			
	Sintered metal fibers (d = fiber diameter)		$k_{\text{eff}} = \varepsilon^2 k_e (1-\varepsilon)^2 k_s + \dfrac{4\varepsilon(1-\varepsilon)k_e k_s}{k_e + k_s}$	Use manufacturers data	$r_c = \dfrac{d}{2(1-\varepsilon)}$	$k = C_1 \dfrac{y^2 - 1}{y^2 - 1}$ where $y = 1 + \dfrac{C_2 d^2 \varepsilon^3}{(1-\varepsilon)^2}$ $C_1 = 6.0 \times 10^{-10}$ m^2 $C_2 = 3.3 \times 10^7$ 1/m^2

[a] The axis of the pipe and direction of fluid flow are normal to the paper.

[b] These wicks are positioned so that the layers follow the contours of the iner surface of the pipe wall.

Revised from Peterson, G.P., *An Introduction to Heat Pipes Modeling, Testing, and Applications*, John Wiley & Sons, New York, 1994.

lower permeabilities, the capillary limitation should be lower as well. Let's try a sintered particle wick made of copper with the following properties (see Table 4.8.5):

$$d = 1.91 \times 10^{-4} \text{ m}$$

$$r_{c,3} = 0.21d = 4.0 \times 10^{-5} \text{ m (same as before)}$$

$$\varepsilon = 0.48$$

$$K = \frac{\left(1.91\times10^{-4}\right)^2 (0.48)}{150(1-0.48)^2} = 2.07\times10^{-10} \text{ m}^2$$

$$k_s = 400 \frac{\text{W}}{\text{mK}} \text{ (copper)}$$

$$k_1 = 0.630 \frac{\text{W}}{\text{mK}} \text{ (water)}$$

$$k_{\text{eff}} = \frac{400\left[2(400) + 0.63 - 2(0.48)(400 - 0.63)\right]}{2(400) + 0.63 + 0.48(400 - 0.63)} = 168 \text{ W/mK}$$

All other geometric and thermophysical properties are the same. The heat transfer limitations affected by the change in wick structure are the capillary and boiling limitations. The sintered metal wick produces a capillary limitation of

$$Q_{c,\max} = \left[\frac{(992.1)(0.07)\left(2.402\times10^6\right)}{6.5\times10^{-3}}\right]\left[\frac{\left(9.11\times10^{-5}\right)\left(2.07\times10^{-10}\right)}{0.75}\right]\left[\frac{2}{4.0\times10^{-5}} + \frac{992.1}{0.07}9.8(1.0)\right] \tag{4.8.37}$$

$$= 122 \text{ W}$$

The boiling limitation for the sintered wick is

$$Q_{b,\max} = \frac{4\pi(0.75)(168)(313)(0.07)}{\left(2.402\times10^6\right)(992.1)\ln\left(\frac{0.015}{0.014}\right)}\left[\frac{1}{2.0\times10^{-6}} - \frac{1}{4.0\times10^{-5}}\right] \tag{4.8.38}$$

$$= 100 \text{ W}$$

This design now meets all the specifications defined in the problem statement.

Application of Heat Pipes

Heat pipes have been applied to a wide variety of thermal processes and technologies. It would be an impossible task to list all the applications of heat pipes; therefore, only a few important industrial applications are given in this section. In the aerospace industry, heat pipes have been used successfully in controlling the temperature of vehicles, instruments, and space suits. Cryogenic heat pipes have been applied in (1) the electronics industry for cooling various devices (e.g., infrared sensors, parametric amplifiers) and (2) the medical field for cryogenic eye and tumor surgery. Heat pipes have been employed to keep the Alaskan tundra frozen below the Alaskan pipeline. Other cooling applications include (1) turbine blades, generators, and motors; (2) nuclear and isotope reactors; and (3) heat collection from exhaust gases, solar and geothermal energy.

In general, heat pipes have advantages over many traditional heat-exchange devices when (1) heat has to be transferred isothermally over relatively short distances, (2) low weight is essential (the heat pipe is a passive pumping device and therefore does not require a pump), (3) fast thermal-response times are required, and (4) low maintenance is mandatory.

Defining Terms

Capillary force: The force caused by a curved vapor-liquid interface. The interfacial curvature is dependent on the surface tension of the liquid, the contact angle between the liquid wick structure, the vapor pressure, and the liquid pressure.

Effective thermal conductivity: The heat transfer rate divided by the temperature difference between the evaporator and condenser outer surfaces.

Heat transfer limitations: Limitations on the axial heat transfer capacity imposed by different physical phenomena (i.e., vapor pressure, sonic, entrainment, capillary, and boiling limitations).

Wettability: The ability of a liquid to spread itself over a surface. A wetting liquid spreads over a surface whereas a nonwetting liquid forms droplets on a surface.

Wick: A porous material used to generate the capillary forces that circulate fluid in a heat pipe.

References

Chi, S.W. 1976. *Heat Pipe Theory and Practice*, Hemisphere Publishing, Washington, D.C.

Dunn, P.D. and Reay, D.A. 1982. *Heat Pipes*, 3rd ed., Pergamon Press, Oxford, U.K.

Gaugler, R.S. 1944. Heat Transfer Device. U.S. Patent No. 2350348.

Grover, G.M. 1963. Evaporation-Condensation Heat Transfer Device. U.S. Patent No. 3229759.

Peterson, G.P. 1994. *An Introduction to Heat Pipes Modeling, Testing, and Applications*, John Wiley & Sons, New York.

Further Information

Recent developments in heat pipe research and technology can be found in the proceedings from a number of technical conferences: (1) The International Heat Pipe Conference (2) The National Heat Transfer Conference, (3) The ASME Winter Annual Meeting, (4) The AIAA Thermophysics Conference.

Books particularly useful for the design of heat pipes include (1) *Heat Pipe Design Handbook* by Brennan and Kroliczek available from B&K Engineering in Baltimore, M.D. (2) *The Heat Pipe* by Chisholm available from Mills and Boon Limited in London, England, and (3) *Heat Pipes: Construction and Application* by Terpstra and Van Veen available from Elsevier Applied Science in New York, N.Y.

An additional book particularly strong in heat pipe theory is *The Principles of Heat Pipes* by Ivanovskii, Sorokin, and Yagodkin available from Clarendon Press in Oxford, England.

Cooling Electronic Equipment

Vincent W. Antonetti

Introduction

In electronic packages, the thermal resistances to heat transfer from heat source to heat sink are often grouped into an internal resistance and an external resistance. The **internal thermal resistance** R_{int} is conductive and exists between the chip and the module case:

$$R_{int} = \frac{T_{chip} - T_{case}}{P_{chip}} \tag{4.8.39}$$

where P_{chip} is the chip power.

The **external thermal resistance** R_{ext} is primarily convective and exists between the surface of the case of the module and some reference point, typically the temperature of the cooling fluid near the module. In a multichip module, the module power P_m is the sum of the individual chip powers, and the external resistance is

$$R_{\text{ext}} = \frac{T_{\text{case}} - T_{\text{coolant}}}{P_m} \tag{4.8.40}$$

The internal and external resistances are related to the chip junction temperature T_j through the following expression:

$$T_j = \Delta T_{j-\text{chip}} + P_{\text{chip}} R_{\text{int}} + P_m R_{\text{ext}} + \Delta T_{\text{coolant}} + T_{\text{coolant in}} \tag{4.8.41}$$

Many factors are involved in determining the appropriate cooling mode to be used. If the component junction temperature is constrained to approximately 85°C, Table 4.8.6 may be used to make a preliminary selection of the cooling mode. Extended surfaces can often be used to increase the allowable heat fluxes.

TABLE 4.8.6 Maximum Component Heat Flux for Various Cooling Modes

Cooling Mode	W/cm²
Free convection air	0.05
Forced convection air	0.5
Impingement air	1.0
Free convection immersion	1.0
Forced convection immersion	50
Pool boiling	20
Forced convection boiling	100
Jet immersion (single phase)	40
Boiling jet immersion	90

Free Convection Air Cooling of Vertical Printed Circuit Boards

Data have been collected from rack-mounted simulated printed circuit boards (PCBs) (see Figure 4.8.15) and from several actual electronic systems at AT&T Bell Laboratories. Results indicated that existing parallel plate correlations for symmetric isoflux plates (separated by distance "b") could be adapted to PCB conditions. Specifically, for $\text{Ra}_b < 10$ use the equation corresponding to the fully developed laminar boundary layer condition:

$$\text{Nu}_b = 0.144\text{Ra}_b^{0.5} \tag{4.8.42}$$

For $10 < \text{Ra}_b < 1000$, use

$$\text{Nu}_b = \left[\frac{48}{\text{Ra}_b} + \frac{2.5}{\text{Ra}_b^{0.4}}\right]^{-0.5} \tag{4.8.43}$$

where

$$\text{Ra}_b = \frac{g\beta c_p \rho^2 b^5 q''}{\mu k L}$$

For Ra > 1000, the expression for an isolated plate in infinite media is recommended:

$$\text{Nu}_b = 0.524\text{Ra}_b^{0.2} \tag{4.8.44}$$

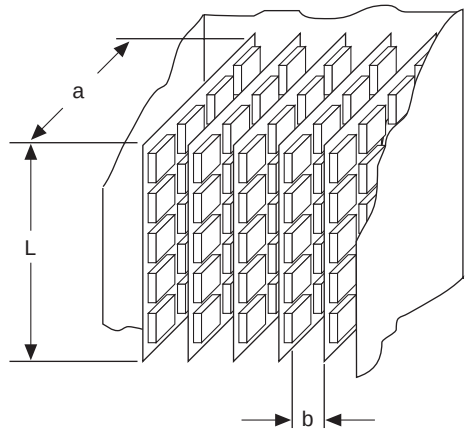


FIGURE 4.8.15 Typical PCB array.

In the previous three expressions air properties are evaluated at the average of the module case and ambient temperatures.

The PCB spacing b_{max} for a given power dissipation which yields the lowest PCB component case temperatures (or which maximizes the rate of heat transfer while maintaining PCB temperatures below the maximum allowable) occurs when the developing boundary layers from adjacent PCBs do not interfere, i.e., so the isolated plate condition is approached as follows: If heat is transferred from both sides of the PCB, let Ra_{ab} = 17,000 and the recommended PCB spacing is $b_{max} = 7.02\xi^{-0.2}$. If heat is transferred from only one side of the PCB, let Ra_{ab} = 5400 and the recommended PCB spacing is $b_{max} = 5.58\xi^{-0.2}$. In both cases

$$\xi = \frac{g\beta\rho^2 \mathrm{Pr} q''}{\mu^2 kL} \tag{4.8.45}$$

Forced Convection Air Cooling of Vertical PCBs

Sparrow et al. (1982, 1984) studied vertical arrays with simulated modules of uniform size, which was 4 modules wide by 17 module rows deep in the flow direction; the modules were 26.7 mm square and 10 mm high; the space between modules was 6.67 mm, and the distance from the top of the module to the adjoining card *Hc* = 16.7 mm. The Nusselt number as a function of module row position for a fully populated array may be determined from Figure 4.8.16. Correction factors to the fully developed Nusselt numbers for the effect of missing modules and the presence of modules whose height differs from others in the array are presented in the cited references.

In actual electronic packages, conditions differ from the relatively ideal setups in laboratories because in real packages the flow is disturbed by the PCB supporting hardware and may extend the entry region effect.

Data from actual computer hardware with PCBs containing a 6 × 4 array of 28 mm modules (4 in the flow direction) were used to develop the following expressions:

$$\mathrm{Nu}_x = C\left\{\mathrm{Re}_{D_h}\left[1 + x/\left(D_h\right)^{-0.836}\right]\right\}^m \tag{4.8.46}$$

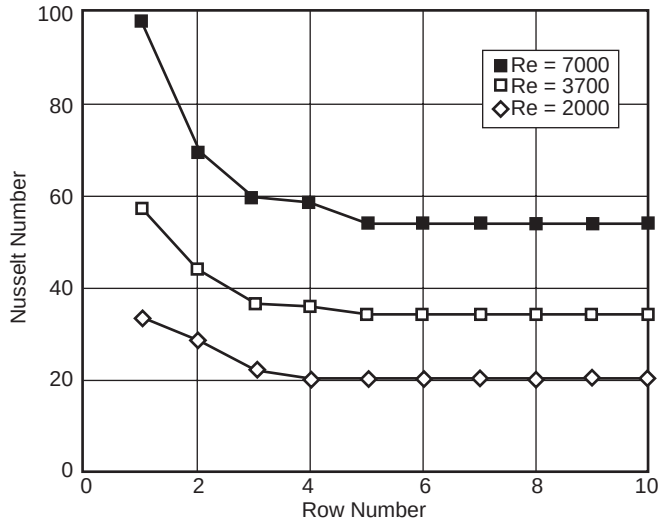


FIGURE 4.8.16 Nusselt number for fully populated array of modules.

For Re < 2000, $C = 0.072$ and $m = 0.70$, and for 2000 < Re < 10,000, $C = 0.056$ and $m = 1.02$, where x is the distance in the flow direction. Because the array was only 4 modules deep, all the modules on the PCB were in the entry region.

Tests have been conducted on a 9 × 7 array of 25-mm-square by 6.4-mm-high blocks cooled by a 3 × 3 array of air jets directed normal to the face of each block. The spent flow exited along the channel formed by the orifice plate and the heat transfer surfaces. Test results were correlated by the relation:

$$N_d = 0.213(z/d)^{-0.376}\,\mathrm{Re}_d^{0.743} \tag{4.8.47}$$

where d is the jet orifice diameter, s is the orifice-to-orifice spacing, and z is the distance from the orifice outlet to the face of the module.

Immersion Cooling

The highly inert perfluorinated liquids, called FC coolants by the 3M Company, are frequently used in **immersion cooling** applications. FC coolants are available with boiling points from 30 to 172°C at atmospheric pressure. FC-75 are FC-77 with boiling points of 100°C are often used in single-phase applications, while FC-72 and FC-87, with boiling points of 56 and 30°C, respectively, are used in systems involving phase change.

Data exist for free convection immersion cooling of a 3 × 3 array of simulated chips mounted on a vertical wall in an enclosure filled with FC-75. Each heat source was 8 mm high by 24 mm wide by 6 mm thick. With the Nusselt and modified Rayleigh numbers based on the heater height, L, the best fit to the data is

$$\mathrm{Nu}_L = 0.279\mathrm{Ra}_b^{0.224} \tag{4.8.48}$$

Air cooling expressions have been modified to make them applicable to free convection immersion cooling of vertical PCB arrays with FC coolants. The Nusselt number (based on PCB spacing "b") at the top of the PCB is

$$\mathrm{Nu}_L = \left[\frac{C}{\mathrm{Ra}_b} + \frac{2.78}{\mathrm{Ra}_b^{0.4}}\right]^{-0.5} \tag{4.8.49}$$

C = 24 when heat transfer is from one side of the PCB, and C = 48 when from both sides.

Nucleate Pool Boiling

A number of investigators have tested small flush heat sources boiling in a pool of dielectric liquid. The heaters ranged from 4 × 4 mm to 12.7 × 12.7 mm. Typical saturated pool boiling data for FC-72 and FC-87 are shown in Figure 4.8.17. Note that a temperature overshoot up to 25°C has been regularly observed for silicon chips in dielectric liquid pools. To estimate the temperature excursion at boiling incipience (q_i''), the following approximation is recommended

$$\Delta T_{\mathrm{ex}} = T_{\mathrm{sat}}\left(p - \frac{2\sigma}{r_b} - p_g\right) - T_{\mathrm{sat}} - C\left(q_i''\right)^n \tag{4.8.50}$$

where

$$C = \mu h_{fg}\left[\frac{c_p}{h_{fg}\mathrm{Pr}^b C_{sf}}\right]^{1/a}\left[\frac{\alpha}{g\left(\rho - \rho_g\right)}\right]^{0.5} \tag{4.8.51}$$

with r_b = 0.25 μm, C_{sf} = 0.003, a = 0.33, and b = 1.7. (Note that $n = 1/a = 3$.)

Park and Bergles (1988) determined the critical heat flux (CHF) as a function of size for flush heaters operating in a saturated pool of R-113. For a 5-mm-wide heater, and for heater heights from 5 to 80 mm, use

$$\frac{q''_{c_{\mathrm{sat}}}}{q''_{c_z}} = 0.86\left[1 + \frac{152}{L^{*3.29}}\right]^{0.14} \tag{4.8.52}$$

where the CHF prediction of Zuber, $q''_{c_z} = \rho_g^{0.5} h_{fg}[\sigma g(\rho_f - \rho_g)]^{0.5}$, and $L^* = L[g(\rho_f - \rho_g)/\sigma]^{0.5}$.

For a 5-mm-high heater, and for heater widths from 2.5 to 70 mm, use

$$\frac{q''_{c_{\mathrm{sat}}}}{q''_{c_z}} = 0.93\left[1 + \frac{52}{I^{1.02}}\right]^{0.14} \tag{4.8.53}$$

where the induced convection parameter is $I = (\rho_f W\sigma/\mu^2)^{0.5}$.

For flush 12.7 × 12.7 mm heaters in FC-72, test data yield $q''_{c_{\mathrm{sat}}}/q''_{c_z} \approx 1.45$. These subcooling data were correlated by

$$\frac{q''_{c_{\mathrm{sub}}}}{q''_{c_{\mathrm{sat}}}} = 1 + \frac{0.0643\rho_f c_{p,f}}{\rho_g h_{fg}}\left[\frac{\rho_g}{\rho_f}\right]^{1/4}\Delta T_{\mathrm{sub}} \tag{4.8.54}$$

Single-Phase and Boiling Forced Convection in Channel Flow

The average Nusselt numbers for 12 flush 12.7 × 12.7 mm heaters (4 rows of 3 sources per row) operating in FC-77 has been correlated by the following expression:

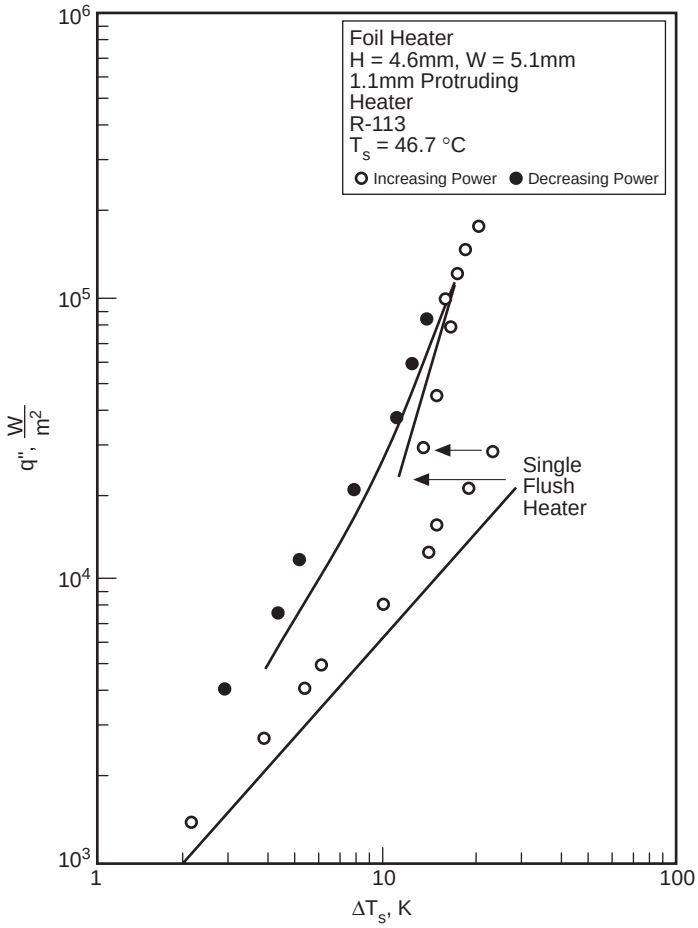


FIGURE 4.8.17 Typical pool boiling curve for small heater.

$$\overline{\mathrm{Nu}}_L = C\left(\mathrm{Re}_{D_h}\right)^m \mathrm{Pr}^{0.11} \tag{4.8.55}$$

For row 1: $C = 0.194$ and $m = 0.60$; for row 2: $C = 0.069$ and $m = 0.69$; for row 3: $C = 0.041$ and $m = 0.74$; and for row 4: $C = 0.029$ and $m = 0.78$. All properties are determined at the inlet temperature except for μ_h, which is evaluated at the heater temperature. Note that when heat sinks are employed, forced convection immersion cooling can support a heat flux of approximately 50 W/cm^2.

Test data have been obtained for a vertical column of ten 6.4 mm × 6.4 mm heaters, with subcooled R-113 flowing in the upward direction. In general, CHF occurred first at the last downstream element in the array, and that CHF values of 100 W/cm^2 were easily attained. The CHF data yielded the following equation:

$$q_c'' = C_5 \mathrm{We}^n V \rho_f h_{fg} \left(\frac{\rho_f}{\rho_g}\right)^{15/23} \left(\frac{L}{D_h}\right)^{1/23} \left[1 + \frac{c_{pf}\Delta T_{\mathrm{sub}}}{h_{fg}}\right]^{7/23} \left[1 + \frac{0.021 \rho_f c_{pf} \Delta T_{\mathrm{sub}}}{\rho_g h_{fg}}\right] \tag{4.8.56}$$

where the Weber number, $\mathrm{We} = \rho_f V^2 L/\sigma$. Note that for We > 100, $C_5 = 0.125$ and $n = -8/23$, and for We < 10, $C_5 = 0.254$ and $n = -1/2$.

Immersion Cooling Using Jets

Two modes have been studied. In the first, a dielectric liquid jet discharges into a miscible liquid and is said to be submerged; in the second, a liquid jet discharges into an immiscible gas (air) and is said to be free. In general, the average Nusselt number can be expressed as

$$\overline{\mathrm{Nu}} = f\left(\mathrm{Re}^m, \mathrm{Pr}^n, L/d, z/d\right) \tag{4.8.57}$$

where L/d is the ratio of the chip length to orifice diameter, and z/d is the ratio of the jet to heat source distance to the orifice diameter. A free jet is virtually unaffected by the orifice-to-chip distance, and as a consequence the (z/d) term drops out.

Data for single-phase forced convection cooling with free jets are available for 2×2 and 3×3 heat source arrays. The heat sources were 12.7 $\times$ 12.7 mm and the cooling fluid was FC-77. Each heat source was cooled either by a single jet or by a 2×2 or 3×3 array of jets per source. For all the configurations tested, the average Nusselt number was correlated by a single expression:

$$\overline{\mathrm{Nu}_L} = 3.84\left(0.008\frac{L}{d}n + 1\right)\mathrm{Re}^{1/2}\mathrm{Pr}^{1/3} \tag{4.8.58}$$

where fluid properties are to be evaluated at an average of the heat source and jet inlet temperatures.

Data for single-phase forced convection using submerged jets are available for a 5×5 mm vertical heat source cooled by a 1.0-mm-diameter submerged jet of R-113. The Nusselt number at the stagnation point was correlated by

$$\mathrm{Nu}_d = 1.29\mathrm{Re}_d^{1/2}\mathrm{Pr}^{0.4} \tag{4.8.59}$$

Also note that the performance of a submerged liquid jet should be approximately equivalent to gas jet impingement.

Data for two-phase forced convection using free jets have been collected for a single 12.7×12.7 mm heat source cooled by either a single jet or a 2×2 or 3×3 array of jets. The jet diameter, velocity, and jet-to-source distance were varied. The CHF data was correlated by

$$q_c'' = 0.0742\mathrm{We}^{-0.365}V\rho_f h_{fg}\left(\frac{\rho_g}{\rho_f}\right)^{0.239}\left[1 + 0.952\left(\frac{\rho_f}{\rho_g}\right)^{0.118}\left(\frac{c_{pf}\Delta T_{\mathrm{sub}}}{h_{fg}}\right)\right]^{1.414} \tag{4.8.60}$$

Experimental evidence in two-phase forced convection using submerged jets indicates that (1) the temperature overshoot at incipient boiling was very small compared with pool or forced boiling; (2) the boiling curves at various velocities merge to a single curve and that this curve coincides approximately with an upward extrapolation of the pool boiling curve; (3) the CHF varies as the cube of the jet velocity; (4) the CHF is greatly improved by increasing the subcooling; and (5) powers in excess of 20 W ($5 \times$ 5-mm chip) could be accommodated within a 85°C chip temperature.

Defining Terms

External thermal resistance: The thermal resistance from a convenient reference point on the outside of the electronic package to the local ambient.

Internal thermal resistance: The thermal resistance from the device junction inside an electronic package to a convenient reference point on the outside surface of the package.

Immersion cooling: Concerns applications where the coolant is in direct physical contact with the electronic components.

References

Antonetti, V.W. 1993. Cooling electronic equipment — section 517, *Heat Transfer and Fluid Flow Data Books*, Kreith, F., Ed., Genium Publishing, Schenectady, NY.

Antonetti, V.W. and Simons, R.E. 1985. Bibliography of heat transfer in electronic equipment, *IEEE Trans. Components, Hybrids, Manuf. Tech.*, CHMT-8(2), 289–295.

Park, K.A. and Bergles, A.E. 1988. Effects of size of simulated microelectron chips on boiling and critical heat flux, *J. Heat Transfer*, 110, 728–734.

Simons, R.E. 1988. Bibliography of heat transfer in electronic equipment, in *Advances in Thermal Modeling of Electronic Components and Systems*, Vol. 1, Bar-Cohen, A. and Kraus, A.D., Eds., Hemisphere Publishing, New York, 413–441.

Simons, R.E. 1990. Bibliography of heat transfer in electronic equipment, in *Advances in Thermal Modeling of Electronic Components and Systems*, Vol. 2, Bar-Cohen, A. and Kraus, A.D., Eds., ASME Press, New York, 343–412.

Sparrow, E.M., Niethammer, J.E., and Chaboki, A. 1982. Heat transfer and pressure-drop characteristics of arrays of rectangular modules in electronic equipment, *Int. J. Heat Mass Transfer*, 25, 961–973.

Sparrow, E.M., Yanezmoreno, A.A., and Otis, D.R. 1984. Convective heat transfer response to height differences in an array of block-like electronic components, *Int. J. Heat Mass Transfer*, 27, 469–473.

4.9 Non-Newtonian Fluids — Heat Transfer

Thomas F. Irvine, Jr., and Massimo Capobianchi

Introduction

The general characteristics of non-Newtonian fluids are described in Section 3.9 and will not be repeated here. Topics to be included in this section are laminar and turbulent heat transfer in fully developed duct flow, and laminar free convection heat transfer in vertical channels and plates and several other common geometries.

For non-Newtonian flows, except for certain classes of fluids which exhibit a slip phenomenon at solid boundaries, the boundary condition is taken as no-slip or zero velocity at all solid surfaces. For heat transfer analyses, however, the situation is more complicated because there are many different ways to heat a wall, which in turn affects the type of thermal boundary conditions.

In general, the rate of heat transfer from a surface, or the temperature difference between the wall and the fluid, is calculated using the equation $q_c = h_c A_q \Delta T$. Since the heat transfer coefficient can vary considerably for different thermal boundary conditions, it is important that the boundary conditions be specified correctly. Although the number of thermal boundary conditions is in principle infinite, several classical types have been identified and are in common use. They are usually identified in terms of the Nusselt number, $\mathrm{Nu} = h_c L/k$, with a particular subscript. For example, for duct flow, the symbol Nu_T is used to specify the Nusselt number when the wall temperature is constant in both the flow and peripheral directions. Other thermal boundary conditions are described in Table 4.9.1 for duct heat transfer and will be used throughout this section.

TABLE 4.9.1 Thermal Boundary Conditions for Duct Heat Transfer

1.	Constant wall temperature in both the flow and circumferential direction	Nu_T
2.	Constant heat flux in the flow direction and constant temperature in the circumferential direction	Nu_{H1}
3.	Constant heat flux in the flow and circumferential directions	Nu_{H2}
4.	Constant heat flux per unit volume in the wall with circumferential wall heat conduction	Nu_{H4}

It should be noted that because of the symmetry in circular and parallel plate ducts, Nu_{H1} and Nu_{H2} are identical and are referred to simply as Nu_H. Nu_{H4} with wall conduction is a more-complicated problem where the energy equations must be solved simultaneously in both the wall and the fluid. Such problems are called conjugated. In the Nu_{H4} situation, the designer has the flexibility of affecting the heat transfer by varying either or both the characteristics of the duct wall or the convective fluid. In the heat transfer relations to be considered later, care will be taken to identify the proper thermal boundary conditions using the nomenclature in Table 4.9.1.

Laminar Duct Heat Transfer — Purely Viscous, Time-Independent Non-Newtonian Fluids

As discussed in Section 3.9, a convenient and comprehensive constitutive equation for pseudoplastic fluids (flow index, $n < 1$) is the modified power law equation:

$$\mu_a = \frac{\mu_o}{1 + \dfrac{\mu_o}{K}\left(\dot{\gamma}\right)^{1-n}} \tag{4.9.1}$$

Equation (4.9.1) has the characteristic that at low shear rates, the equation approaches that for a Newtonian fluid while at large shear rates it describes a power law fluid. In addition, solutions using

Equation (4.9.1) generate a shear rate parameter, β, which describes whether any particular system is in the Newtonian, transitional, or power law region. For duct flow, β is given by

$$\beta = \frac{\mu_o}{K}\left(\frac{\bar{u}}{D_H}\right)^{1-n} \tag{4.9.2}$$

If $\log_{10} \beta > 2$: Power law region
If $\log_{10} \beta < -2$: Newtonian region
If $-2 \le \log_{10} \beta \le 2$: Transition region

For fully developed flow, the characteristic length is the hydraulic diameter, D_H, and the fluid temperature is the "bulk" temperature defined as

$$T_b = \frac{1}{A_c \bar{u}} \int_{A_c} uTdA_c \tag{4.9.3}$$

Figure 4.9.1 illustrates the values of Nu_T vs. β for a circular duct with the flow index, n, as a parameter. It is seen from the figure that the effect of β on Nu_T is only moderate, but for some applications it may be important to know at what value of β the system is operating. The situation is similar for boundary condition Nu_H.

Although Figure 4.9.1 shows the Nusselt number relation graphically, it is convenient to have simple correlation equations to represent the solutions for both boundary conditions. For fully developed Nusselt numbers with values of $0.5 \le n \le 1.0$ and $10^{-4} \le \beta \le 10^4$, Irvine et al. (1988) present the following equation which represents both solutions with a maximum difference of 1.5%:

$$\mathrm{Nu} = \frac{\mathrm{Nu}_N(1+\beta)}{1+\dfrac{\mathrm{Nu}_N \beta}{\mathrm{Nu}_P}} \tag{4.9.4}$$

The Newtonian Nusselt numbers are $Nu_N = 3.6568$ for Nu_T, and $Nu_N = 4.3638$ for Nu_H. In addition, Table 4.9.2 lists the power law Nusselt numbers, Nu_{TP} and Nu_{HP}, for $\log_{10} \beta = 4$.

Graetz solutions for the thermal entrance lengths are also available. They assume that the velocity profile is fully developed at the duct entrance and present the duct lengths required for the Nusselt numbers to reach within 1% of the fully developed values. Figure 4.9.2 shows these thermal entrance lengths for Nu_T thermal boundary condition. The situation is similar for boundary condition Nu_H.

A correlation equation for the thermal entrance lengths for both the Nu_T and Nu_H boundary conditions by Irvine et al. (1988) represents the numerical solutions within 0.5% for $0.5 \le n \le 1.0$ and $-4 \le \log_{10} \beta \le 4$. Table 4.9.3 lists the power law thermal entrance lengths which are needed to evaluate the following correlation equation:

$$x^+_{\mathrm{ent},\beta,n} = \frac{x^+_{\mathrm{ent},N}(1+\beta)}{1+\dfrac{x^+_{\mathrm{ent},N}(\beta)}{x^+_{\mathrm{ent},P}}} \tag{4.9.5}$$

where $x^+_{\mathrm{ent},\beta,n}$ is the modified power law dimensionless entrance length defined as $x^+_{\mathrm{ent},\beta,n} = (x_{\mathrm{ent},\beta,n}/D_H)/Pe$, and $x^+_{\mathrm{ent},N}$ and $x^+_{\mathrm{ent},P}$ are the Newtonian and power law values, respectively. The Newtonian dimensionless entrance lengths are $x^+_{\mathrm{ent},N} = 0.03347$ for Nu_T and $x^+_{\mathrm{ent},N} = 0.04309$ for Nu_H.

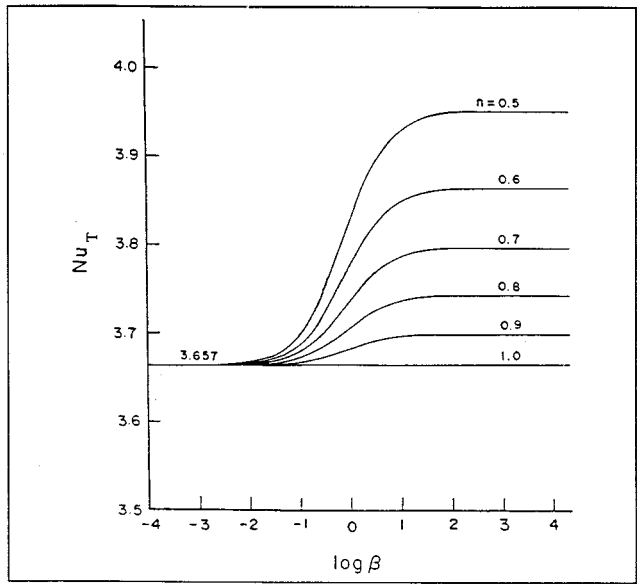


FIGURE 4.9.1 Variation of the fully developed circular duct Nusselt numbers, Nu_T, with the shear rate parameter β and *n*. (From Irvine, T.F., Jr. et al., in *ASME Symposium on Fundamentals of Forced Convection Heat Transfer*, ASME publ. HTD 101, 1988, 123–127. With permission.)

TABLE 4.9.2 Power Law Nu_T and Nu_H Solutions for a Circular Duct ($\log_{10}\beta = 4$)

n	Nu_{TP}	Nu_{HP}
1.0 (Newtonian)	3.6568	4.3638
0.9	3.6934	4.4109
0.8	3.7377	4.4679
0.7	3.7921	4.5385
0.6	3.8605	4.6281
0.5	3.9494	4.7456

Source: Irvine, T.F., Jr. et al., in *ASME Symposium on Fundamentals of Forced Convection Heat Transfer*, ASME publ. HTD 101, 1988, 123–127.

Only one noncircular geometry using the modified power law equation has been published in the archival literature for laminar fully developed heat transfer (Capobianchi and Irvine, 1992). A correlation equation for Nu_{H1} for annuli with constant heat flux at the inner wall and the outer wall insulated is

$$n<1 \qquad Nu_{H1} = \frac{1+\beta}{\dfrac{1}{Nu_{H1,N}} + \dfrac{\beta}{Nu_{H1,P}}} \tag{4.9.6}$$

Nusselt numbers for square ducts and power law fluids can be found in Chandrupatla and Sastri (1977) and, for isosceles triangular ducts, in Cheng (1984). Thermally developing and thermally developed laminar heat transfer in rectangular channels has been studied by Hartnett and Kostic (1989).

For other cross-sectional shapes, a power law approximate correlation has been proposed by Cheng (1984):

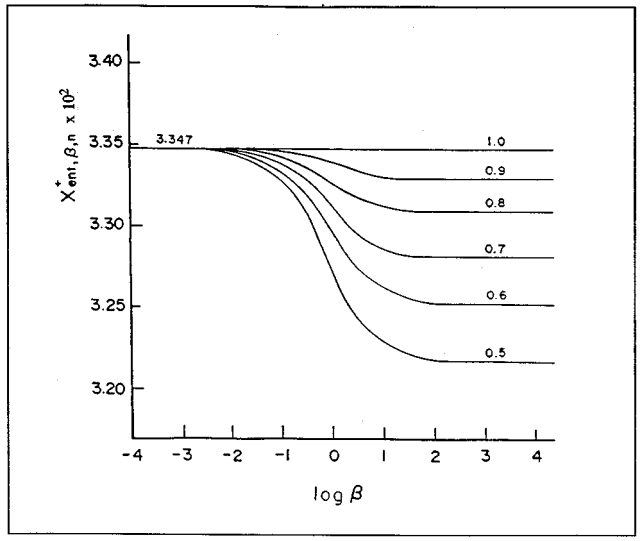


FIGURE 4.9.2 Thermal entrance lengths vs. shear rate parameter β and *n* for Nu_T in circular ducts. (From Irvine, T.F., Jr. et al., in *ASME Symposium on Fundamentals of Forced Convection Heat Transfer*, ASME publ. HTD 101, 1988, 123–127. With permission.)

TABLE 4.9.3 Values of Circular Duct Thermal Entrance Lengths for Nu_T and Nu_H for Use in Equation 4.9.5

n	Nu_T, $x^+_{ent,P} \times 10^2$	Nu_H, $x^+_{ent,P} \times 10^2$
1.0 (Newtonian)	3.347	4.309
0.9	3.326	4.281
0.8	3.306	4.248
0.7	3.279	4.210
0.6	3.250	4.166
0.5	3.213	4.114

Source: Irvine, T.F., Jr., et al., in *ASME Symposium on Fundamentals of Forced Convection Heat Transfer*, ASME publ. HTD 101, 1988, 123–127.

$$Nu_P = Nu_N \left[\frac{(a+bn)}{(a+b)n} \right]^{1/3} \tag{4.9.7}$$

where *a* and *b* are the Kozicki geometric constants listed in Table 3.9.3 in the section on non-Newtonian flows. Equation (4.9.7) applies to any thermal boundary condition. For circular ducts, Equation 4.9.7 predicts the correct solution for both Nu_T and Nu_H.

Turbulent Duct Flow for Purely Viscous Time-Independent Non-Newtonian Fluids

It is known that in turbulent flow, the type of thermal boundary conditions has much less effect than in laminar flow. Therefore, turbulent flow heat transfer investigations are often reported without specifying the thermal boundary conditions. Yoo (1974) has presented an empirical correlation for turbulent heat transfer in circular ducts for purely viscous time-independent power law fluids.

$$\mathrm{St}\,\mathrm{Pr}_a^{2/3} = 0.0152\,\mathrm{Re}_a^{-0.155} \tag{4.9.8}$$

Equation (4.9.8) describes all of the experimental data available in the literature at the time with a mean deviation of 2.3%. Equation (4.9.8) is recommended in order to predict the turbulent fully developed heat transfer in the ranges $0.2 \le n \le 0.9$ and $3000 \le \mathrm{Re}_a \le 90{,}000$. The Reynolds number and Prandtl numbers in Equation (4.9.8) are based on the apparent viscosity at the wall, μ_a, i.e.,

$$\mathrm{Re}_a = \frac{\rho \bar{u} D_H}{\mu_a} \tag{4.9.9}$$

$$\mathrm{Pr}_a = \frac{\mu_a c_p}{k} \tag{4.9.10}$$

In order to evaluate Equations (4.9.9) and (4.9.10) in terms of the rheological properties and operating parameters, an expression must be obtained for μ_a in terms of these quantities. The value of μ_a is evaluated by considering that μ_a is determined from fully developed laminar circular tube power law fluid flow for which it can be shown that (Irvine and Karni, 1987)

$$\mu_a = K\left(\frac{3n+1}{4n}\right)^{n-1}\left(\frac{8\bar{u}}{D_H}\right)^{n-1} \tag{4.9.11}$$

assuming that the quantities K, n, c_p, and k are constant. It is also of interest that the Prandtl number is no longer a thermophysical property for power law fluids but depends upon the average velocity, $\bar{u}$, and the hydraulic diameter, D_H.

Hartnett and Rao (1987) have investigated fully developed turbulent heat transfer for a rectangular duct with a 2:1 aspect ratio and propose the following equation which generally agreed with their experimental data within ±20%:

$$\mathrm{Nu} = (0.0081 + 0.0149n)\,\mathrm{Re}_a^{0.8}\,\mathrm{Pr}_a^{0.4} \tag{4.9.12}$$

Viscoelastic Fluids

An important characteristic of viscoelastic fluids is their large hydrodynamic and thermal entrance lengths. Cho and Hartnett (1982) have reported hydrodynamic entrance lengths of up to 100 diameters and thermal entrance lengths up to 200 to 800 diameters depending upon the Reynolds and Prandtl numbers. These can be compared with Newtonian fluids entrance lengths which are of the order of 10 to 15 diameters. Therefore, care must be used in applying fully developed relations to practical situations.

Cho et al. (1980) reported heat transfer measurements in the thermal entrance region and recommend the following empirical equation for saturated aqueous polymer solutions for $6000 \le \mathrm{Re}_a$ and x/D_H values up to 450:

$$J_H = 0.13\left(x/D_H\right)^{-0.24}\,\mathrm{Re}_a^{-0.45} \tag{4.9.13}$$

where $J_H = \mathrm{St}\,\mathrm{Pr}_a^{2/3}$ and $\mathrm{St} - h_c/\rho c_p \bar{u}$.

All of the reported fully developed turbulent flow heat transfer measurements have been plagued by solute and solvent, thermal entrance, and degradation effects, and thus there is considerable scatter in the results. Degradation effects can be reduced or eliminated by using large amounts of polymer (500

to 10,000 wppm) so that the solution becomes saturated. Cho and Hartnett (1982) attempted to eliminate these effects by using a thermal entrance length of 430 diameters and saturated polymer solutions which should yield maximum heat transfer reductions. Their experimental results for fully developed heat transfer were correlated for a Reynolds number range $3500 \le \mathrm{Re}_a \le 40{,}000$ and concentration solutions of 500 to 5000 wppm of polyacrylamide and polyethylene oxide by

$$J_H = 0.03\mathrm{Re}_a^{-0.45} \tag{4.9.14}$$

For viscoelastic fluids in fully developed (hydrodynamically and thermally) *laminar flow in circular ducts* there is no apparent viscoelastic effect. Thus, the heat transfer relations are the same as those for time-independent fluids such as power law or modified power law fluids. The same situation holds for thermal entrance region heat transfer (Graetz problem). Relations for laminar Nusselt numbers in thermal entrance regions are presented by Cho and Hartnett (1982).

Free Convection Flows and Heat Transfer

Free convection information available in the heat transfer literature up to the present time is concentrated on heat transfer to power law fluids for vertical plates and parallel plate channels. For free convection flows, however, the velocities and thus the shear rates are low and care must be taken that the flow for a particular fluid is in the power law shear rate region before using power law solutions or correlations. Comprehensive review articles on free convection with non-Newtonian fluids have been presented by Shenoy and Mashelkar (1982) and Irvine and Karni (1987).

For a single vertical plate with a modified power law fluid and a thermal boundary condition $\overline{N}u_T$, in laminar flow, the following relation is recommended by Shenoy and Mashelkar (1982):

$$\overline{N}u_{TL} = T(n)Gr_{TL}^{1/(2n+2)}\mathrm{Pr}_{TL}^{n/(3n+1)} \tag{4.9.15}$$

where $\overline{N}u_{TL}$ is the average Nusselt number and

$$\mathrm{Gr}_{TL} = \frac{\rho^2 L^{n+2}}{K^2}\left[g\alpha\left(T_s - T_\infty\right)\right]^{2-n} \tag{4.9.16}$$

$$\mathrm{Pr}_{TL} = \frac{\rho c_p}{k}\left(\frac{K}{\rho}\right)^{2/(n+1)} L^{(n-1)/(2n+2)}\left[g\alpha\left(T_s - T_\infty\right)\right]^{(3n-3)/(2n+2)} \tag{4.9.17}$$

where α is the isobaric thermal expansion coefficient.

In the range $0.5 \le n \le 1$, $T(n)$ can be approximated by

$$T(n) = 0.1636n + 0.5139 \tag{4.9.18}$$

The characteristic dimension in the Nusselt and Grashof numbers is the plate height, L.

For thermal boundary conditions Nu_H, the following relation is also recommended by Shenoy and Mashelkar (1982). Since the heat flux, q_w is specified in this case, the local plate temperature at any x (measured from the bottom of the plate) can be obtained from the local Nusselt number Nu_{Hx}. The heat transfer coefficient is defined in terms of the difference between the wall and free-stream temperatures.

$$\mathrm{Nu}_{Hx} = 0.619\left[\mathrm{Gr}_{Hx}^{(3n+2)/(n+4)}\mathrm{Pr}_{Hx}^{n}\right]^{0.213} \tag{4.9.19}$$

where

$$\mathrm{Gr}_{Hx} = \frac{\rho^2 x^4}{k^2}\left(\frac{g\alpha q_w}{k}\right)^{2-n} \tag{4.9.20}$$

$$\mathrm{Pr}_{Hx} = \frac{\rho c_p}{K}\left(\frac{K}{\rho}\right)^{5/(n+4)} x^{(2n-2)/(n+4)}\left(\frac{g\alpha q_w}{k}\right)^{(3n-3)/(n+4)} \tag{4.9.21}$$

Vertical Parallel Plates

For *power law fluids* and laminar flow, Figure 4.9.3 presents the graphical results of a numerical solution. Of interest are the average Nusselt number $\bar{N}u_{Tb}$ and the dimensionless average flow velocity between the plates, U_o^+. These are shown on the left and right ordinates respectively in Figure 4.9.3 (Irvine et al., 1982). The characteristic dimension in the Nusselt and Grashof numbers is the plate spacing, *b*. The dimensionless quantities used in Figure 4.9.3 are defined as follows:

$$\bar{N}u_{Tb} = \frac{\bar{h}_c b}{k} \qquad U_o^+ = \frac{bu_o}{Lu^*}$$

$$\mathrm{Pr}_g = \frac{\rho c_p}{k}\left[\frac{v_k^{1/(2-n)}}{\left(\frac{L}{b}\right)^{(1-n)/(2-n)} b^{(2n-2)/(2-n)}}\right] \qquad v_K = \frac{K}{\rho}$$

$$\mathrm{Gr}_g = \frac{g\alpha\left(T_s - T_\infty\right)b^{(n+2)/(2-n)}}{v_K^{2/(2-n)}\left(\frac{L}{b}\right)^{n/(2-n)}} \qquad u^* = \frac{v_K^{1/(2-n)}\, b^{(1-2n)/(2-n)}}{L^{(1-n)/(2-n)}}$$

For vertical parallel plates for the average Nusselt number, $\bar{N}u_{Hb}$, and the between plate average velocity, Schneider and Irvine (1984) have presented graphical results similar to Figure 4.9.3.

Lee (1992) has presented a numerical solution for laminar flow of a *modified power law fluid* between vertical plates. Lee has also calculated thermal entrance regions and shown that if a parallel plate system is actually operating in the transition region and if the power law solution is used, both the total heat transfer and the velocity between plates can differ by over an order of magnitude. It is important to consider the shear rate parameter in order to determine which free convection solution to use.

Sphere and Horizontal Cylinder — Power Law Fluids

For flow over a sphere, the correlation for power law fluids by Amato and Tien (1976) is

$$\bar{N}u_{Tr} = CZ^D \tag{4.9.22}$$

where

$$Z = \mathrm{Gr}_{Tr}^{1/(2n+2)}\,\mathrm{Pr}_{Tr}^{n/(3n+1)} \tag{4.9.23}$$

and

$$C = 0.996 \pm 0.120, \quad D = 0.682 \qquad \text{for } Z < 10$$

$$C = 0.489 \pm 0.005, \quad D = 1.10 \qquad \text{for } 10 \le Z \le 40$$

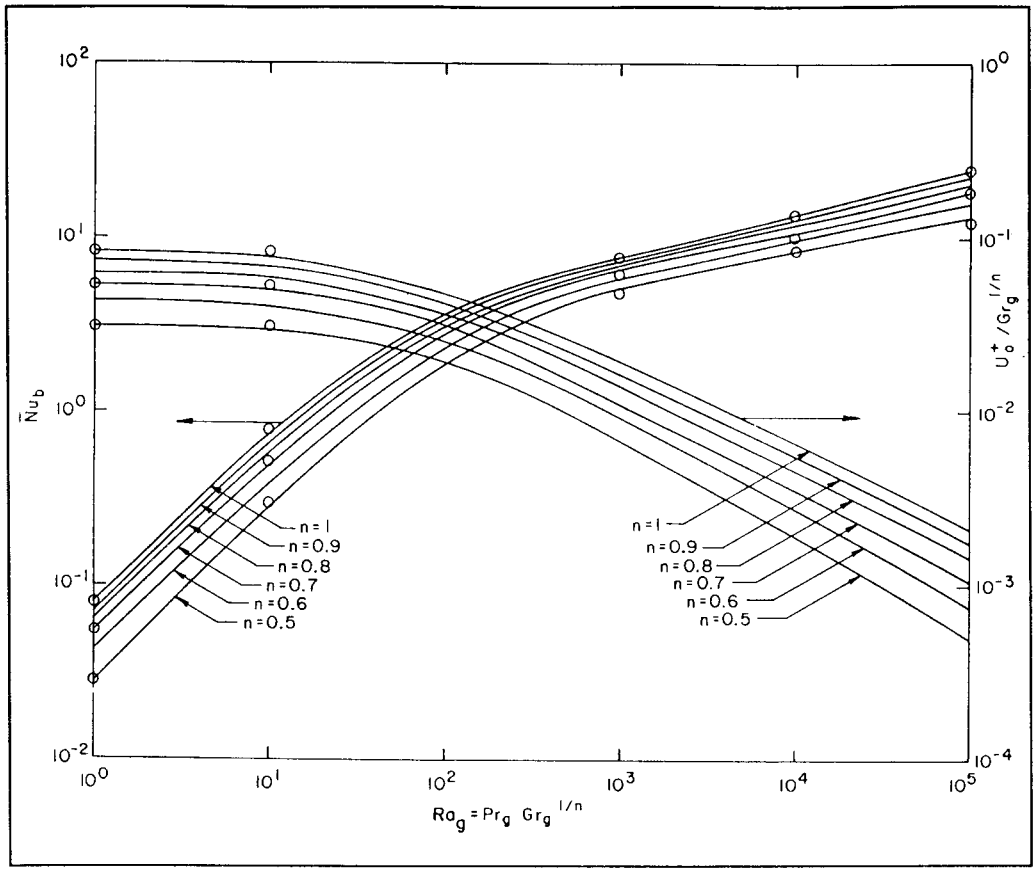


FIGURE 4.9.3 Free convection average Nusselt number, $\bar{N}u_b$ and dimensionless average velocity U_o^+ between vertical plates for a power law fluid vs. generalized Raleigh number for the Nu_T boundary condition. (From Irvine, T.F., Jr. et al., ASME Paper 82-WA/HT-69, 1982. With permission.)

where the characteristic dimension in all dimensionless variables is the sphere radius, *r*, and Gr_{Tr} and Pr_{Tr} are defined in Equations (4.9.16) and (4.9.17).

For pseudoplastic fluids flowing over a cylinder, an experimental correlation proposed by Gentry and Worllersheim (1974) for the average Nusselt number, $\bar{N}u_{TD}$, is

$$\bar{N}u_{TD} = \frac{\bar{h}_c D}{k} = 1.19\left(\mathrm{Gr}_{TD}\,\mathrm{Pr}_{TD}\right)^{0.2} \tag{4.9.24}$$

where Gr_{TD} and Pr_{TD} are defined as in Equations (4.9.16) and (4.9.17) with the cylinder diameter, *D*, being used instead of *L*.

References

Acrivos, A. 1960. A theoretical analysis of laminar natural convection heat transfer to non-Newtonian fluids, *AIChE J.*, 6, 584–590.

Amato, W.S. and Tien, C. 1976. Free convection heat transfer from isothermal spheres in polymer solutions, *Int. J. Heat Mass Transfer*, 19, 1257–1266.

Capobianchi, M. and Irvine, T.F., Jr. 1992. Predictions of pressure drop and heat transfer in concentric annular ducts with modified power law fluids, *Wärme Stoffübertragung*, 27,209–215.

Chandrupatla, A.R. and Sastri, V.M. 1977. Laminar forced convection heat transfer of a non-Newtonian fluid in a square duct, *Int. J. Heat Mass Transfer*, 20, 1315–1324.

Cheng, J.A. 1984. Laminar Forced Convection Heat Transfer of Power Law Fluids in Isosceles Triangular Ducts, Ph.D. Thesis, Mechanical Engineering Department, State University of New York at Stony Brook.

Cho, Y.I. and Hartnett, J.P. 1982. Non-Newtonian fluids in circular pipe flow, *Adv. Heat Transfer*, 15, 59–141.

Cho, Y.I., Ng, K.S., and Hartnett, J.P. 1980. Viscoelastic fluids in turbulent pipe flow — a new heat transfer correlation, *Lett. Heat Mass Transfer*, 7, 347.

Gentry, C.C. and Wollersheim, D.E. 1974. Local free convection to non-Newtonian fluids from a horizontal isothermal cylinder, *ASME J. Heat Transfer*, 96, 3–8.

Hartnett, J.P. and Kostic, M. 1989. Heat transfer to Newtonian and non-Newtonian fluids in rectangular ducts, *Adv. Heat Transfer*, 19, 247–356.

Hartnett, J.P. and Rao, B.K. 1987. Heat transfer and pressure drop for purely viscous non-Newtonian fluids in turbulent flow through rectangular passages, *Wärme Stoffüberetragung*, 21, 261.

Irvine, T.F., Jr. and Karni, J. 1987. Non-Newtonian flow and heat transfer, in *Handbook of Single Phase Convective Heat Transfer*, John Wiley & Sons, New York, 20-1–20-57.

Irvine, T.F., Jr., Wu, K.C., and Schneider, W.J. 1982. Vertical Channel Free Convection to a Power Law Fluid, ASME Paper 82-WA/HT-69.

Irvine, T.F., Jr., Kim, S.C., and Gui, F.L. 1988. Graetz problem solutions for a modified power law fluid, in *ASME Symposium on Fundamentals of Forced Convection Heat Transfer*, ASME publ. HTD 101, pp. 123–127.

Lee, S.R. 1992, A Computational Analysis of Natural Convection in a Vertical Channel with a Modified Power Law Fluid, Ph.D. Thesis, Mechanical Engineering Department, State University of New York at Stony Brook.

Schneider, W.J. and Irvine, T.F., Jr. 1984. Vertical Channel Free Convection for a Power Law Fluid with Constant Heat Flux, ASME Paper 84-HT-16.

Shenoy, A.V. and Mashelkar, R.A. 1982 Thermal convection in non-Newtonian fluids, *Adv. Heat Transfer*, 15, 143–225.

Yoo, S.S. 1974. Heat Transfer and Friction Factors for Non-Newtonian Fluids in Turbulent Pipe Flow, Ph.D. Thesis, University of Illinois at Chicago Circle.

Further Information

Other sources which may be consulted for more detailed information are Cho and Hartnett (1982), Shenoy and Mashelkar (1982), Irvine and Karni (1987), and Hartnett and Kostic (1989).

Rizzoni, G. “Electrical Engineering*”
Mechanical Engineering Handbook
Ed. Frank Kreith
Boca Raton: CRC Press LLC, 1999

Electrical Engineering*

Giorgio Rizzoni
Ohio State University

*From Rizzoni, G., *Principles and Applications of Electrical Engineering, 2nd ed.,* 1996; reproduced with permission of MacGraw-Hill, Inc.

5.1 Introduction

The role played by electrical and electronic engineering in mechanical systems has dramatically increased in importance in the past two decades, thanks to advances in integrated circuit electronics and in materials that have permitted the integration of sensing, computing, and actuation technology into industrial systems and consumer products. Examples of this integration revolution, which has been referred to as a new field called *Mechatronics,* can be found in consumer electronics (auto-focus cameras, printers, microprocessor-controlled appliances), in industrial automation, and in transportation systems, most notably in passenger vehicles. The aim of this chapter is to review and summarize the foundations of electrical engineering for the purpose of providing the practicing mechanical engineer a quick and useful reference to the different fields of electrical engineering. Special emphasis has been placed on those topics that are likely to be relevant to product design.

5.2 Fundamentals of Electric Circuits

This section presents the fundamental laws of circuit analysis and serves as the foundation for the study of electrical circuits. The fundamental concepts developed in these first pages will be called on through the chapter.

The fundamental electric quantity is **charge**, and the smallest amount of charge that exists is the charge carried by an electron, equal to

$$q_e = -1.602 \times 10^{-19} \text{ coulomb} \tag{5.2.1}$$

As you can see, the amount of charge associated with an electron is rather small. This, of course, has to do with the size of the unit we use to measure charge, the **Coulomb** (C), named after Charles Coulomb. However, the definition of the coulomb leads to an appropriate unit when we define electric current, since current consists of the flow of very large numbers of charge particles. The other charge-carrying particle in an atom, the proton, is assigned a positive sign and the same magnitude. The charge of a proton is

$$q_p = +1.602 \times 10^{-19} \text{ coulomb} \tag{5.2.2}$$

Electrons and protons are often referred to as **elementary charges**.

Electric current is defined as the time rate of change of charge passing through a predetermined area. If we consider the effect of the enormous number of elementary charges actually flowing, we can write this relationship in differential form:

$$i = \frac{dq}{dt} \frac{\text{C}}{\text{s}} \tag{5.2.3}$$

The units of current are called **amperes** (A), where 1 A = 1 C/sec. The electrical engineering convention states that the positive direction of current flow is that of positive charges. In metallic conductors,

however, current is carried by negative charges; these charges are the free electrons in the conduction band, which are only weakly attracted to the atomic structure in metallic elements and are therefore easily displaced in the presence of electric fields.

In order for current to flow there must exist a closed circuit, Figure 5.2.1 depicts a simple circuit, composed of a battery (e.g., a dry-cell or alkaline 1.5-V battery) and a light bulb.

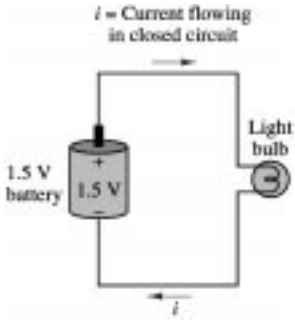


FIGURE 5.2.1 A simple electrical circuit.

Note that in the circuit of Figure 5.2.1, the current, i, flowing from the battery to the resistor is equal to the current flowing from the light bulb to the battery. In other words, no current (and therefore no charge) is "lost" around the closed circuit. This principle was observed by the German scientist G.R. Kirchhoff and is now known as **Kirchhoff's current law** (KCL). KCL states that because charge cannot be created but must be conserved, *the sum of the currents at a node must equal zero* (in an electrical circuit, a **node** is the junction of two or more conductors). Formally:

$$\sum_{n=1}^{N} i_n = 0 \qquad \text{Kirchhoff's current law} \tag{5.2.4}$$

The significance of KCL is illustrated in Figure 5.2.2, where the simple circuit of Figure 5 has been augmented by the addition of two light bulbs (note how the two nodes that exist in this circuit have been emphasized by the shaded areas). In applying KCL, one usually defines currents entering a node as being negative and currents exiting the node as being positive. Thus, the resulting expression for the circuit of Figure 5.2.2 is

$$-i + i_1 + i_2 + i_3 = 0$$

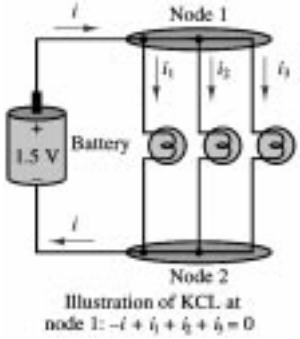


FIGURE 5.2.2 Illustration of Kirchhoff's current law.

Charge moving in an electric circuit gives rise to a current, as stated in the preceding section. Naturally, it must take some work, or energy, for the charge to move between two points in a circuit, say, from

point *a* to point *b*. The total *work per unit charge* associated with the motion of charge between two points is called **voltage**. Thus, the units of voltage are those of energy per unit charge:

$$1 \text{ volt} = \frac{1 \text{ joule}}{\text{coulomb}} \tag{5.2.5}$$

The voltage, or **potential difference,** between two points in a circuit indicates the energy required to move charge from one point to the other. As will be presently shown, the direction, or polarity, of the voltage is closely tied to whether energy is being dissipated or generated in the process. The seemingly abstract concept of work being done in moving charges can be directly applied to the analysis of electrical circuits; consider again the simple circuit consisting of a battery and a light bulb. The circuit is drawn again for convenience in Figure 5.2.3, and nodes are defined by the letters *a* and *b*. A series of carefully conducted experimental observations regarding the nature of voltages in an electric circuit led Kirchhoff to the formulation of the second of his laws, **Kirchhoff's voltage law,** or KVL. The principle underlying KVL is that no energy is lost or created in an electric circuit; in circuit terms, the sum of all voltages associated with sources must equal the sum of the load voltages, so that *the net voltage around a closed circuit is zero*. If this were not the case, we would need to find a physical explanation for the excess (or missing) energy not accounted for in the voltages around a circuit. KVL may be stated in a form similar to that used for KCL:

$$\sum_{n=1}^{N} v_n = 0 \quad \text{Kirchhoff's voltage law} \tag{5.2.6}$$

where the v_n are the individual voltages around the closed circuit. Making reference to Figure 5.2.3, we can see that it must follow from KVL that the work generated by the battery is equal to the energy dissipated in the light bulb to sustain the current flow and to convert the electric energy to heat and light:

$$v_{ab} = -v_{ba}$$

or

$$v_1 = v_2$$

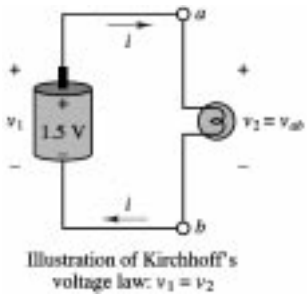


FIGURE 5.2.3 Voltages around a circuit.

One may think of the work done in moving a charge from point *a* to point *b* and the work done moving it back from *b* to *a* as corresponding directly to the *voltages across individual circuit elements*. Let *Q* be the total charge that moves around the circuit per unit time, giving rise to the current *i*. Then the work done in moving *Q* from *b* to *a* (i.e., across the battery) is

$$W_{ba} = Q \times 1.5 \text{ V} \tag{5.2.7}$$

Similarly, work is done in moving Q from a to b, that is, across the light bulb. Note that the word *potential* is quite appropriate as a synonym of voltage, in that voltage represents the potential energy between two points in a circuit: if we remove the light bulb from its connections to the battery, there still exists a voltage across the (now disconnected) terminals b and a.

A moment's reflection upon the significance of voltage should suggest that it must be necessary to specify a sign for this quantity. Consider, again, the same dry-cell or alkaline battery, where, by virtue of an electrochemically induced separation of charge, a 1.5-V potential difference is generated. The potential generated by the battery may be used to move charge in a circuit. The rate at which charge is moved once a closed circuit is established (i.e., the current drawn by the circuit connected to the battery) depends now on the circuit element we choose to connect to the battery. Thus, while the voltage across the battery represents the potential for *providing energy* to a circuit, the voltage across the light bulb indicates the amount of work done in *dissipating energy*. In the first case, energy is generated; in the second, it is consumed (note that energy may also be stored, by suitable circuit elements yet to be introduced). This fundamental distinction required attention in defining the sign (or polarity) of voltages.

We shall, in general, refer to elements that provide energy as **sources**, and to elements that dissipate energy as **loads**. Standard symbols for a generalized source-and-load circuit are shown in Figure 5.2.4. Formal definitions will be given in a later section.

A symbolic representation of the battery–light bulb circuit of Figure 2.5.

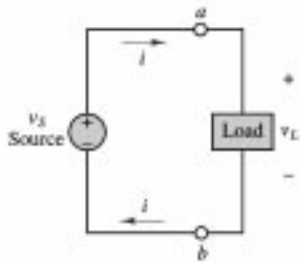


FIGURE 5.2.4 Sources and loads in an electrical circuit.

Electric Power and Sign Convention

The definition of voltage as work per unit charge lends itself very conveniently to the introduction of power. Recall that power is defined as the work done per unit time. Thus, the power, P, either generated or dissipated by a circuit element can be represented by the following relationship:

$$\text{Power} = \frac{\text{Work}}{\text{Time}} = \frac{\text{Work}}{\text{Unit charge}} \frac{\text{Charge}}{\text{Time}} = \text{Voltage} \times \text{Current} \tag{5.2.8}$$

Thus, the electrical power generated by an active element, or that dissipated or stored by a passive element, is equal to the product of the voltage across the element and the current flowing through it.

$$P = VI \tag{5.2.9}$$

It is easy to verify that the units of voltage (joules/coulomb) times current (coulombs/second) are indeed those of power (joules/second, or watts).

It is important to realize that, just like voltage, power is a signed quantity, and that it is necessary to make a distinction between *positive* and *negative power*. This distinction can be understood with reference to Figure 5.2.5, in which a source and a load are shown side by side. The polarity of the voltage across the source and the direction of the current through it indicate that the voltage source *is doing work in moving charge from a lower potential to a higher potential*. On the other hand, the load is dissipating

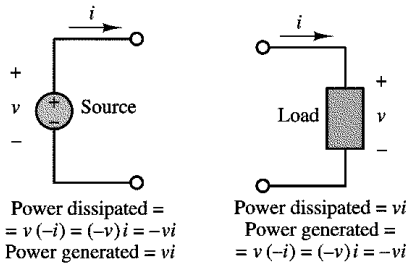


FIGURE 5.2.5 The passive sign convention.

energy, because the direction of the current indicates that *charge is being displaced from a higher potential to a lower potential*. To avoid confusion with regard to the sign of power, the electrical engineering community uniformly adopts the **passive sign convention**, which simply states that *the power dissipated by a load is a positive quantity* (or, conversely, that the power generated by a source is a positive quantity). Another way of phrasing the same concept is to state that if current flows from a higher to a lower voltage (+ to –), the power dissipated will be a positive quantity.

Circuit Elements and Their *i-v* Characteristics

The relationship between current and voltage at the terminals of a circuit element defines the behavior of that element within the circuit. In this section, we shall introduce a graphical means of representing the terminal characteristics of circuit elements. Figure 5.2.6 depicts the representation that will be employed throughout the chapter to denote a generalized circuit element: the variable i represents the current flowing through the element, while ν is the potential difference, or voltage, across the element.

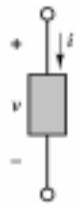


FIGURE 5.2.6 Generalized representation of circuit elements.

Suppose now that a known voltage were imposed across a circuit element. The current that would flow as a consequence of this voltage, and the voltage itself, form a unique pair of values. If the voltage applied to the element were varied and the resulting current measured, it would be possible to construct a functional relationship between voltage and current known as the ***i-v* characteristic** (or **volt-ampere characteristic**). Such a relationship defines the circuit element, in the sense that if we impose any prescribed voltage (or current), the resulting current (or voltage) is directly obtainable from the *i-v* characteristic. A direct consequence is that the power dissipated (or generated) by the element may also be determined from the *i-v* curve.

The *i-v* characteristics of ideal current and voltage sources can also be useful in visually representing their behavior. An ideal voltage source generates a prescribed voltage independent of the current drawn from the load; thus, its *i-v* characteristic is a straight vertical line with a voltage axis intercept corresponding to the source voltage. Similarly, the *i-v* characteristic of an ideal current source is a horizontal line with a current axis intercept corresponding to the source current. Figure 5.2.7 depicts this behavior.

Resistance and Ohm's Law

When electric current flows through a metal wire or through other circuit elements, it encounters a certain amount of **resistance,** the magnitude of which depends on the electrical properties of the material. Resistance to the flow of current may be undesired — for example, in the case of lead wires and

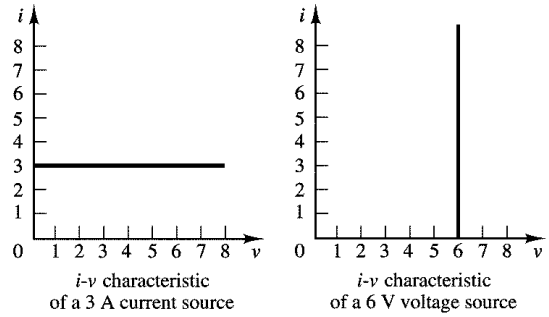


FIGURE 5.2.7 *i-v* characteristics of ideal sources.

connection cable — or it may be exploited in an electrical circuit in a useful way. Nevertheless, practically all circuit elements exhibit some resistance; as a consequence, current flowing through an element will cause energy to be dissipated in the form of heat. An ideal **resistor** is a device that exhibits linear resistance properties according to **Ohm's law,** which states that

$$V = IR \tag{5.2.10}$$

that is, that the voltage across an element is directly proportional to the current flow through it. R is the value of the resistance in units of **ohms** (Ω), where

$$1\,\Omega = 1\,\text{V/A} \tag{5.2.11}$$

The resistance of a material depends on a property called **resistivity,** denoted by the symbol ρ; the inverse of resistivity is called **conductivity** and is denoted by the symbol σ. For a cylindrical resistance element (shown in Figure 5.2.8), the resistance is proportional to the length of the sample, l, and inversely proportional to its cross-sectional area, A, and conductivity, σ.

$$v = \frac{l}{\sigma A} i \tag{5.2.12}$$

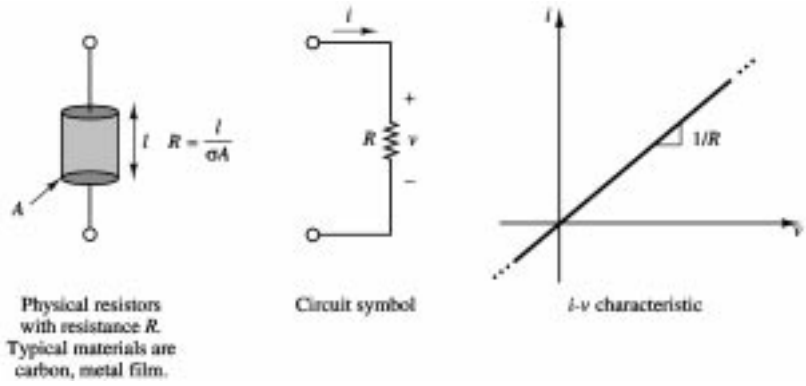


FIGURE 5.2.8 The resistance element.

It is often convenient to define the **conductance** of a circuit element as the inverse of its resistance. The symbol used to denote the conductance of an element is G, where

$$G = \frac{1}{R} \text{ siemens (S)} \quad \text{where} \quad 1\ \text{S} = 1\ \text{A/V} \tag{5.2.13}$$

Thus, Ohm's law can be rested in terms of conductance, as

$$I = GV \tag{5.2.14}$$

Ohm's law is an empirical relationship that finds widespread application in electrical engineering because of its simplicity. It is, however, only an approximation of the physics of electrically conducting materials. Typically, the linear relationship between voltage and current in electrical conductors does not apply at very high voltages and currents. Further, not all electrically conducting materials exhibit linear behavior even for small voltages and currents. It is usually true, however, that for some range of voltages and currents, most elements display a linear *i-v characteristic*.

The typical construction and the circuit symbol of the resistor are shown in Figure 5.2.8. Resistors made of cylindrical sections of carbon (with resistivity $\rho = 3.5 \times 10^{-5}$ Ω-m) are very common and are commercially available in a wide range of values for several power ratings (as will be explained shortly). Another commonly employed construction technique for resistors employs metal film. A common power rating for resistors used in electronic circuits (e.g., in most consumer electronic appliances such as radios and television sets) is $^1/_4$ W. Table 5.2.1 lists the standard values for commonly used resistors and the color code associated with these values (i.e., the common combinations of the digits $b_1b_2b_3$ as defined in Figure 5.2.9. For example, if the first three color bands on a resistor show the colors red ($b_1 = 2$), violet ($b_2 = 7$), and yellow ($b_3 = 4$), the resistance value can be interpreted as follows:

$$R = 27 \times 10^4 = 270{,}000\ \Omega = 270\ \text{k}\Omega$$

TABLE 5.2.1 Common Resistor Values ($^1/_8$-, $^1/_4$-, $^1/_2$-, 1-, 2-W Rating)

Ω	Code	Ω	Multiplier	kΩ	Multiplier	kΩ	Multiplier	kΩ	Multiplier
10	Brn-blk-blk	100	Brown	1.0	Red	10	Orange	100	Yellow
12	Brn-red-blk	120	Brown	1.2	Red	12	Orange	120	Yellow
15	Brn-grn-blk	150	Brown	1.5	Red	15	Orange	150	Yellow
18	Brn-gry-blk	180	Brown	1.8	Red	18	Orange	180	Yellow
22	Red-red-blk	220	Brown	2.2	Red	22	Orange	220	Yellow
27	Red-vlt-blk	270	Brown	2.7	Red	27	Orange	270	Yellow
33	Org-org-blk	330	Brown	3.3	Red	33	Orange	330	Yellow
39	Org-wht-blk	390	Brown	3.9	Red	39	Orange	390	Yellow
47	Ylw-vlt-blk	470	Brown	4.7	Red	47	Orange	470	Yellow
56	Grn-blu-blk	560	Brown	5.6	Red	56	Orange	560	Yellow
68	Blu-gry-blk	680	Brown	6.8	Red	68	Orange	680	Yellow
82	Gry-red-blk	820	Brown	8.2	Red	82	Orange	820	Yellow

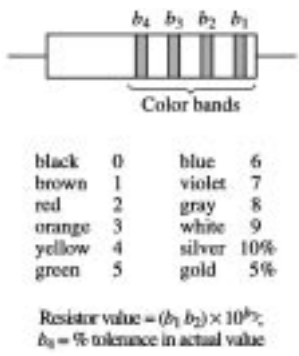


FIGURE 5.2.9 Resistor color code.

In Table 5.2.1, the leftmost column represents the complete color code; columns to the right of it only show the third color, since this is the only one that changes. For example, a 10-Ω resistor has the code brown-black-*black*, while a 100-Ω resistor has brown-black-*brown*.

In addition to the resistance in ohms, the maximum allowable power dissipation (or **power rating**) is typically specified for commercial resistors. Exceeding this power rating leads to overheating and can cause the resistor to literally start on fire. For a resistor R, the power dissipated is given by

$$P = VI = I^2R = \frac{V^2}{R} \tag{5.2.15}$$

That is, the power dissipated by a resistor is proportional to the square of the current flowing through it, as well as the square of the voltage across it. The following example illustrates how one can make use of the power rating to determine whether a given resistor will be suitable for a certain application.

Example 5.2.1 Resistance Strain Gauges

A common application of the resistance concept to engineering measurements is the resistance **strain gauge**. Strain gauges are devices that are bonded to the surface of an object, and whose resistance varies as a function of the surface strain experienced by the object. Strain gauges may be used to perform measurements of strain, stress, force, torque, and pressure. Recall that the resistance of a cylindrical conductor of cross-sectional area A, length L, and conductivity σ is given by the expression

$$R = \frac{L}{\sigma A}$$

If the conductor is compressed or elongated as a consequence of an external force, its dimensions will change, and with them its resistance. In particular, if the conductor is stretched, its cross-sectional area will decrease and the resistance will increase. If the conductor is compressed, its resistance decreases, since the length, L, will decrease. The relationship between change in resistance and change in length is given by the gauge factor, G, defined by

$$\mathrm{G} = \frac{\Delta R/R}{\Delta L/L}$$

and since the strain ε is defined as the fractional change in length of an object by the formula

$$\varepsilon = \frac{\Delta L}{L}$$

the change in resistance due to an applied strain ε is given by the expression

$$\Delta R = R_0 G\varepsilon$$

where R_0 is the resistance of the strain gauge under no strain and is called the zero strain resistance. The value of G for resistance strain gauges made of metal foil is usually about 2.

Figure 5.2.10 depicts a typical foil strain gauge. The maximum strain that can be measured by a foil gauge is about 0.4 to 0.5%; that is , $\Delta L/L = 0.004$ to 0.005. For a 120-Ω gauge, this corresponds to a change in resistance of the order of 0.96 to 1.2 Ω. Although this change is resistance is very small, it can be detected by means of suitable circuitry. Resistance strain gauges are usually connected in a circuit called the Wheatstone bridge, which we analyze later in this section.

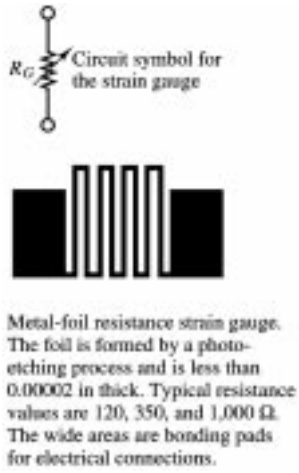


FIGURE 5.2.10 The resistance strain gauge.

Open and Short Circuits

Two convenient idealizations of the resistance element are provided by the limiting cases of Ohm's law as the resistance of a circuit element approaches zero or infinity. A circuit element with resistance approaching zero is called a **short circuit**. Intuitively, one would expect a short circuit to allow for unimpeded flow of current. In fact, metallic conductors (e.g., short wires of large diameter) approximate the behavior of a short circuit. Formally, a short circuit is defined as a circuit element across which the voltage is zero, regardless of the current flowing through it. Figure 5.2.11 depicts the circuit symbol for an ideal short circuit.

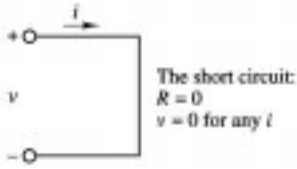


FIGURE 5.2.11 The short circuit.

Physically, any wire or other metallic conductor will exhibit some resistance, though small. For practical purposes, however, many elements approximate a short circuit quite accurately under certain conditions. For example, a large-diameter copper pipe is effectively a short circuit in the context of a residential electrical power supply, while in a low-power microelectronic circuit (e.g., an FM radio) a short length of 24 gauge wire (refer to Table 5.2.2 for the resistance of 24 gauge wire) is a more than adequate short circuit.

TABLE 5.2.2 Resistance of Copper Wire

AWG Size	Number of Strands	Diameter per Strand	Resistance per 1000 ft (Ω)
24	Solid	0.0201	28.4
24	7	0.0080	28.4
22	Solid	0.0254	18.0
22	7	0.0100	19.0
20	Solid	0.0320	11.3
20	7	0.0126	11.9
18	Solid	0.0403	7.2
18	7	0.0159	7.5
16	Solid	0.0508	4.5
16	19	0.0113	4.7

A circuit element whose resistance approaches infinity is called an **open circuit**. Intuitively, one would expect no current to flow through an open circuit, since it offers infinite resistance to any current. In an open circuit, we would expect to see zero current regardless of the externally applied voltage. Figure 5.2.12 illustrates this idea .

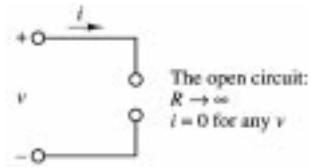


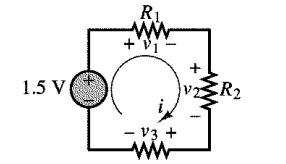
FIGURE 5.2.12 The open circuit.

In practice, it is not too difficult to approximate an open circuit; any break in continuity in a conducting path amounts to an open circuit. The idealization of the open circuit, as defined in Figure 5.2.12 does not hold, however, for very high voltages. The insulating material between two insulated terminals will break down at a sufficiently high voltage. If the insulator is air, ionized particles in the neighborhood of the two conducting elements may lead to the phenomenon of arcing; in other words, a pulse of current may be generated that momentarily jumps a gap between conductors (thanks to this principle, we are able to ignite the air-fuel mixture in a spark-ignition internal combustion engine by means of spark plugs). The ideal open and short circuits are useful concepts and find extensive use in circuit analysis.

Series Resistors and the Voltage Divider Rule

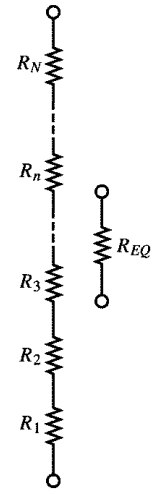
Although electrical circuits can take rather complicated forms, even the most involved circuits can be reduced to combinations of circuit elements *in parallel* and *in series*. Thus, it is important that you become acquainted with parallel and series circuits as early as possible, even before formally approaching the topic of network analysis. Parallel and series circuits have a direct relationship with Kirchhoff's laws. The objective of this section and the next is to illustrate two common circuits based on series and parallel combinations of resistors: the voltage and current dividers. These circuits form the basis of all network analysis; it is therefore important to master these topics as early as possible.

For an example of a series circuit, refer to the circuit of Figure 5.2.13, where a battery has been connected to resistors R_1, R_2, and R_3. The following definition applies.



The current i flows through each of the four series elements. Thus, by KVL,

$$1.5 = v_1 + v_2 + v_3$$



N series resistors are equivalent to a single resistor equal to the sum of the individual resistances.

FIGURE 5.2.13 Voltage divider rule.

Definition. Two or more circuit elements are said to be **in series** if the same current flows through each of the elements.

The three resistors could thus be replaced by a single resistor of value R_{EQ} without changing the amount of current required of the battery. From this result we may extrapolate to the more general relationship defining the equivalent resistance of N series resistors:

$$R_{EQ} = \sum_{n=1}^{N} R_n \tag{5.2.16}$$

which is also illustrated in Figure 5.2.13. A concept very closely tied to series resistors is that of the **voltage divider**.

The general form of the voltage divider rule for a circuit with N series resistors and a voltage source is

$$v_n = \frac{R_n}{R_1 + R_2 + \ldots + R_n + \ldots + R_N} v_S \tag{5.2.17}$$

Parallel Resistors and the Current Divider Rule

A concept analogous to that of the voltage may be developed by applying Kirchhoff's current law to a circuit containing only parallel resistances.

Definition. Two or more circuit elements are said to be **in parallel** if the same voltage appears across each of the elements. (See Figure 5.2.14.)

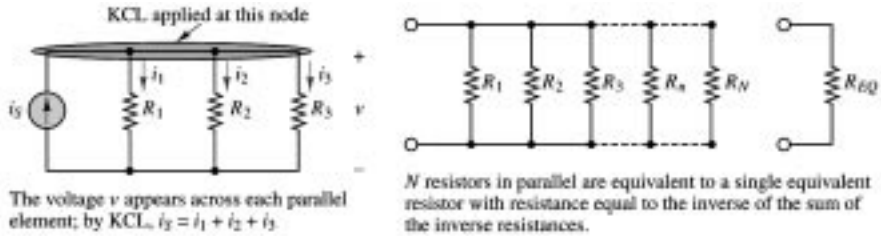


FIGURE 5.2.14 Parallel circuits.

N resistors in parallel act as a single equivalent resistance, R_{EQ}, given by the expression

$$\frac{1}{R_{EQ}} = \frac{1}{R_1} + \frac{1}{R_2} + \ldots + \frac{1}{R_N} \tag{5.2.18}$$

or

$$R_{EQ} = \frac{1}{1/R_1 + 1/R_2 + \ldots + 1/R_N} \tag{5.2.19}$$

Very often in the remainder of this book we shall refer to the parallel combination of two or more resistors with the following notation:

$$R_1 \parallel R_2 \parallel \cdots$$

where the symbol $\parallel$ signifies "in parallel with."

The general expression for the current divider for a circuit with N parallel resistors is the following:

$$i_n = \frac{1/R_n}{1/R_1 + 1/R_2 + \ldots + 1/R_n + \ldots + 1/R_N} i_S \quad \text{Current divider} \tag{5.2.20}$$

Example 5.2.2 The Wheatstone Bridge

The **Wheatstone bridge** is a resistive circuit that is frequently encountered in a variety of measurement circuits. The general form of the bridge is shown in Figure 5.2.15(a), where R_1, R_2, and R_3 are known, while R_x is an unknown resistance, to be determined. The circuit may also be redrawn as shown in Figure 5.2.15(b). The latter circuit will be used to demonstrate the use of the voltage divider rule in a mixed series-parallel circuit. The objective is to determine the unknown resistance R_x.

1. Find the value of the voltage $v_{ad} = v_{ad} - v_{bd}$ in terms of the four resistances and the source voltage, v_s. Note that since the reference point d is the same for both voltages, we can also write $v_{ab} = v_a - v_b$.
2. If $R_1 = R_2 = R_3 = 1$ kΩ, $v_s = 12$ V, and $v_{ab} = 12$ mV, what is the value of R_x?

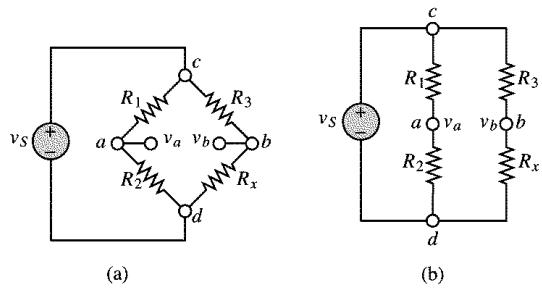


FIGURE 5.2.15 Wheatstone bridge circuits.

Solution.

1. First, we observe that the circuit consists of the parallel combination of three subcircuits: the voltage source, the series combination of R_1 and R_2, and the series combination of R_3 and R_x. Since these three subcircuits are in parallel, the same voltage will appear across each of them, namely, the source voltage, v_s.
   Thus, the source voltage divides between each resistor pair, R_1-R_2 and R_3-R_x, according to the voltage divider rule: v_a is the fraction of the source voltage appearing across R_2, while v_b is the voltage appearing across R_x:

$$v_a = v_S \frac{R_2}{R_1 + R_2} \quad \text{and} \quad v_b = v_S \frac{R_x}{R_3 + R_x}$$

   Finally, the voltage difference between points a and b is given by:

$$v_{ab} = v_a - v_b = v_S \left(\frac{R_2}{R_1 + R_2} - \frac{R_x}{R_3 + R_x} \right)$$

   This result is very useful and quite general, and it finds application in numerous practical circuits.
2. In order to solve for the unknown resistance, we substitute the numerical values in the preceding equation to obtain

$$0.012 = 12\left(\frac{1{,}000}{2{,}000} - \frac{R_x}{1{,}000 + R_x}\right)$$

which may be solved for R_x to yield

$$R_x = 996\ \Omega$$

Practical Voltage and Current Sources

Idealized models of voltage and current sources fail to take into consideration the finite-energy nature of practical voltage and current sources. The objective of this section is to extend the ideal models to models that are capable of describing the physical limitations of the voltage and current sources used in practice. Consider, for example, the model of an ideal voltage source. As the load resistance (R) decreases, the source is required to provide increasing amounts of current to maintain the voltage $v_s(t)$ across its terminal:

$$i(t) = \frac{v_S(t)}{R} \tag{5.2.21}$$

This circuit suggests that the ideal voltage source is required to provide an infinite amount of current to the load, in the limit as the load resistance approaches zero.

Figure 5.2.16 depicts a model for a practical voltage source; this is composed of an ideal voltage source, v_s, in series with a resistance, r_s. The resistance r_s in effect poses a limit to the maximum current the voltage source can provide:

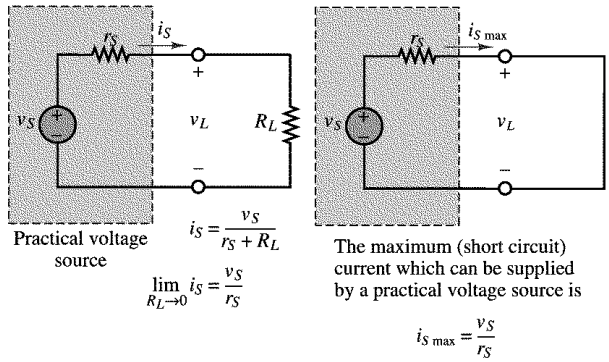
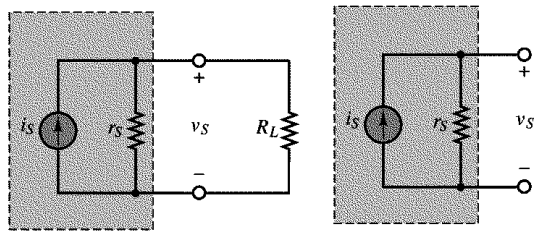


FIGURE 5.2.16 Practical voltage source.

$$i_{S\,\max} = \frac{v_S}{r_S} \tag{5.2.22}$$

It should be apparent that a desirable feature of an ideal voltage source is a very small internal resistance, so that the current requirements of an arbitrary load may be satisfied.

A similar modification of the ideal current source model is useful to describe the behavior of a practical current source. The circuit illustrated in Figure 5.2.17 depicts a simple representation of a practical current source, consisting of an ideal source in parallel with a resistor. Note that as the load resistance approaches infinity (i.e., an open circuit), the output voltage of the current source approaches its limit,



A model for practical current sources consists of an ideal source in *parallel* with an internal resistance.

Maximum output voltage for practical current source with open-circuit load:

$$v_{S\,\max} = i_S r_S$$

FIGURE 5.2.17 Practical current source.

$$v_{S\,\max} = i_S r_S \tag{5.2.23}$$

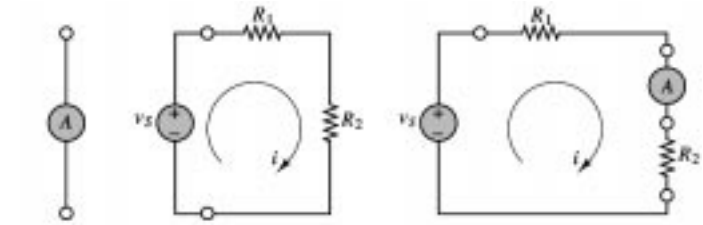
A good current source should be able to approximate the behavior of an ideal current source. Therefore, a desirable characteristic for the internal resistance of a current source is that it be as large as possible.

Measuring Devices

The Ammeter

The **ammeter** is a device that, when connected in series with a circuit element, can measure the current flowing through the element. Figure 5.2.18 illustrates this idea. From Figure 5.2.18, two requirements are evident for obtaining a correct measurement of current:

1. The ammeter must be placed in series with the element whose current is to be measured (e.g., resistor R_2).
2. The ammeter should not resist the flow of current (i.e., cause a voltage drop), or else it will not be measuring the true current flowing the circuit. *An ideal ammeter has zero internal resistance*.



Symbol for ideal ammeter

A series circuit

Circuit for the measurement of the current i

FIGURE 5.2.18 Measurement of current.

The Voltmeter

The **voltmeter** is a device that can measure the voltage across a circuit element. Since voltage is the difference in potential between two points in a circuit, the voltmeter needs to be connected across the element whose voltage we wish to measure. A voltmeter must also fulfill two requirements:

1. The voltmeter must be placed in parallel with the element whose voltage it is measuring.
2. The voltmeter should draw no current away from the element whose voltage it is measuring, or else it will not be measuring the true voltage across that element. Thus, *an ideal voltmeter has infinite internal resistance*.

Figure 5.2.19 illustrates these two points.

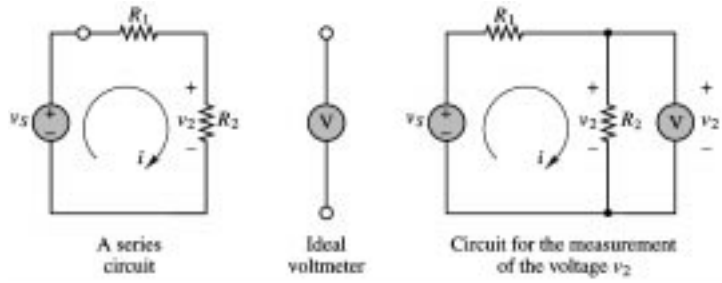


FIGURE 5.2.19 Measurement of voltage.

Once again, the definitions just stated for the ideal voltmeter and ammeter need to be augmented by considering the practical limitations of the devices. A practical ammeter will contribute some series resistance to the circuit in which it is measuring current; a practical voltmeter will not act as an ideal open circuit but will always draw some current from the measured circuit. Figure 5.2.20 depicts the circuit models for the practical ammeter and voltmeter.

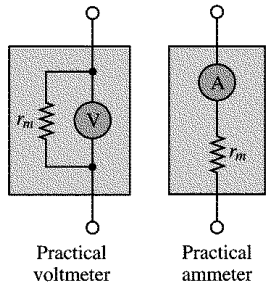


FIGURE 5.2.20 Models for practical ammeter and voltmeter.

All of the considerations that pertain to practical ammeters and voltmeters can be applied to the operation of a **wattmeter**, a measuring instrument that provides a measurement of the power dissipated by a circuit element, since the wattmeter is in effect made up of a combination of a voltmeter and an ammeter.

Figure 5.2.21 depicts the typical connection of a wattmeter in the same series circuit used in the preceding paragraphs. In effect, the wattmeter measures the current flowing through the load and, simultaneously, the voltage across it and multiplies the two to provide a reading of the power dissipated by the load.

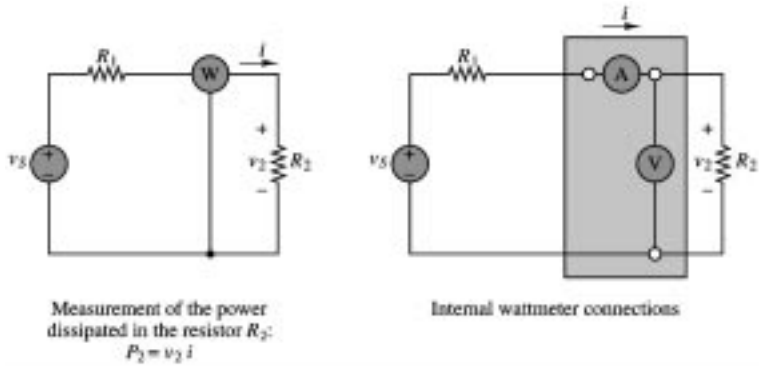


FIGURE 5.2.21 Measurement of power.

5.3 Resistive Network Analysis

This section will illustrate the fundamental techniques for the analysis of resistive circuits. The methods introduced are based on Kirchhoff's and Ohm's laws. The main thrust of the section is to introduce and illustrate various methods of circuit analysis that will be applied throughout the book.

The Node Voltage Method

Node voltage analysis is the most general method for the analysis of electrical circuits. In this section, its application to linear resistive circuits will be illustrated. The **node voltage method** is based on defining the voltage at each node as an independent variable. One of the nodes is selected as a **reference node** (usually — but not necessarily — ground), and each of the other node voltages is referenced to this node. Once each node voltage is defined, Ohm's law may be applied between any two adjacent nodes in order to determine the current flowing in each branch. In the node voltage method, *each branch current is expressed in terms of one or more node voltages;* thus, currents do not explicitly enter into the equations. Figure 5.3.1 illustrates how one defines branch currents in this method.

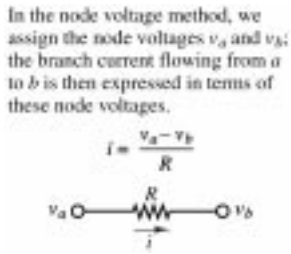


FIGURE 5.3.1 Branch current formulation in nodal analysis.

Once each branch current is defined in terms of the node voltages, Kirchhoff's current law is applied at each node. The particular form of KCL employed in the nodal analysis equates the sum of the currents into the node to the sum of the currents leaving the node:

$$\sum i_{\text{in}} = \sum i_{\text{out}} \tag{5.3.1}$$

Figure 5.3.2 illustrates this procedure.

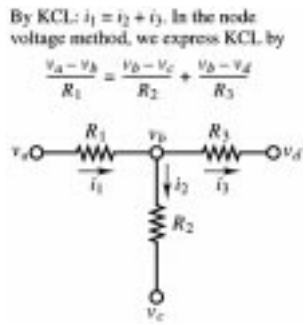


FIGURE 5.3.2 Use of KCL in nodal analysis.

The systematic application of this method to a circuit with n nodes would lead to writing n linear equations. However, one of the node voltages is the reference voltage and is therefore already known, since it is usually assumed to be zero. Thus, we can write $n - 1$ *independent linear equations* in the $n - 1$ independent variables (the node voltages). Nodal analysis provides the minimum number of equations

required to solve the circuit, since any branch voltage or current may be determined from knowledge of nodal voltages.

The nodal analysis method may also be defined as a sequence of steps, as outlined below.

Node Voltage Analysis Method

1. Select a reference node (usually ground). All other node voltages will be referenced to this node.
2. Define the remaining $n - 1$ node voltages as the independent variables.
3. Apply KCL at each of the $n - 1$ nodes, expressing each current in terms of the adjacent node voltages.
4. Solve the linear system of $n - 1$ equations in $n - 1$ unknowns.

In a circuit containing n nodes we can write at most $n - 1$ independent equations.

The Mesh Current Method

In the mesh current method, we observe that a current flowing through a resistor in a specified direction defines the polarity of the voltage across the resistor, as illustrated in Figure 5.3.3, and that the sum of the voltages around a closed circuit must equal zero, by KVL. Once a convention is established regarding the direction of current flow around a mesh. Simple application of KVL provides the desired equation. Figure 5.3.4 illustrates this point.

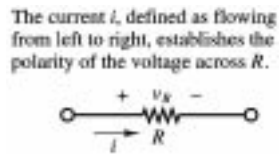


FIGURE 5.3.3 Basic principle of mesh analysis.

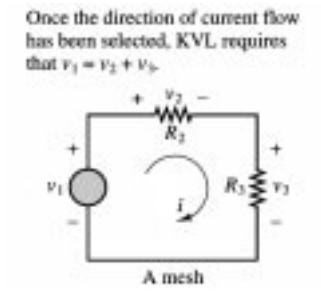


FIGURE 5.3.4 Use of KVL in mesh analysis.

The number of equations one obtains by this technique is equal to the number of meshes in the circuit. All branch currents and voltages may subsequently be obtained from the mesh currents, as will presently be shown. Since meshes are easily identified in a circuit, this method provides a very efficient and systematic procedure for the analysis of electrical circuits. The following section outlines the procedure used in applying the mesh current method to a linear circuit.

Mesh Current Analysis Method

1. Define each mesh current consistently. We shall always define mesh currents clockwise, for convenience.
2. Apply KVL around each mesh, expressing each voltage in terms of one or more mesh currents.
3. Solve the resulting linear system of equations with mesh currents as the independent variables.

In mesh analysis, it is important to be consistent in choosing the direction of current flow. To avoid confusion in writing the circuit equations, mesh currents will be defined exclusively clockwise when we are using this method.

One-Port Networks and Equivalent Circuits

This general circuit representation is shown in Figure 5.3.5. This configuration is called a **one-port network** and is particularly useful for introducing the notion of equivalent circuits. Note that the network of Figure 5.3.5 is completely described by its *i-v* characteristic.

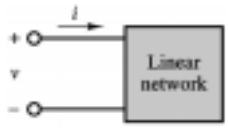


FIGURE 5.3.5 One-port network.

Thévenin and Norton Equivalent Circuits

This section discusses one of the most important topics in the analysis of electrical circuits: the concept of an **equivalent circuit**. It will be shown that it is always possible to view even a very complicated circuit in terms of much simpler *equivalent* source and load circuits, and that the transformations leading to equivalent circuits are easily managed, with a little practice. In studying node voltage and mesh current analysis, you may have observed that there is a certain correspondence (called **duality**) between current sources and voltage sources, on the one hand, and parallel and series circuits, on the other. This duality appears again very clearly in the analysis of equivalent circuits: it will shortly be shown that equivalent circuits fall into one of two classes, involving either voltage or current sources and (respectively) either series or parallel resistors, reflecting this same principle of duality. The discussion of equivalent circuits begins with the statement of two very important theorems, summarized in Figures 5.3.6 and 5.3.7.

FIGURE 5.3.6 Illustration of Thévenin theorem.

FIGURE 5.3.7 Illustration of Norton theorem.

The Thévenin Thorem. As far as a load is concerned, any network composed of ideal voltage and current sources, and of linear resistors, may be represented by an equivalent circuit consisting of an ideal voltage source, v_T, in series with an equivalent resistance, R_T.
The Norton Theorem. As far as a load is concerned, any network composed of ideal voltage and current sources, and of linear resistors, may be represented by an equivalent circuit consisting of an ideal current source, i_N, in parallel with an equivalent resistance, R_N.

Determination of Norton or Thévenin Equivalent Resistance

The first step in computing a Thévenin or Norton equivalent circuit consists of finding the equivalent resistance presented by the circuit at its terminals. This is done by setting all sources in the circuit equal to zero and computing the effective resistance between terminals. The voltage and current sources present in the circuit are set to zero as follows: voltage sources are replaced by short circuits, current sources by open circuits. We can produce a set of simple rules as an aid in the computation of the Thévenin (or Norton) equivalent resistance for a linear resistive circuit.

Computation of Equivalent Resistance of a One-Port Network:

1. Remove the load.
2. Zero all voltage and current sources
3. Compute the total resistance between load terminals, *with the load removed*. This resistance is equivalent to that which would be encountered by a current source connected to the circuit in place of the load.

These rules are summarized in Figure 5.3.8.

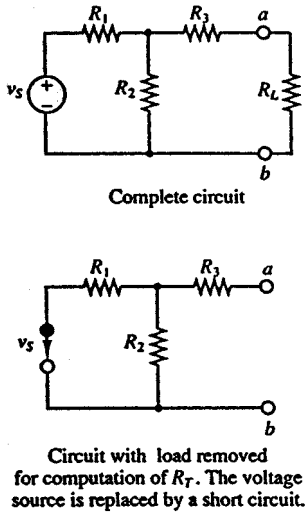


FIGURE 5.3.8 Computation of Thévenin resistance.

Computing the Thévenin Voltage

The Thévenin equivalent voltage is defined as follows: the equivalent (Thévenin) source voltage is equal to the **open-circuit voltage** present at the load terminals with the load removed.

This states that in order to compute v_T, it is sufficient to remove the load and to compute the open-circuit voltage at the one-port terminals. Figure 5.3.9 illustrates that the open-circuit voltage, v_{OC}, and the Thévenin voltage, v_T, must be the same if the Thévenin theorem is to hold. This is true because in the circuit consisting of v_T and R_T, the voltage v_{OC} must equal v_T, since no current flows through R_T and therefore the voltage across R_T is zero. Kirchhoff's voltage law confirms that

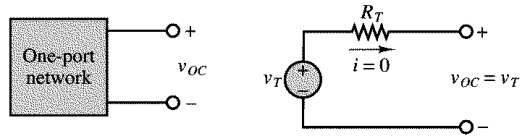


FIGURE 5.3.9 Equivalence of open-circuit and Thévenin voltage.

$$v_T = R_T(0) + v_{OC} = v_{OC} \tag{5.3.2}$$

Computing the Norton Current

The computation of the Norton equivalent current is very similar in concept to that of the Thévenin voltage. The following definition will serve as a starting point.

Definition. The Norton equivalent current is equal to the **short-circuit current** that would flow were the load replaced by a short circuit.

An explanation for the definition of the Norton current is easily found by considering, again, an arbitrary one-port network, as shown in Figure 5.3.10, where the one-port network is shown together with its Norton equivalent circuit.

It should be clear that the current, i_{SC}, flowing through the short circuit replacing the load is exactly the Norton current, i_N, since all of the source current in the circuit of Figure 5.3.10 must flow through the short circuit.

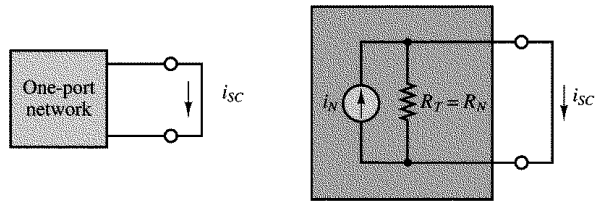


FIGURE 5.3.10 Illustration of Norton equivalent circuit.

Experimental Determination of Thévenin and Norton Equivalents

Figure 5.3.11 illustrates the measurement of the open-circuit voltage and short-circuit current for an arbitrary network connected to any load and also illustrates that the procedure requires some special attention, because of the nonideal nature of any practical measuring instrument. The figure clearly illustrates that in the presence of finite meter resistance, r_m, one must take this quantity into account in the computation of the short-circuit current and open-circuit voltage; v_{OC} and i_{SC} appear between quotation marks in the figure specifically to illustrate that the measured "open-circuit voltage" and "short-circuit current" are, in fact, affected by the internal resistance of the measuring instrument and are not the true quantities.

The following are expressions for the true short-circuit current and open-circuit voltage.

$$\begin{aligned} i_N &= "i_{SC}"\left(1 + \frac{r_m}{R_T}\right) \\ v_T &= "v_{OC}"\left(1 + \frac{R_T}{r_m}\right) \end{aligned} \tag{5.3.3}$$

where i_N is the ideal Norton current, v_T the Thévenin voltage, and R_T the true Thévenin resistance.

Nonlinear Circuit Elements

Description of Nonlinear Elements

There are a number of useful cases in which a simple functional relationship exists between voltage and current in a nonlinear circuit element. For example, Figure 5.3.12 depicts an element with an exponential *i-v* characteristic, described by the following equations:

$$\begin{aligned} i &= I_0 e^{\alpha v} \qquad v > 0 \\ i &= -I_0 \qquad\quad v \le 0 \end{aligned} \tag{5.3.4}$$

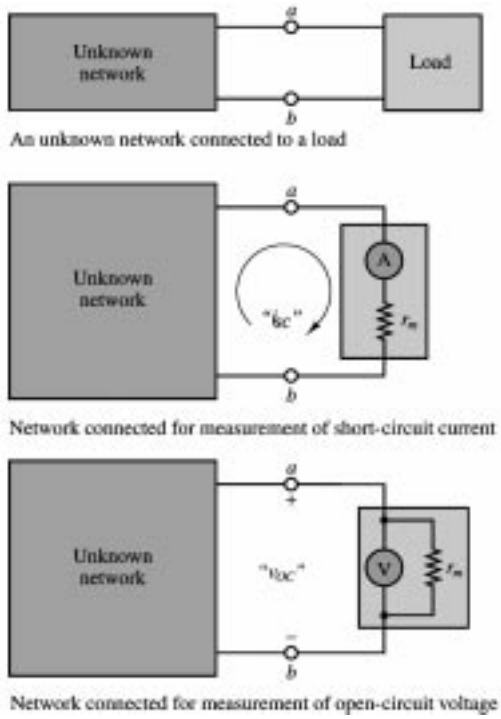


FIGURE 5.3.11 Measurement of open-circuit voltage and short-circuit current.

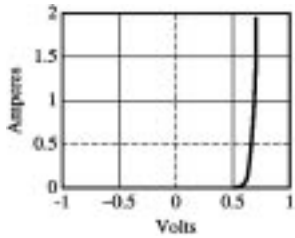


FIGURE 5.3.12 *i-v* characteristic of exponential resistor.

There exists, in fact, a circuit element (the semiconductor diode) that very nearly satisfies this simple relationship. The difficulty in the *i-v* relationship of Equation (5.3.4) is that it is not possible, in general, to obtain a closed-form analytical solution, even for a very simple circuit.

One approach to analyzing a circuit containing a nonlinear element might be to treat the nonlinear element as a load, and to compute the Thévenin equivalent of the remaining circuit, as shown in Figure 5.3.13. Applying KVL, the following equation may then be obtained:

$$v_T = R_T i_x + v_x \tag{5.3.5}$$

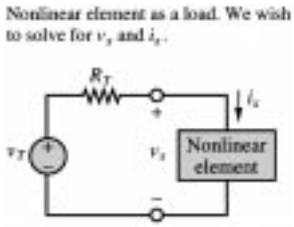


FIGURE 5.3.13 Representation of nonlinear element in a linear circuit.

To obtain the second equation needed to solve for both the unknown voltage, v_x, and the unknown current, i_x, it is necessary to resort to the *i-v* description of the nonlinear element, namely, Equation (5.3.4). If, for the moment, only positive voltages are considered, the circuit is completely described by the following system:

$$\begin{aligned} i_x &= I_0 e^{\alpha v_x} \qquad v > 0 \\ v_T &= R_T i_x + v_x \end{aligned} \tag{5.3.6}$$

The two parts of Equation (5.3.6) represent a system of two equations in two unknowns. Any numerical method of choice may now be applied to solve the system of Equations (5.3.6).

5.4 AC Network Analysis

In this section we introduce energy-storage elements, dynamic circuits, and the analysis of circuits excited by sinusoidal voltages and currents. Sinusoidal (or AC) signals constitute the most important class of signals in the analysis of electrical circuits. The simplest reason is that virtually all of the electric power used in households and industries comes in the form of sinusoidal voltages and currents.

Energy-Storage (Dynamic) Circuit Elements

The ideal resistor was introduced through Ohm's law in Section 5.2 as a useful idealization of many practical electrical devices. However, in addition to resistance to the flow of electric current, which is purely a dissipative (i.e., an energy-loss) phenomenon, electric devices may also exhibit energy-storage properties, much in the same way a spring or a flywheel can store mechanical energy. Two distinct mechanisms for energy storage exist in electric circuits: **capacitance** and **inductance**, both of which lead to the storage of energy in an electromagnetic field.

The Ideal Capacitor

A physical capacitor is a device that can store energy in the form of a charge separation when appropriately polarized by an electric field (i.e., a voltage). The simplest capacitor configuration consists of two parallel conducting plates of cross-sectional area A, separated by air (or another **dielectric**[*] material, such as mica or Teflon). Figure 5.4.1 depicts a typical configuration and the circuit symbol for a capacitor.

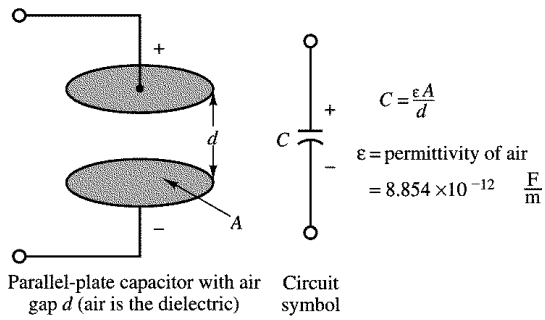


FIGURE 5.4.1 Structure of parallel-plate capacitor.

The presence of an insulating material between the conducting plates does not allow for the flow of DC current; thus, *a capacitor acts as an open circuit in the presence of DC currents*. However, if the voltage present at the capacitor terminals changes as a function of time, so will the charge that has accumulated at the two capacitor plates, since the degree of polarization is a function of the applied electric field, which is time-varying. In a capacitor, the charge separation caused by the polarization of the dielectric is proportional to the external voltage, that is, to the applied electric field:

$$Q = CV \tag{5.4.1}$$

where the parameter C is called the *capacitance* of the element and is a measure of the ability of the device to accumulate, or store, charge. The unit of capacitance is the coulomb/volt and is called the **farad** (F). The farad is an unpractically large unit; therefore, it is common to use microfarads (1 μF =

[*] A dielectric material contains a large number of electric dipoles, which become polarized in the presence of an electric field.

10^{-6} F) or picofarads (1 pF = 10^{-12} F). From Equation (5.4.1) it becomes apparent that if the external voltage applied to the capacitor plates changes in time, so will the charge that is internally stored by the capacitor:

$$q(t) = Cv(t) \tag{5.4.2}$$

Thus, although no current can flow through a capacitor if the voltage across it is constant, a time-varying voltage will cause charge to vary in time. The change with time in the stored charge is analogous to a current. The relationship between the current and voltage in a capacitor is as follows:

$$i(t) = C\frac{dv(t)}{dt} \tag{5.4.3}$$

If the above differential equation is integrated, one can obtain the following relationship for the voltage across a capacitor:

$$v_C(t) = \frac{1}{C}\int_{-\infty}^{t_0} i_C\, dt \tag{5.4.4}$$

Equation (5.4.4) indicates that the capacitor voltage depends on the past current through the capacitor, up until the present time, t. Of course, one does not usually have precise information regarding the flow of capacitor current for all past time, and so it is useful to define the initial voltage (or *initial condition*) for the capacitor according to the following, where t_0 is an arbitrary initial time:

$$V_0 = v_C\left(t = t_0\right) = \frac{1}{C}\int_{-\infty}^{t} i_C\, dt \tag{5.4.5}$$

The capacitor voltage is now given by the expression

$$v_C(t) = \frac{1}{C}\int_{t_0}^{t} i_C\, dt + V_0 \qquad t \geq t_0 \tag{5.4.6}$$

The significance of the initial voltage, V_0, is simply that at time t_0 some charge is stored in the capacitor, giving rise to a voltage, $v_c\,(t_0)$, according to the relationship $Q = CV$. Knowledge of this initial condition is sufficient to account for the entire past history of the capacitor current. (See Figure 5.4.2.)

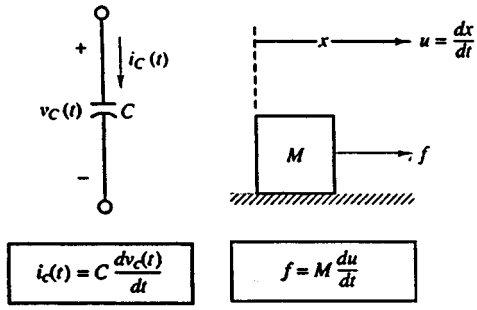


FIGURE 5.4.2 Defining equation for the ideal capacitor, and analogy with force-mass system.

From the standpoint of circuit analysis, it is important to point out that capacitors connected in series and parallel can be combined to yield a single equivalent capacitance. The rule of thumb, which is

illustrated in Figure 5.4.3, is the following: capacitors in parallel add; capacitors in series combine according to the same rules used for resistors connected in parallel.

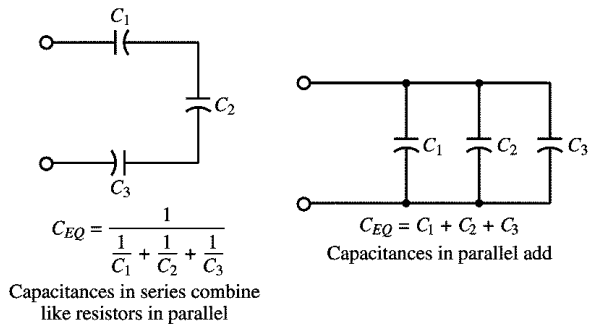


FIGURE 5.4.3 Combining capacitors in a circuit.

Physical capacitors are rarely constructed of two parallel plates separated by air, because this configuration yields very low values of capacitance, unless one is willing to tolerate very large plate areas. In order to increase the capacitance (i.e., the ability to store energy), physical capacitors are often made of tightly rolled sheets of metal film, with a dielectric (paper or Mylar) sandwiched in-between. Table 5.4.1 illustrates typical values, materials, maximum voltage ratings, and useful frequency ranges for various types of capacitors. The voltage rating is particularly important, because any insulator will break down if a sufficiently high voltage is applied across it.

TABLE 5.4.1 Capacitors

Material	Capacitance Range	Maximum Voltage (V)	Frequency Range (Hz)
Mica	1 pF to 0.1 μF	100–600	10^3–10^{10}
Ceramic	10 pF to 1 μF	50–1000	10^3–10^{10}
Mylar	0.001 to 10 μF	50–500	10^2–10^8
Paper	1,000 pF to 50 μF	100–105	10^2–10^8
Electrolytic	0.1 μF to 0.2 F	3–600	10–10^4

The energy stored in a capacitor is given by:

$$W_C(t) = \frac{1}{2} C v_C^2(t) \quad \text{(J)}$$

Example 5.4.1 Capacitive Displacement Transducer and Microphone

As shown in Figure 5.3.5, the capacitance of a parallel-plate capacitor is given by the expression

$$C = \frac{\varepsilon A}{d}$$

where ε is the **permittivity** of the dielectric material, A the area of each of the plates, and d their separation. The permittivity of air is $\varepsilon_0 = 8.854 \times 10^{-12}$ F/m, so that two parallel plates of area 1 m^2, separated by a distance of 1 mm, would give rise to a capacitance of 8.854×10^{-3} μF, a very small value for a very large plate area. This relative inefficiency makes parallel-plate capacitors impractical for use in electronic circuits. On the other hand, parallel-plate capacitors find application as *motion transducers*, that is, as devices that can measure the motion or displacement of an object. In a capacitive motion

transducer, the air gap between the plates is designed to be variable, typically by fixing one plate and connecting the other to an object in motion. Using the capacitance value just derived for a parallel-plate capacitor, one can obtain the expression

$$C = \frac{8.854 \times 10^{-3}\ A}{x}$$

where C is the capacitance in pF, A is the area of the plates in mm^2, and x is the (variable) distance in mm. It is important to observe that the change in capacitance caused by the displacement of one of the plates is nonlinear, since the capacitance varies as the inverse of the displacement. For small displacements, however, the capacitance varies approximately in a linear fashion.

The *sensitivity, S,* of this motion transducer is defined as the slope of the change in capacitance per change in displacement, x, according to the relation

$$S = \frac{dC}{dx} = -\frac{8.854 \times 10^{-3}\ A}{2x^2}\ \frac{\text{pF}}{\text{mm}}$$

Thus, the sensitivity increases for small displacements. This behavior can be verified by plotting the capacitance as a function of x and noting that as x approaches zero, the slope of the nonlinear $C(x)$ curve becomes steeper (thus the greater sensitivity). Figure 5.4.4 depicts this behavior for a transducer with area equal to 10 mm^2.

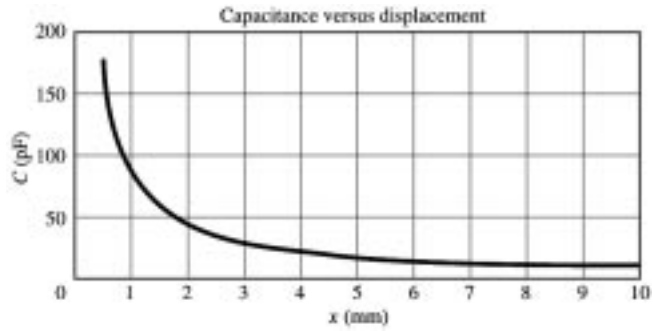


FIGURE 5.4.4 Response of a capacitive displacement transducer.

This simple capacitive displacement transducer actually finds use in the popular *capacitive (or condenser) microphone*, in which the sound pressure waves act to displace one of the capacitor plates. The change in capacitance can then be converted into a change in voltage or current by means of a suitable circuit. An extension of this concept that permits measurement of differential pressures is shown in simplified form in Figure 5.4.5. In the figure, a three-terminal variable capacitor is shown to be made up of two fixed surfaces (typically, spherical depressions ground into glass disks and coated with a conducting material) and of a deflecting plate (typically made of steel) sandwiched between the glass disks. Pressure inlet orifices are provided, so that the deflecting plate can come into contact with the fluid whose pressure it is measuring. When the pressure on both sides of the deflecting plate is the same, the capacitance between terminals b and d, C_{bd}, will be equal to that between terminals b and c, C_{bc}. If any pressure differential exists, the two capacitances will change, with an increase on the side where the deflecting plate has come closer to the fixed surface and a corresponding decrease on the other side.

This behavior is ideally suited for the application of a bridge circuit, similar to the Wheatstone bridge circuit illustrated in Example 5.3.1, and also shown in Figure 5.4.5. In the bridge circuit, the output voltage, v_{out}, is precisely balanced when the differential pressure across the transducer is zero, but it will

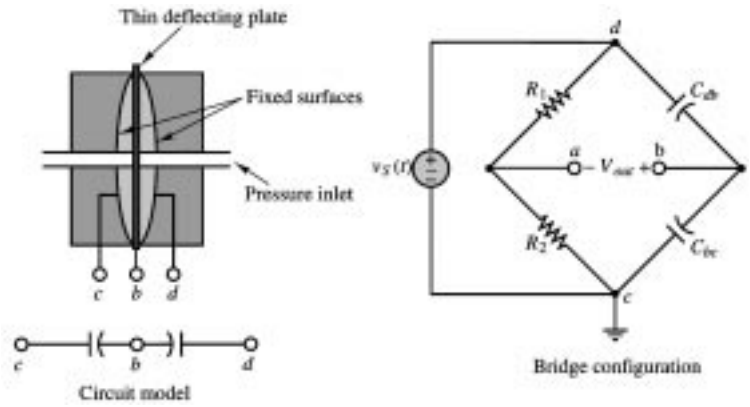


FIGURE 5.4.5 Capacitive pressure transducer and related bridge circuit.

deviate from zero whenever the two capacitances are not identical because of a pressure differential across the transducer. We shall analyze the bridge circuit later in Example 5.4.2.

The Ideal Inductor

The ideal inductor is an element that has the ability to store energy in a magnetic field. Inductors are typically made by winding a coil of wire around a core, which can be an insulator or a ferromagnetic material, shown in Figure 5.4.6. When a current flows through the coil, a magnetic field is established,

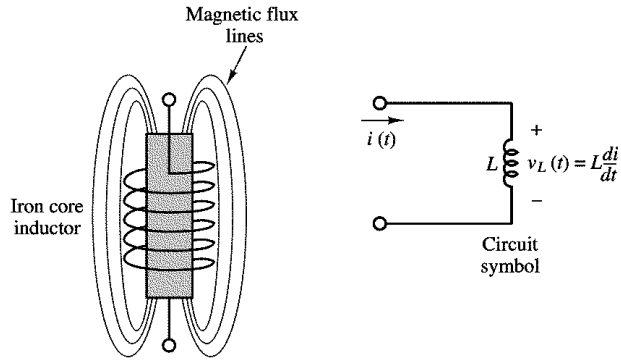


FIGURE 5.4.6 Iron-core inductor.

as you may recall from early physics experiments with electromagnets. In an ideal inductor, the resistance of the wire is zero, so that a constant current through the inductor will flow freely without causing a voltage drop. In other words, *the ideal inductor acts as a short circuit in the presence of DC currents*. If a time-varying voltage is established across the inductor, a corresponding current will result, according to the following relationship:

$$v_L(t) = L\frac{di_L}{dt} \tag{5.4.7}$$

where L is called the *inductance* of the coil and is measured **henry** (H), where

$$1\text{ H} = 1\text{ V-sec/A} \tag{5.4.8}$$

Henrys are reasonable units for practical inductors; millihenrys (mH) and microhenrys (μH) are also used.

The inductor current is found by integrating the voltage across the inductor:

$$i_L(t) = \frac{1}{L}\int_{-\infty}^{t} v_L \, dt \tag{5.4.9}$$

If the current flowing through the inductor at time $t = t_0$ is known to be I_0, with

$$I_0 = i_L(t = t_0) = \frac{1}{L}\int_{-\infty}^{t_0} v_L \, dt \tag{5.4.10}$$

then the inductor current can be found according to the equation

$$i_L(t) = \frac{1}{L}\int_{t_0}^{t} v_L \, dt + I_0 \qquad t \geq t_0 \tag{5.4.11}$$

Inductors in series add. Inductors in parallel combine according to the same rules used for resistors connected in parallel. See Figures 5.4.7 to 5.4.9.

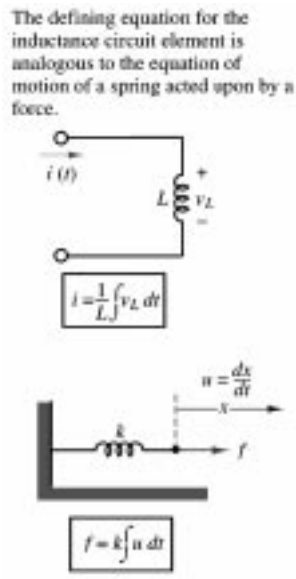


FIGURE 5.4.7 Defining equation for the ideal inductor and analogy with force-spring system.

Table 5.4.2 and Figures 5.4.2, 5.4.7, and 5.4.9 illustrate a useful analogy between ideal electrical and mechanical elements.

Time-Dependent Signal Sources

Figure 5.4.10 illustrates the convention that will be employed to denote time-dependent signal sources.

One of the most important classes of time-dependent signals is that of **periodic signals**. These signals appear frequently in practical applications and are a useful approximation of many physical phenomena. A periodic signal $x(t)$ is a signal that satisfies the following equation:

$$x(t) = x(t + nT) \qquad n = 1, 2, 3, \ldots \tag{5.4.12}$$

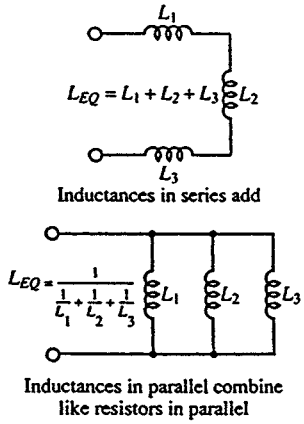


FIGURE 5.4.8 Combining inductors in a circuit.

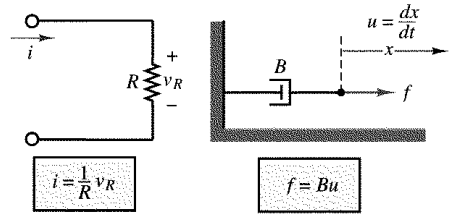


FIGURE 5.4.9 Analogy between electrical and mechanical elements.

TABLE 5.4.2 Analogy Between Electrical and Mechanical Variables

Mechanical System	Electrical System
Force, f (N)	Current, i (A)
Velocity, μ (m/sec)	Voltage, v (V)
Damping, B (N-sec/m)	Conductance, 1/R (S)
Compliance, 1/k (m/N)	Inductance, L (H)
Mass, M (kg)	Capacitance, C (F)

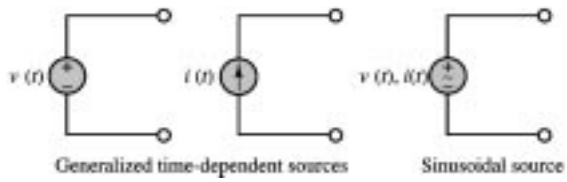


FIGURE 5.4.10 Time-dependent signal sosurces.

where T is the **period** of $x(t)$. Figure 5.4.11 illustrates a number of the periodic waveforms that are typically encountered in the study of electrical circuits. Waveforms such as the sine, triangle, square, pulse, and sawtooth waves are provided in the form of voltages (or, less frequently, currents) by commercially available **signal** (or **waveform**) **generators**. Such instruments allow for selection of the waveform peak amplitude, and of its period.

As stated in the introduction, sinusoidal waveforms constitute by far the most important class of time-dependent signals. Figure 5.4.12 depicts the relevant parameters of a sinusoidal waveform. A generalized sinusoid is defined as follows:

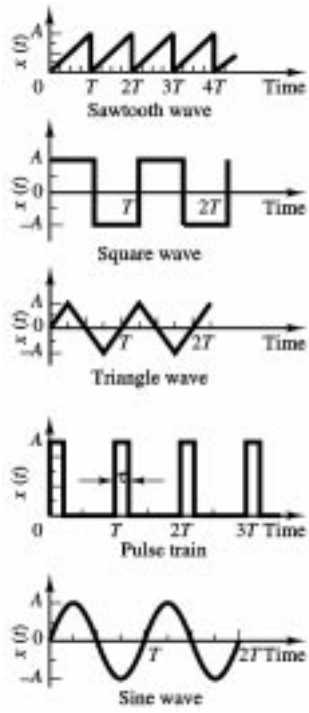


FIGURE 5.4.11 Periodic signal waveforms.

$$x(t) = A\cos(\omega t + \phi) \tag{5.4.13}$$

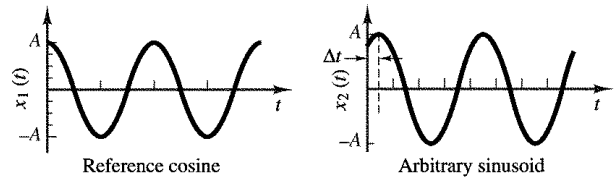


FIGURE 5.4.12 Sinusoidal waveforms.

where A is the **amplitude**, ω the **radian frequency**, and ϕ the **phase**. Figure 5.4.12 summarizes the definitions of A, ω, and ϕ for the waveforms

$$x_1(t) = A\cos(\omega t) \quad \text{and} \quad x_2(t) = A\cos(\omega t + \phi)$$

where

$$f = \text{natural frequency} = \frac{1}{T}(\text{cycles/sec, or Hz})$$

$$\omega = \text{radian frequency} = 2\pi f(\text{radians/sec}) \tag{5.4.14}$$

$$\phi = 2\pi\frac{\Delta T}{T}(\text{radians}) = 360\frac{\Delta T}{T}(\text{degrees})$$

The phase shift, ϕ, permits the representation of an arbitrary sinusoidal signal. Thus, the choice of the reference cosine function to represent sinusoidal signals — arbitrary as it may appear at first — does

not restrict the ability to represent all sinusoids. For example, one can represent a sine wave in terms of a cosine wave simply by introducing a phase shift of π/2 radians:

$$A\sin(\omega t) = A\cos\left(\omega t - \frac{\pi}{2}\right) \tag{5.4.15}$$

It is important to note that, although one usually employs the variable ω (in units of radians per second) to denote sinusoidal frequency, it is common to refer to natural frequency, f, in units of cycles per second, or **hertz** (Hz). The relationship between the two is the following:

$$\omega = 2\pi f \tag{5.4.16}$$

Average and RMS Values

Now that a number of different signal waveforms have been defined, it is appropriate to define suitable measurements for quantifying the strength of a time-varying electrical signal. The most common types of measurements are the **average** (or **DC**) **value** of a signal waveform — which corresponds to just measuring the mean voltage or current over a period of time — and the **root-mean-square** (or **rms**) **value**, which takes into account the fluctuations of the signal about its average value. Formally, the operation of computing the average value of a signal corresponds to integrating the signal waveform over some (presumably, suitably chosen) period of time. We define the time-averaged value of a signal $x(t)$ as

$$\langle x(t)\rangle = \frac{1}{T}\int_0^T x(t)\,dt \tag{5.4.17}$$

where T is the period of integration. Figure 5.4.13 illustrates how this process does, in fact, correspond to computing the average amplitude of $x(t)$ over a period of T seconds.

FIGURE 5.4.13 Averaging a signal waveform.

$$\langle A\cos(\omega t + \phi)\rangle = 0$$

A useful measure of the voltage of an AC waveform is the root-mean-square, or rms, value of the signal, $x(t)$, defined as follows:

$$x_{rms}\sqrt{\frac{1}{T}\int_0^T x^2(t)\,dt} \tag{5.4.18}$$

Note immediately that if $x(t)$ is a voltage, the resulting x_{rms} will also have units of volts. If you analyze Equation (5.4.18), you can see that, in effect, the rms value consists of the square root of the average (or mean) of the square of the signal. Thus, the notation *rms* indicates exactly the operations performed on $x(t)$ in order to obtain its rms value.

Solution of Circuits Containing Dynamic Elements

The major difference between the analysis of the resistive circuits and circuits containing capacitors and inductors is now that the equations that result from applying Kirchhoff's laws are differential equations,

as opposed to the algebraic equations obtained in solving resistive circuits. Consider, for example, the circuit of Figure 5.4.14 which consists of the series connection of a voltage source, a resistor, and a capacitor. Applying KVL around the loop, we may obtain the following equation:

$$v_S(t) = v_R(t) + v_C(t) \tag{5.4.19}$$

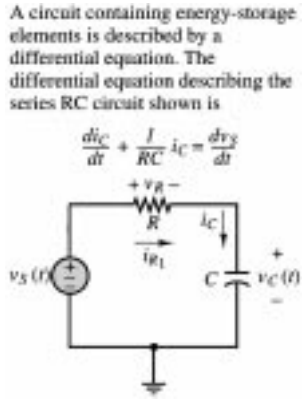


FIGURE 5.4.14 Circuit containing energy-storage element.

Observing that $i_R = i_C$, Equation (5.4.19) may be combined with the defining equation for the capacitor (Equation 4.6.6) to obtain

$$v_S(t) = Ri_C(t) + \frac{1}{C}\int_{-\infty}^{t} i_C\, dt \tag{5.4.20}$$

Equation (5.4.20) is an integral equation, which may be converted to the more familiar form of a differential equation by differentiating both sides of the equation, and recalling that

$$\frac{d}{dt}\left(\int_{-\infty}^{t} i_C\, dt\right) = i_C(t) \tag{5.4.21}$$

to obtain the following differential equation:

$$\frac{di_C}{dt} + \frac{1}{RC} i_C = \frac{1}{R}\frac{dv_S}{dt} \tag{5.4.22}$$

where the argument (t) has been dropped for ease of notation.

Observe that in Equation (5.4.22), the independent variable is the series current flowing in the circuit, and that this is not the only equation that describes the series *RC* circuit. If, instead of applying KVL, for example, we had applied KCL at the node connecting the resistor to the capacitor, we would have obtained the following relationship:

$$i_R = \frac{v_S - v_C}{R} = i_C = C\frac{dv_C}{dt} \tag{5.4.23}$$

or

$$\frac{dv_C}{dt} + \frac{1}{RC} v_C = \frac{1}{RC} v_S \tag{5.4.24}$$

Note the similarity between Equations (5.4.22) and (5.4.24). The left-hand side of both equations is identical, except for the dependent variable, while the right-hand side takes a slightly different form. The solution of either equation is sufficient, however, to determine all voltages and currents in the circuit.

We can generalize the results above by observing that any circuit containing a single energy-storage element can be described by a differential equation of the form

$$a_1 \frac{dy(t)}{dt} + a_0(t) = F(t) \tag{5.4.25}$$

where $y(t)$ represents the capacitor voltage in the circuit of Figure 5.4.14 and where the constants a_0 and a_1 consist of combinations of circuit element parameters. Equation (5.4.25) is a **first-order ordinary differential equation** with constant coefficients.

Consider now a circuit that contains two energy-storage elements, such as that shown in Figure 5.4.15. Application of KVL results in the following equation:

$$Ri(t) + L\frac{di(t)}{dt} + \frac{1}{C}\int_{-\infty}^{t} i(t)\,dt = v_S(t) \tag{5.4.26}$$

Equation (5.4.26) is called an integro-differential equation because it contains both an integral and a derivative. This equation can be converted into a differential equation by differentiating both sides, to obtain:

$$R\frac{di(t)}{dt} + L\frac{d^2i(t)}{dt^2} + \frac{1}{C}i(t) = \frac{dv_S(t)}{dt} \tag{5.4.27}$$

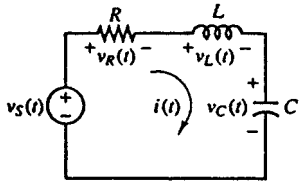


FIGURE 5.4.15 Second-order circuit.

or, equivalently, by observing that the current flowing in the series circuit is related to the capacitor voltage by $i(t) = Cdv_C/dt$, and that Equation (5.4.26) can be rewritten as:

$$RC\frac{dv_C}{dt} + LC\frac{d^2v_C(t)}{dt^2} + v_C(t) = v_S(t) \tag{5.4.28}$$

Note that although different variables appear in the preceding differential equations, both Equations (5.4.26) and (5.4.28) can be rearranged to appear in the same general form, as follows:

$$a_2\frac{d^2y(t)}{dt^2} + a_1\frac{dy(t)}{dt} + a_0y(t) = F(t) \tag{5.4.29}$$

where the general variable $y(t)$ represents either the series current of the circuit of Figure 5.4.15 or the capacitor voltage. By analogy with Equation (5.4.25), we call Equation (5.4.29) a **second-order ordinary differential equation** with constant coefficients. As the number of energy-storage elements in a circuit increases, one can therefore expect that higher-order differential equations will result.

Phasors and Impedance

In this section, we introduce an efficient notation to make it possible to represent sinusoidal signals as *complex numbers*, and to eliminate the need for solving differential equations.

Phasors

Let us recall that it is possible to express a generalized sinusoid as the real part of a complex vector whose **argument**, or **angle**, is given by $(\omega t + \phi)$ and whose length, or **magnitude**, is equal to the peak amplitude of the sinusoid. The **complex phasor** corresponding to the sinusoidal signal $A\cos(\omega t + \phi)$ is therefore defined to be the complex number $Ae^{j\phi}$:

$$Ae^{j\phi} = \text{complex phasor notation for } A\cos(\omega t + \phi) \tag{5.4.30}$$

1. Any sinusoidal signal may be mathematically represented in one of two ways: a **time-domain form**,

$$v(t) = A\cos(\omega t + \phi)$$

and a **frequency-domain** (or **phasor**) **form**.

$$\mathbf{V}(j\omega) = Ae^{j\phi}$$

2. A phasor is a complex number, expressed in polar form, consisting of a *magnitude* equal to the peak amplitude of the sinusoidal signal and a *phase angle* equal to the phase shift of the sinusoidal signal *referenced to a cosine signal*.
3. When using phasor notation, it is important to make a note of the specific frequency, ω, of the sinusoidal signal, since this is not explicitly apparent in the phasor expression.

Impedance

We now analyze the *i-v* relationship of the three ideal circuit elements in light of the new phasor notation. The result will be a new formulation in which resistors, capacitors, and inductors will be described in the same notation. A direct consequence of this result will be that the circuit theorems of Section 5.3 will be extended to AC circuits. In the context of AC circuits, any one of the three ideal circuit elements defined so far will be described by a parameter called **impedance**, which may be viewed as a *complex resistance*. The impedance concept is equivalent to stating that capacitors and inductors act as *frequency-dependent resistors*, that is, as resistors whose resistance is a function of the frequency of the sinusoidal excitation. Figure 5.4.16 depicts the same circuit represented in conventional form (top) and in phasor-impedance form (bottom); the latter representation explicitly shows phasor voltages and currents and treats the circuit element as a generalized "impedance". It will presently be shown that each of the three ideal circuit elements may be represented by one such impedance element.

Let the source voltage in the circuit of Figure 5.4.16 be defined by

$$v_S(t) = A\cos\omega t \quad \text{or} \quad \mathbf{V}_S(j\omega) = Ae^{j0^\circ} \tag{5.4.31}$$

without loss of generality. Then the current $i(t)$ is defined by the *i-v* relationship for each circuit element. Let us examine the frequency-dependent properties of the resistor, inductor, and capacitor, one at a time.

The *impedance* of the resistor is defined as the ratio of the phasor voltage across the resistor to the phasor current flowing through it, and the symbol Z_R is used to denote it:

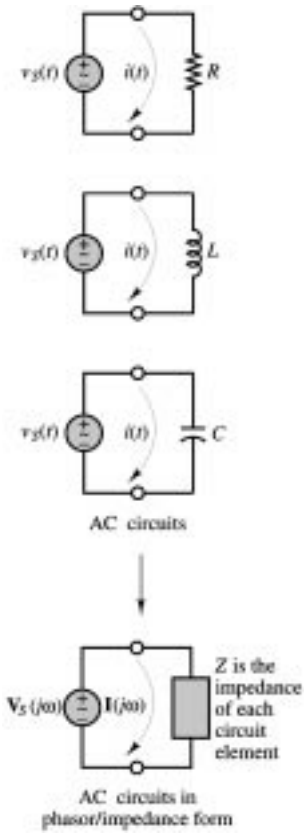


FIGURE 5.4.16 The impedance element.

$$Z_R(j\omega) = \frac{\mathbf{V}_S(j\omega)}{\mathbf{I}(j\omega)} = R \tag{5.4.32}$$

The impedance of the inductor is defined as follows:

$$Z_L(j\omega) = \frac{\mathbf{V}_S(j\omega)}{\mathbf{I}(j\omega)} = \omega L e^{j90^\circ} = j\omega L \tag{5.4.33}$$

Note that the inductor now appears to behave like a *complex frequency-dependent resistor*, and that the magnitude of this complex resistor, ωL, is proportional to the signal frequency, ω. Thus, an inductor will "impede" current flow in proportion to the sinusoidal frequency of the source signal. This means that at low signal frequencies, an inductor acts somewhat like a short circuit, while at high frequencies it tends to behave more as an open circuit. Another important point is that *the magnitude of the impedance of an inductor is always positive*, since both L and ω are positive numbers. You should verify that the units of this magnitude are also ohms.

The impedance of the ideal capacitor, $Z_C(j\omega)$, is therefore defined as follows:

$$Z_C(j\omega) = \frac{\mathbf{V}_S(j\omega)}{\mathbf{I}(j\omega)} = \frac{1}{\omega C} e^{-j90^\circ} = \frac{-j}{\omega C} = \frac{1}{j\omega C} \tag{5.4.34}$$

where we have used the fact that $1/j = e^{-j90^\circ} = -j$. Thus, the impedance of a capacitor is also a frequency-dependent complex quantity, with the impedance of the capacitor varying as an inverse function of

frequency; and so a capacitor acts like a short circuit at high frequencies, whereas it behaves more like an open circuit at low frequencies. Another important point is that *the impedance of a capacitor is always negative*, since both C and ω are positive numbers. You should verify that the units of impedance for a capacitor are ohms. Figure 5.4.17 depicts $Z_C(j\omega)$ in the complex plane, alongside $Z_R(j\omega)$ and $Z_L(j\omega)$.

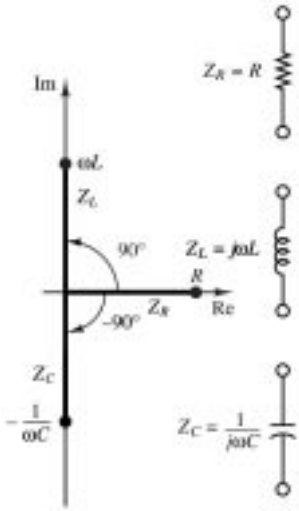


FIGURE 5.4.17 Impedances of R, L, and C in the complex plane.

The impedance parameter defined in this section is extremely useful in solving AC circuit analysis problems, because it will make it possible to take advantage of most of the network theorems developed for DC circuits by replacing resistances with complex-valued impedances. In its most general form, the impedance of a circuit element is defined as the sum of a real part and an imaginary part:

$$Z(j\omega) = R(j\omega) + jX(j\omega) \tag{5.4.35}$$

where R is called the **AC resistance** and X is called the **reactance**. The frequency dependence of R and X has been indicated explicitly, since it is possible for a circuit to have a frequency-dependent resistance. The examples illustrate how a complex impedance containing both real and imaginary parts arises in a circuit.

Example 5.4.2 Capacitive Displacement Transducer

In Example 5.4.1, the idea of a capacitive displacement transducer was introduced when we considered a parallel-plate capacitor composed of a fixed plate and a movable plate. The capacitance of this variable capacitor was shown to be a *nonlinear* function of the position of the movable plate, x (see Figure 5.4.5). In this example, we show that under certain conditions the impedance of the capacitor varies as a *linear* function of displacement — that is, the movable-plate capacitor can serve as a linear transducer.

Recall the expression derived in Example 5.4.1:

$$C = \frac{8.854 \times 10^{-3} A}{x}$$

where C is the capacitance in pF, A is the area of the plates in mm^2, and x is the (variable) distance in mm. If the capacitor is placed in an AC circuit, its impedance will be determined by the expression

$$Z_C = \frac{1}{j\omega C}$$

so that

$$Z_C = \frac{x}{j\omega 8.854A}$$

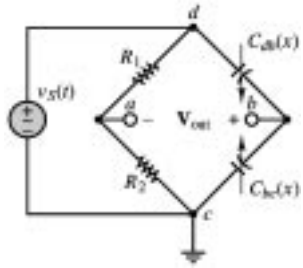


FIGURE 5.4.18 Bridge circuit for capacitive displacement transducer.

Thus, at a fixed frequency ω, the impedance of the capacitor will vary linearly with displacement. This property may be exploited in the bridge circuit of Example 5.4.1, where a differential pressure transducer was shown as being made of two movable-plate capacitors, such that if the capacitance of one increased as a consequence of a pressure differential across the transducer, the capacitance of the other had to decrease by a corresponding amount (at least for small displacements). The circuit is shown again in Figure 5.4.18 where two resistors have been connected in the bridge along with the variable capacitors (denoted by $C(x)$). The bridge is excited by a sinusoidal source.

Using phasor notation, we can express the output voltage as follows:

$$\mathbf{V}_{\text{out}}(j\omega) = \mathbf{V}_S(j\omega)\left(\frac{Z_{C_{bc}(x)}}{Z_{C_{db}(x)} + Z_{C_{bc}(x)}} - \frac{R_2}{R_1 + R_2}\right)$$

If the nominal capacitance of each movable-plate capacitor with the diaphragm in the center position is given by

$$C = \frac{\varepsilon A}{d}$$

where d is the nominal (undisplaced) separation between the diaphragm and the fixed surfaces of the capacitors (in mm), the capacitors will see a change in capacitance given by

$$C_{db} = \frac{\varepsilon A}{d - x} \quad \text{and} \quad C_{bc} = \frac{\varepsilon A}{d + x}$$

when a pressure differential exists across the transducer, so that the impedances of the variable capacitors change according to the displacement:

$$Z_{C_{db}} = \frac{d - x}{j\omega 8.854A} \quad \text{and} \quad Z_{C_{bc}} = \frac{d + x}{j\omega 8.854A}$$

and we obtain the following expression for the phasor output voltage, if we choose $R_1 = R_2$.

$$\mathbf{V}_{\text{out}}(j\omega) = \mathbf{V}_S(j\omega)\left(\frac{\dfrac{d+x}{j\omega 8.854A}}{\dfrac{d-x}{j\omega 8.854A}+\dfrac{d+x}{j\omega 8.854A}} - \frac{R_2}{R_1+R_2}\right)$$

$$= \mathbf{V}_S(j\omega)\left(\frac{1}{2}+\frac{x}{2d}-\frac{R_2}{R_1+R_2}\right)$$

$$= \mathbf{V}_S(j\omega)\frac{x}{2d}$$

Thus, the output voltage will vary as a scaled version of the input voltage in proportion to the displacement.

5.5 AC Power

The aim of this section is to introduce the student to simple AC power calculations, and to the generation and distribution of electric power.

Instantaneous and Average Power

The most general expressions for the voltage and current delivered to an arbitrary load are as follows:

$$\begin{aligned} v(t) &= V\cos(\omega t + \theta_v) \\ i(t) &= I\cos(\omega t + \theta_I) \end{aligned} \tag{5.5.1}$$

where V and I are the peak amplitudes of the sinusoidal voltage and current, respectively, and θ_V and θ_I are their angles. Two such waveforms are plotted in Figure 5.5.1, with unit amplitude and with phase angles $\theta_V = \pi/6$ and $\theta_I = \pi/3$.

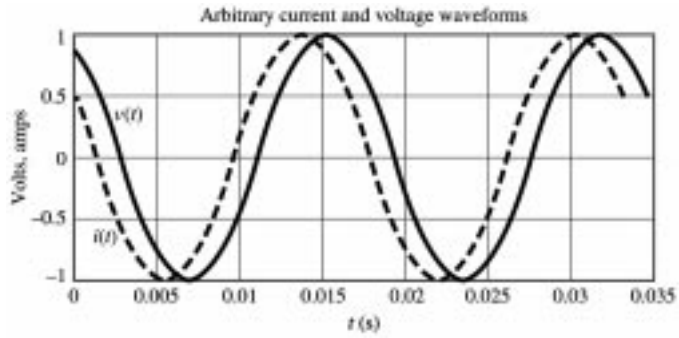


FIGURE 5.5.1 Current and voltage waveforms for illustration of AC power.

Since the **instantaneous power** dissipated by a circuit element is given by the product of the instantaneous voltage and current, it is possible to obtain a general expression for the power dissipated by an AC circuit element:

$$\begin{aligned} p(t) &= v(t)i(t) \\ &= VI\cos(\omega t + \theta_v)\cos(\omega t + \theta_I) \end{aligned} \tag{5.5.2}$$

Equation (5.5.2) can be further simplified with the aid of trigonometric identities to yield

$$p(t) = \frac{VI}{2}\cos(\theta) + \frac{VI}{2}\cos(2\omega t + \theta) \tag{5.5.3}$$

Equation (5.5.3) illustrates how the instantaneous power dissipated by an AC circuit element is equal to the sum of an average component, $VI/2\cos(\theta)$, plus a sinusoidal component, $VI/2\cos(2\omega t + \theta)$, oscillating at a frequency double that of the original source frequency.

The **average power** corresponding to the voltage and current signals of Equation (5.5.1) can be obtained by integrating the instantaneous power over one cycle of the sinusoidal signal.

$$P_{av} = \frac{VI}{2}\cos(\theta)$$

AC Power Notation

It has already been noted that AC power systems operate at a fixed frequency; in North America, this frequency is 60 cycles per second (Hz), corresponding to a radian frequency

$$\omega = 2\pi \cdot 60 = 377 \text{ rad/sec} \quad \text{AC power frequency} \tag{5.5.4}$$

In Europe and most other parts of the world, AC power is generated at a frequency of 50 Hz (this is the reason why some appliances will not operate under one of the two systems). It will therefore be understood that for the remainder of this section the radian frequency, ω, is fixed at 377 rad/sec. With knowledge of the radian frequency of all voltages and currents, it will always be possible to compute the exact magnitude and phase of any impedance in a circuit.

Power Factor

The phase angle of the load impedance plays a very important role in the absorption of power by a load impedance. As illustrated in Equation (5.5.3) and in the preceding examples, the average power dissipated by an AC load is dependent on the cosine of the angle of the impedance. To recognize the importance of this factor in AC power computations, the term cos(θ) is referred to as the **power factor** (pf). Note that the power factor is equal to 0 for a purely inductive or capacitive load and equal to 1 for a purely resistive load; in every other case,

$$0 < \text{pf} < 1 \tag{5.5.5}$$

Two equivalent expressions for the power factor are given in the following:

$$\text{pf} = \cos(\theta) = \frac{P_{av}}{V_f I_f} \quad \text{Power factor} \tag{5.5.6}$$

where V_f and I_f are the rms values of the load voltage and current.

Complex Power

The expression for the instantaneous power given in Equation (5.5.3) may be further expanded to provide further insight into AC power.

$$p(t) = I_f^2 R + I_f^2 R\cos(2\omega t) - I_f^2 X\sin(2\omega t) \tag{5.5.7}$$

The physical interpretation of this expression for the instantaneous power should be intuitively appealing at this point. As Equation (5.5.7) suggests, the instantaneous power dissipated by a complex load consists of the following three components:

1. An average component, which is constant; this is called the *average power* and is denoted by the symbol P_{av}:

$$P_{av} = I_f^2 R \tag{5.5.8}$$

where R = Re (Z).

2. A time-varying (sinusoidal) component with zero average value that is contributed by the power fluctuations in the resistive component of the load and is denoted by $p_R(t)$:

$$p_R(t) = I^2 R\cos(2\omega t) = P_{av}\cos 2\omega t \tag{5.5.9}$$

3. A time-varying (sinusoidal component with zero average value, due to the power fluctuation in the reactive component of the load and denoted by $p_X(t)$:

$$p_X(t) = -I^2 X\sin(2\omega t) = Q\sin 2\omega t \tag{5.5.10}$$

where X = Im (Z) and Q is called the **reactive power**. *Note that since reactive elements can only store energy and not dissipate it, there is no net average power absorbed by X.*

Since P_{av} corresponds to the power absorbed by the load resistance, it is also called the **real power**, measured in units of watts (W). On the other hand, Q takes the name of *reactive power*, since it is associated with the load reactance. Table 5.5.1 shows the general methods of calculating P and Q.

TABLE 5.5.1 Real and Reactive Power

Real Power, P_{av}	Reactive Power, Q
$VI\cos(\theta)$	$VI\sin(\theta)$
I^2R	I^2X

The units of Q are **volt-amperes reactive**, or VAR. Note that Q represents an exchange of energy between the source and the reactive part of the load; thus, no net power is gained or lost in the process since the average reactive power is zero. In general, it is desirable to minimize the reactive power in a load.

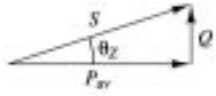
The computation of AC power is greatly simplified by defining a fictitious but very useful quantity called the **complex power**, S:

$$S = \mathbf{VI}^* \quad \text{Complex power} \tag{5.5.11}$$

or

$$S = P_{av} + jQ$$

The complex power S may be interpreted graphically as a vector in the complex plane, as shown in Figure 5.5.2.



$$|S| = \sqrt{P_{av}^2 + Q^2} = \bar{V}\cdot\bar{I}$$

$$P_{av} = \bar{V}\bar{I}\cos\theta$$

$$Q = \bar{V}\bar{I}\sin\theta$$

FIGURE 5.5.2 The complex power triangle.

The magnitude of S, $|S|$, is measured in units of **volt-amperes** (VA) and is called **apparent power**, because this is the quantity one would compute by measuring the rms load voltage and currents without regard for the phase angle of the load. The complex power may also be expressed by the product of the square of the rms current through the load and the complex load impedance:

$$S = \mathbf{I}^2 Z$$

or

$$\mathbf{I}^2 R + j\mathbf{I}^2 R = \mathbf{I}^2 Z \tag{5.5.12}$$

or, equivalently, by the ratio of the rms voltage across the load to the complex conjugate of the load impedance:

$$S = \frac{\mathbf{V}^2}{Z^*} \tag{5.5.13}$$

Although the reactive power does not contribute to any average power dissipation in the load, it may have an adverse effect on power consumption because it increases the overall rms current flowing in the circuit. The presence of any source resistance (typically, the resistance of the line wires in AC power circuits) will cause a loss of power; the power loss due to this line resistance is unrecoverable and constitutes a net loss for the electric company, since the user never receives this power.

Power Factor, Revisited

The power factor, defined earlier as the cosine of the angle of the load impedance, plays a very important role in AC power. A power factor close to unity signifies an efficient transfer from the AC source to the load, while a small power factor corresponds to inefficient use of energy. It should be apparent that if a load requires a fixed amount of real power, P, the source will be providing the smallest amount of current when the power factor is the greatest, that is, when $\cos\theta = 1$. If the power factor is less than unity, some additional current will be drawn from the source, lowering the efficiency of power transfer from the source to the load. However, it will be shown shortly that it is possible to correct the power factor of a load by adding an appropriate reactive component to the load itself.

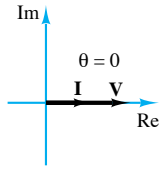
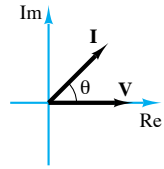
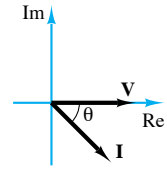
Since the reactive power, Q, is related to the reactive part of the load, its sign depends on whether the load reactance in inductive or capacitive. This leads to the following important statement:

> If the load has an inductive reactance, then θ is positive and the current *lags* (or *follows*) the voltage. Thus, when θ and Q are positive, the corresponding power factor is termed *lagging*. Conversely, a capacitive load will have a negative Q, and hence a negative θ. This corresponds to a *leading* power factor, meaning that the load current *leads* the load voltage.

Table 5.5.2 illustrates the concept and summarizes all of the important points so far. In the table, the phasor voltage V has a zero phase angle and the current phasor is referenced to the phase of V.

The distinction between leading and lagging power factors made in Table 5.5.2 is important because it corresponds to opposite signs of the reactive power: Q is positive if the load is inductive ($\theta > 0$) and the power factor is lagging: Q is negative if the load is capacitive and the power factor is leading ($\theta < 0$). It is therefore possible to improve the power factor of a load according to a procedure called **power factor correction** — that is, by placing a suitable reactance in parallel with the load so that the reactive power component generated by the additional reactance is of opposite sign to the original load reactive power. Most often the need is to improve the power factor of an inductive load because many common

TABLE 5.5.2 Important Facts Related to Complex Power

	Resistive load	Capacitive load	Inductive load
Ohm's law	$\mathbf{V}_L = Z_L\mathbf{I}_L$	$\mathbf{V}_L = Z_L\mathbf{I}_L$	$\mathbf{V}_L = Z_L\mathbf{I}_L$
Complex impedance	$Z_L = R_L$	$Z_L = R_L - jX_L$	$Z_L = R_L + jX_L$
Phase angle	$\theta_V - \theta_I = \theta = 0$	$\theta_V - \theta_I = \theta < 0$	$\theta_V - \theta_I = \theta > 0$
Complex plane sketch			
Explanation	The current is in phase with the voltage.	The current "leads" the voltage.	The current "lags" the voltage.
Power factor	Unity	Leading, < 1	Lagging, < 1
Reactive power	0	Negative	Positive

industrial loads consist of electric motors, which are predominantly inductive loads. This improvement may be accomplished by placing a capacitance in parallel with the load. The following example illustrates a typical power factor correction for an industrial load.

Example 5.5.1

For the circuit shown in Figure 5.5.3:

1. Calculate the complex power for the load.
2. Correct the power factor by adding a suitable reactance in parallel with the load.

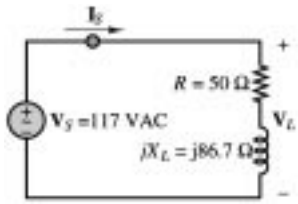


FIGURE 5.5.3 Circuit for Example 5.5.1.

Solution.

1. The circuit of Figure 5.5.4 is an inductive load. The total impedance is

$$Z = R + jX_L = 50 + j86.7\ \Omega = 100\angle 60^\circ$$

The power factor is then

$$\text{pf} = \cos\theta = \cos 60^\circ = 0.5\ (\text{lagging})$$

The current drawn from the source by the load is

$$\mathbf{I}_S = \frac{\mathbf{V}_S}{Z} = \frac{117\angle 0°}{100\angle 60°} = 1.17\angle -60°$$

and the average power is found to be

$$P = V_S I_S \cos\theta = 117 \times 1.17 \cos 60° = 68.4 \text{ W}$$

while the reactive power is

$$Q_L = V_S I_S \sin\theta = 117 \times 1.17 \sin 60° = 119 \text{ VAR}$$

Figure 5.5.4 shows the power triangle for the circuit.

2. The unity power factor for the circuit can be obtained by simply reducing the power factor angle θ to 0°. This can be accomplished by adding a capacitor to the circuit that requires –119 VAR of reactive power. The capacitive power and the inductive power will then cancel each other in the power triangle, resulting in a unity power factor, as shown in Figure 5.5.5.
   The value of capacitive reactance, X_c, required to cancel the reactive power due to the inductance is found most easily by observing that the total reactive power in the circuit must be the sum of the reactive power due to the capacitance and that due to the inductance. Observing that the capacitor sees the same voltage as the RL load, because of the parallel connection, we can write

$$X_C = \frac{V_S^2}{Q_C} = \frac{117^2}{119} 115\ \Omega$$

From the expression for the reactance, it is then possible to compute the value of the capacitor that will cancel the reactive power due to the inductor:

$$C = \frac{1}{\omega X_C} = \frac{1}{377 \times 115} = 23.1\ \mu\text{F}$$

The reactive component of power needed by the inductance is now balanced by the capacitance, and all the power delivered by the source is real power. The power factor is 1.

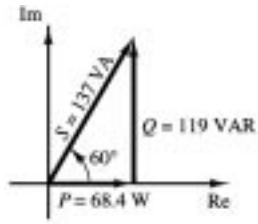


FIGURE 5.5.4 Power triangle for the circuit.

Example 5.5.2

The instrument used to measure power is called a *wattmeter*. The external part of a wattmeter consists of four connections and a metering mechanism that displays the amount of real power dissipated by a circuit. The external and internal appearances of a wattmeter are depicted in Figure 5.5.6. Inside the wattmeter are two coils: a current-sensing coil and a voltage-sensing coil. In this example, we assume for simplicity that the impedance of the current-sensing coil, C_I, is zero and the impedance of the voltage-sensing coil, C_V, is infinite. In practice, this will not necessarily be true; some correction mechanism will be required to account for the impedance of the sensing coils.

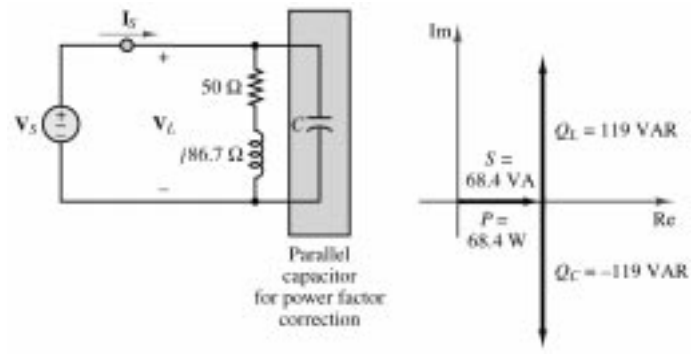


FIGURE 5.5.5 Power factor correction.

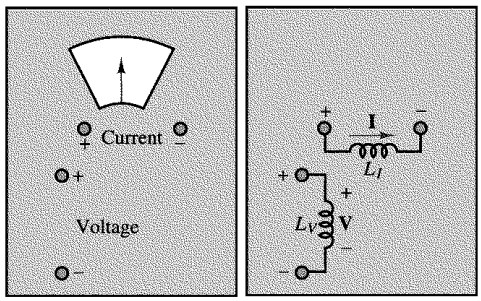


FIGURE 5.5.6 Wattmeter.

A wattmeter should be connected as shown in Figure 5.5.7, to provide both current and voltage measurements. We see that the current-sensing coil is placed in series with the load and the voltage-sensing coil is placed in parallel with the load. In this manner, the wattmeter is seeing the current through and the voltage across the load. Remember that the power dissipated by a circuit element is related to these two quantities. The wattmeter, then, is constructed to provide a readout of the product of the rms values of the load current and the voltage, which is the real power absorbed by the load: P = Re (S) = Re (**VI***).

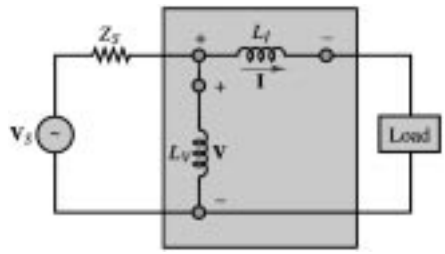


FIGURE 5.5.7 Circuit for Example 5.5.2.

1. For the circuit shown in Figure 5.5.8, show the connections of the wattmeter, and find the power dissipated by the load.
2. Show the connections that will determine the power dissipated by R_2. What should the meter read?

Solution.

1. To measure the power dissipated by the load, we must know the current through and the voltage across the entire load circuit. This means that the wattmeter must be connected as shown in Figure 5.5.9. The wattmeter should read:

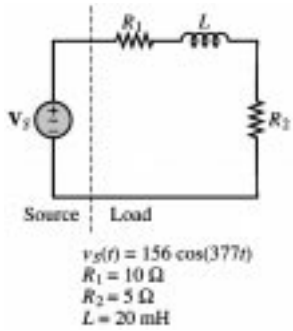


FIGURE 5.5.8 Circuit for Example 5.5.2.

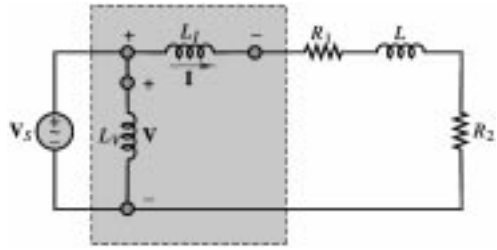


FIGURE 5.5.9 Circuit for Example 5.5.2.

$$P = \mathrm{Re}\left(\mathbf{V}_S \mathbf{I}^*\right)$$

$$= \mathrm{Re}\left\{\left(\frac{156}{\sqrt{2}}\angle 0°\right)\left(\frac{\frac{156}{\sqrt{2}}\angle 0°}{R_1 + R_2 + j\omega L}\right)^*\right\}$$

$$= \mathrm{Re}\left\{110\angle 0°\left(\frac{110\angle 0°}{15 + j7.54}\right)^*\right\}$$

$$\mathrm{Re}\left\{110\angle 0°\left(\frac{110\angle 0°}{16.79\angle 26.69°}\right)^*\right\} = \mathrm{Re}\left(\frac{110^2}{16.79\angle -26.69°}\right)$$

$$= \mathrm{Re}\left(720.67\angle 26.69°\right)$$

$$= 643.88 \text{ W}$$

2. To measure the power dissipated by R_2 alone, we must measure the current through R_2 and the voltage across R_2 *alone*. The connection is shown in Figure 5.5.10. The meter will read:

$$P = I^2 R_2$$

$$= \left(\frac{110}{\left(15^2 + 7.54^2\right)^{1/2}}\right)^2 \times 5$$

$$= \left(\frac{110^2}{\left(15^2 + 7.54^2\right)}\right)^2 \times 5 = 215 \text{ W}$$

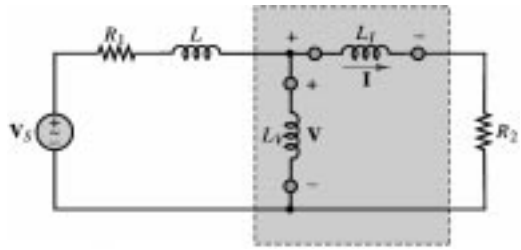


FIGURE 5.5.10 Circuit for Example 5.5.2.

The measurement and correction of the power factor for the load are an extremely important aspect of any engineering application in industry that requires the use of substantial quantities of electric power. In particular, industrial plants, construction sites, heavy machinery, and other heavy users of electric power must be aware of the power factor their loads present to the electric utility company. As was already observed, a low power factor results in greater current draw from the electric utility and in greater line losses. Thus, computations related to the power factor of complex loads are of great practical utility to any practicing engineer.

Transformers

AC circuits are very commonly connected to each other by means of **transformers**. A transformer is a device that couples two AC circuits magnetically rather than through any direct conductive connection and permits a "transformation" of the voltage and current between one circuit and the other (for example, by matching a high-voltage, low-current AC output to a circuit requiring a low-voltage, high-current source). Transformers play a major role in electric power engineering and are a necessary part of the electric power distribution network. The objective of this section is to introduce the ideal transformer and the concepts of impedance reflection and impedance matching.

The Ideal Transformer

The ideal transformer consists of two coils that are coupled to each other by some magnetic medium. There is no electrical connection between the coils. The coil on the input side is termed the **primary**, and that on the output side the **secondary**. The primary coil is wound so that it has n_1 turns, while the secondary has n_2 turns. We define the **turns ratio** N as

$$N = \frac{n_2}{n_1} \tag{5.5.14}$$

Figure 5.5.11 illustrates the convention by which voltages and currents are usually assigned at a transformer. The dots in Figure 5.5.11 are related so the polarity of the coil voltage: coil terminals marked with a dot have the same polarity.

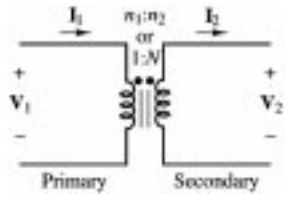


FIGURE 5.5.11 Ideal transformer.

Since an ideal inductor acts as a short circuit in the presence of DC currents, transformers do not perform any useful function when the primary voltage is DC. However, when a time-varying current flows in the primary winding, a corresponding time-varying voltage is generated in the secondary because

of the magnetic coupling between the two coils. This behavior is due to Faraday's law, as will be explained in Section 5.12. The relationship between primary and secondary current in an ideal transformer is very simply stated as follows:

$$\mathbf{V}_2 = N\mathbf{V}_1$$
$$\mathbf{I}_2 = \frac{\mathbf{I}_1}{N} \tag{5.5.15}$$

An ideal transformer multiples a sinusoidal input voltage by a factor of N and divides a sinusoidal input current by a factor of N. If N is greater than 1, the output voltage is greater than the input voltage and the transformer is called a **step-up transformer**. If N is less than 1, then the transformer is called a **step-down transformer**, since V_2 is now smaller than V_1.

An ideal transformer can be used in either direction (i.e., either of its coils may be viewed as the input side or primary). Finally, a transformer with $N = 1$ is called an **isolation transformer** and may perform a very useful function if one needs to electrically isolate two circuits from each other; note that any DC currents at the primary will not appear at the secondary coil. An important property of ideal transformers is conservation of power; one can easily verify that an ideal transformer conserves power, since

$$S_1 = \mathbf{I}_1^* \mathbf{V}_1 = N\mathbf{I}_2^* \frac{\mathbf{V}_2}{N} = \mathbf{I}_2^* \mathbf{V}_2 = S_2 \tag{5.5.16}$$

That is, the power on the primary side equals that on the secondary.

In many practical circuits, the secondary is tapped at two different points, giving rise to two separate output circuits, as shown in Figure 5.5.12. The most common configuration is the **center-tapped transformer**, which splits the secondary voltage into two equal voltages of half the original amplitude. The most common occurrence of this type of transformer is found at the entry of a power line into a household, where the 240-VAC line is split into two 120-VAC lines (this may help explain why both 240- and 120-VAC power are present in your house).

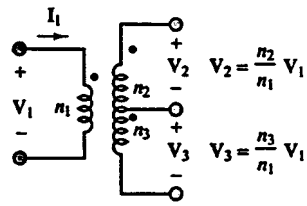


FIGURE 5.5.12 Center-tapped transformer.

Three-Phase Power

The material presented so far in this chapter has dealt exclusively with **single-phase AC power**, that is, with single sinusoidal sources. In fact, most of the AC power used today is generated and distributed as **three-phase power**, by means of an arrangement in which three sinusoidal voltages are generated out of phase with each other. The primary reason is efficiency: the weight of the conductors and other components in a three-phase system is much lower than in a single-phase system delivering the same amount of power. Further, while the power produced by a single-phase system has a pulsating nature, a three-phase system can deliver a steady, constant supply of power. A three-phase generator producing three **balanced voltages** — that is, voltages of equal amplitude and frequency displaced in phase by 120° — has the property of delivering constant instantaneous power.

Another important advantage of three-phase power is that, as will be explained in Section 5.12, three-phase motors have a nonzero starting torque, unlike their single-phase counterpart. The change to three-phase AC power system from the early DC system proposed by Edison was therefore due to a number of reasons: the efficiency resulting from transforming voltages up and down to minimize transmission losses over long distances; the ability to deliver constant power (an ability not shared by single- and two-phase AC systems); a more efficient use of conductors; and the ability to provide starting torque for industrial motors.

To begin the discussion of three-phase power, consider a three-phase source connected in the **wye** (or **Y**) **configuration**, as shown in Figure 5.5.13. Each of the three voltages is 120° out of phase with the others, so that, using phasor notation, we may write:

$$\begin{aligned} V_{an} &= V^f_{an}\angle 0° \\ V_{bn} &= V^f_{bn}\angle -120° \\ V_{cn} &= V^f_{cn}\angle -240° = V^f_{cn}\angle 120° \end{aligned} \tag{5.5.17}$$

where the quantities V^f_{an}, V^f_{bn}, and V^f_{cn} are rms values and are equal to each other. To simplify the notation, it will be assumed from here on that

$$V^f_{an} = V^f_{bn} = V^f_{cn} = V^f \tag{5.5.18}$$

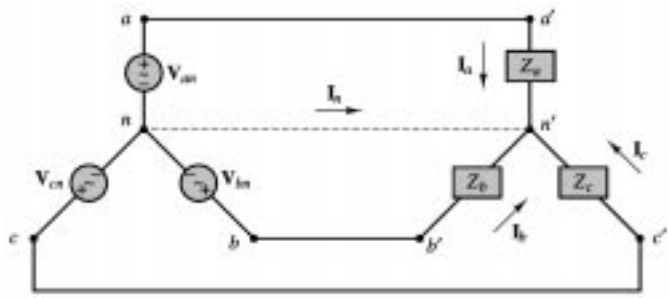


FIGURE 5.5.13 Balanced three-phase AC circuit.

Section 5.12 will discuss how three-phase AC electric generators may be constructed to provide such balanced voltages. In the circuit of Figure 5.5.13, the resistive loads are also wye-connected and balanced (i.e., equal). The three AC sources are all connected together at a node called the *neutral node*, denoted by *n*. The voltages V^f_{an}, V^f_{bn}, and V^f_{cn} are called the *phase voltages* and form a balanced set in the sense that

$$\mathbf{V}_{an} + \mathbf{V}_{bn} + \mathbf{V}_{cn} = 0 \tag{5.5.19}$$

This last statement is easily verified by sketching the phasor diagram. The sequence of phasor voltages shown in Figure 5.5.14 is usually referred to as the **positive** (or *abc*) **sequence**.

Consider now the "lines" connecting each source to the load and observe that it is possible to also define **line voltages** (also called *line-to-line voltages*) by considering the voltages between the lines *aa*′ and *bb*′, *aa*′ and *cc*′, and *bb*′ and *cc*′. Since the line voltage, say, between *aa*′ and *bb*′ is given by

$$\mathbf{V}_{ab} = \mathbf{V}_{an} + \mathbf{V}_{nb} = \mathbf{V}_{an} - \mathbf{V}_{bn} \tag{5.5.20}$$

the line voltages may be computed relative to the phase voltages as follows:

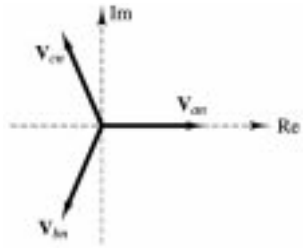


FIGURE 5.5.14 Positive, or *abc*, sequence for balanced three-phase voltages.

$$\mathbf{V}_{ab} = \mathcal{V}\angle 0° - \mathcal{V}\angle -120° = \sqrt{3}\mathcal{V}\angle 30°$$
$$\mathbf{V}_{bc} = \mathcal{V}\angle -120° - \mathcal{V}\angle 120° = \sqrt{3}\mathcal{V}\angle -90° \qquad (5.5.21)$$
$$\mathbf{V}_{ca} = \mathcal{V}\angle 120° - \mathcal{V}\angle 0° = \sqrt{3}\mathcal{V}\angle 150°$$

It can be seen, then, that the magnitude of the line voltages is equal to $\sqrt{3}$ times the magnitude of the phase voltages. It is instructive, at least once, to point out that the circuit of Figure 5.5.13 can be redrawn to have the appearance of the circuit of Figure 5.5.15.

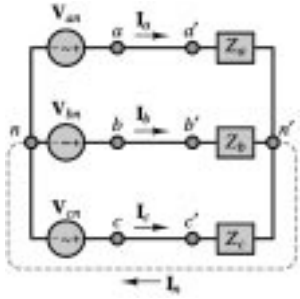


FIGURE 5.5.15 Balanced three-phase AC circuit (redrawn).

One of the important features of a balanced three-phase system is that it does not require a fourth wire (the neutral connection), since the current $\mathbf{I}_n$ is identically zero (for balanced load $Z_a = Z_b = Z_c = Z$). Another, more important characteristic of a balanced three-phase power system is that the total power delivered to the balanced load by the three-phase generator is constant.

$$p(t) = p_a(t) + p_b(t) + p_c(t) = \frac{3\mathcal{V}^2}{R}$$

It is also possible to connect the three AC sources in a three-phase system in a so-called **delta** (or Δ) **connection**, although in practice this configuration is rarely used. Figure 5.5.16 depicts a set of three delta-connected generators.

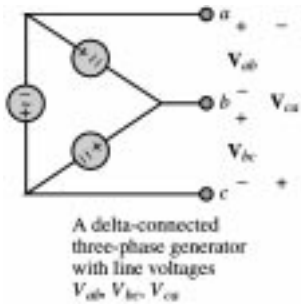


FIGURE 5.5.16 Delta-connected generators.

Balanced Wye Loads

Consider again the circuit of Figure 5.5.13, where now the balanced load consists of the three complex impedances

$$Z_a = Z_b = Z_c = Z_y = |Z_y|\angle\theta \tag{5.5.22}$$

From the diagram of Figure 5.5.13, it can be verified that each impedance sees the corresponding phase voltage across itself; thus, since the currents $\mathbf{I}_a$, $\mathbf{I}_b$, and $\mathbf{I}_c$ have the same rms value, I, the phase angles of the currents will differ by ±120°. It is therefore possible to compute the power for each phase by considering the phase voltage (equal to the load voltage) for each impedance, and the associated line current. Let us denote the complex power for each phase by S:

$$S = \mathbf{V} \cdot \mathbf{I}^* \tag{5.5.23}$$

so that

$$\begin{aligned} S &= P + jQ \\ &= VI\cos\theta + jVI\sin\theta \end{aligned} \tag{5.5.24}$$

where V and I denote, once again, the rms values of each phase voltage and line current. Consequently, the total real power delivered to the balanced wye load is $3P$, and the total reactive power is $3Q$. Thus, the total complex power, S_T, is given by

$$\begin{aligned} S_T &= P_T + jQ_T = 3P + j3Q \\ &= \sqrt{(3P)^2 + (3Q)^2}\,\angle\theta \end{aligned} \tag{5.5.25}$$

and the apparent power is

$$\begin{aligned} |S_T| &= 3\sqrt{(VI)^2\cos^2\theta + (VI)^2\sin^2\theta} \\ &= 3VI \end{aligned}$$

and the total real and reactive power may be expressed in terms of the apparent power:

$$\begin{aligned} P_T &= |S_T|\cos\theta \\ Q_T &= |S_T|\sin\theta \end{aligned} \tag{5.5.26}$$

Balanced Delta Loads

In addition to a wye connection, it is also possible to connect a balanced load in the delta configuration. A wye-connected generator and a delta-connected load are shown in Figure 5.5.17.

It should be noted immediately that now the corresponding line voltage (not phase voltage) appears across each impedance. For example, the voltage across $Z_{c'a'}$ is $\mathbf{V}_{ca}$. Thus, the three load currents are given by the following expressions:

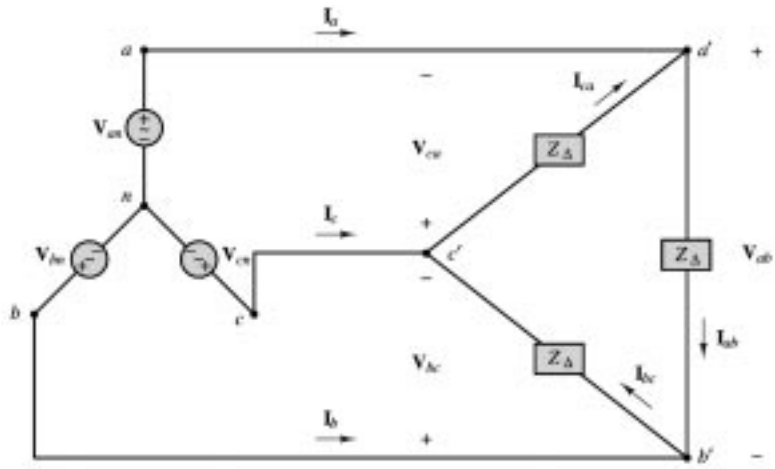


FIGURE 5.5.17 Balanced wye generators with balanced delta load.

$$\mathbf{I}_{ab} = \frac{\mathbf{V}_{ab}}{Z_\Delta} = \frac{\sqrt{3}V\angle 30^\circ}{|Z_\Delta|\angle\theta}$$

$$\mathbf{I}_{bc} = \frac{\mathbf{V}_{bc}}{Z_\Delta} = \frac{\sqrt{3}V\angle -90^\circ}{|Z_\Delta|\angle\theta}$$

$$\mathbf{I}_{ca} = \frac{\mathbf{V}_{ca}}{Z_\Delta} = \frac{\sqrt{3}V\angle 150^\circ}{|Z_\Delta|\angle\theta}$$

One can readily verify that the two currents $(\mathbf{I}_a)_\Delta$ and $(\mathbf{I}_a)_y$ will be equal if the magnitude of the delta-connected impedance is three times larger than Z_y:

$$Z_\Delta = 3Z_y \tag{5.5.27}$$

This result also implies that a delta load will necessarily draw three times as much current (and therefore absorb three times as much power) as we wye load with the same branch impedance.

Generation and Distribution of AC Power

We now conclude the discussion of power systems with a brief description of the various elements of a power system. Electric power originates from a variety of sources; in Section 5.12, electric generators will be introduced as a means of producing electric power from a variety of energy-conversion processes. In general, electric power may be obtained from hydroelectric, thermoelectric, geothermal, wind, solar, and nuclear sources. The choice of a given source is typically dictated by the power requirement for the given application and by economic and environmental factors. In this section, the structure of an AC power network, from the power-generating station to the residential circuits discussed in the previous section, is briefly outlined.

A typical generator will produce electric power at 18 kV, as shown in the diagram of Figure 5.5.18. To minimize losses along the conductors, the output of the generators is processed through a step-up transformer to achieve line voltages of hundreds of kilovolts. Without this transformation, the majority of the power generated would be lost in the **transmission lines** that carry the electric current from the power station.

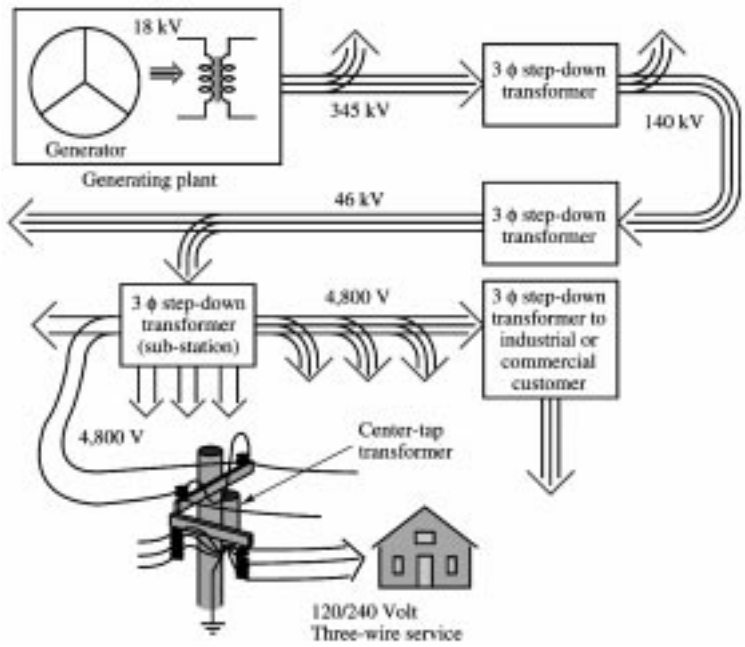


FIGURE 5.5.18 Structure of an AC power distribution network.

The local electric company operates a power-generating plant that is capable of supplying several hundred megavolt-amperes (MVA) on a three-phase basis. For this reason, the power company uses a three-phase step-up transformer at the generation plant to increase the line voltage to around 345 kV. One can immediately see that at the rated power of the generator (in MVA) there will be a significant reduction of current beyond the step-up transformer.

Beyond the generation plant, an electric power network distributes energy to several **substations**. This network is usually referred to as the **power grid**. At the substations, the voltage is stepped down to a lower level (10 to 150 kV, typically). Some very large loads (for example, an industrial plant) may be served directly from the power grid, although most loads are supplied by individual substations in the power grid. At the local substations (one of which you may have seen in your own neighborhood), the voltage is stepped down further by a three-phase step-down transformer to 4800 V. These substations distribute the energy to residential and industrial customers. To further reduce the line voltage to levels that are safe for residential use, step-down transformers are mounted on utility poles. These drop the voltage to the 120/240-V three-wire single-phase residential service discussed in the previous section. Industrial and commercial customers receive 460- and/or 208-V three-phase service.

5.6 Frequency Response, Filters, and Transient Analysis

The aim of the present section is twofold: first, to exploit AC circuit analysis methods to study the frequency response of electric circuits; and second, to continue the discussion of dynamic circuit equations for the purpose of analyzing the transient response of electrical circuits.

The **sinusoidal frequency response** (or, simply, **frequency response**) of a circuit provides a measure of how the circuit responds to sinusoidal inputs of arbitrary frequency. In other words, given the input signal amplitude, phase, and frequency, knowledge of the frequency response of a circuit permits the computation of the output signal.

The frequency response of a circuit is a measure of the variation of a load-related voltage or current as a function of the amplitude, phase, and frequency of the excitation signal.

To express the frequency response of a circuit in terms of variation in output voltage as a function of source voltage, we use the general formula

$$H_V(j\omega) = \frac{\mathbf{V}_L(j\omega)}{\mathbf{V}_S(j\omega)} \tag{5.6.1}$$

One method that allows for representation of the load voltage as a function of the source voltage (this is, in effect, what the frequency response of a circuit implies) is to describe the source and attached circuit by means of the Thévenin equivalent circuit. The frequency response of the circuit shown in Figure 5.6.1 is given by the expression

$$\frac{\mathbf{V}_L}{\mathbf{V}_S}(j\omega) = H_V(j\omega) = \frac{Z_L Z_2}{Z_L(Z_S + Z_1 + Z_2) + (Z_S + Z_1)Z_2} \tag{5.6.2}$$

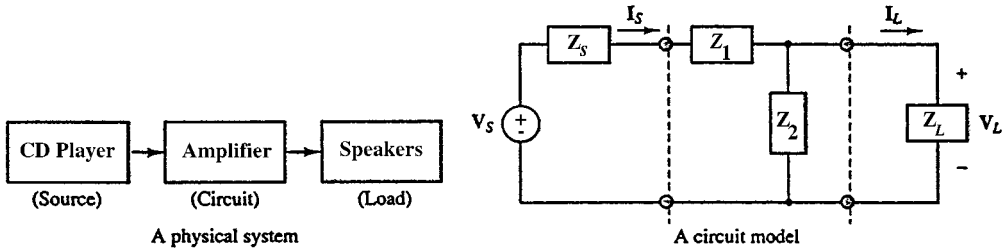


FIGURE 5.6.1 A circuit model.

The expression for $H_V(j\omega)$ could be evaluated for any given $\mathbf{V}_S(j\omega)$ (i.e., for any given source signal amplitude, phase, and frequency) to determine what the resultant load voltage would be. Note that $H_V(j\omega)$ is a complex quantity (dimensionless, because it is the ratio of two voltages), and that it therefore follows that $\mathbf{V}_L(j\omega)$ is a phase-shifted and amplitude-scaled version of $\mathbf{V}_S(j\omega)$:

$$\mathbf{V}_L(j\omega) = H_V(j\omega) \cdot \mathbf{V}_S(j\omega) \tag{5.6.3}$$

$$V_L e^{j\phi_L} = |H_V| e^{j\phi_H} \cdot V_S e^{j\phi_S} \tag{5.6.4}$$

or

$$V_L e^{j\phi_L} = |H_V| \mathbf{V}_S e^{j(\phi_H + \phi_S)} \tag{5.6.5}$$

where

$$V_L = |H_V| \cdot V_S$$

and

$$\phi_L = \phi_H + \phi_S \tag{5.6.6}$$

The effect of inserting a linear circuit between a source and a load is best understood by considering that, at any given frequency, ω, the load voltage is a sinusoid at the same frequency as the source voltage, with amplitude given by $V_L = |H_V| \cdot V_S$ and phase equal to $\phi_L = \phi_H + \phi_S$, where $|H_V|$ is the magnitude of the frequency response and ϕ_H its phase angle. Both $|H_V|$ and ϕ_H are functions of frequency.

The importance and usefulness of the frequency response concept lies in its ability to summarize the response of a circuit in a single function of frequency, $H(j\omega)$, which can predict the load voltage or current at any frequency, given the input. Note that the frequency response of a circuit can be defined in four different ways:

$$H_V(j\omega) = \frac{\mathbf{V}_L(j\omega)}{\mathbf{V}_S(j\omega)} \qquad H_I(j\omega) = \frac{\mathbf{I}_L(j\omega)}{\mathbf{I}_S(j\omega)}$$
$$H_Z(j\omega) = \frac{\mathbf{V}_L(j\omega)}{\mathbf{I}_S(j\omega)} \qquad H_Y(j\omega) = \frac{\mathbf{I}_L(j\omega)}{\mathbf{V}_S(j\omega)} \tag{5.6.7}$$

If $H_V(j\omega)$ and $H_I(j\omega)$ are known, one can directly derive the other two expressions:

$$H_Z(j\omega) = \frac{\mathbf{V}_L(j\omega)}{\mathbf{I}_S(j\omega)} = Z_L(j\omega)\frac{\mathbf{I}_L(j\omega)}{\mathbf{I}_S(j\omega)} = Z_L(j\omega)H_I(j\omega) \tag{5.6.8}$$

$$H_Y(j\omega) = \frac{\mathbf{I}_L(j\omega)}{\mathbf{V}_S(j\omega)} = \frac{1}{Z_L(j\omega)}\frac{\mathbf{V}_L(j\omega)}{\mathbf{V}_S(j\omega)} = \frac{1}{Z_L(j\omega)}H_V(j\omega) \tag{5.6.9}$$

With these definitions in hand, it is now possible to introduce one of the central concepts of electrical circuit analysis: **filters.** The concept of filtering an electrical signal will be discussed in the next section.

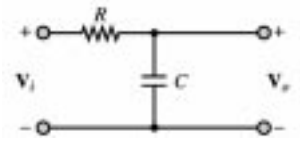
Filters

There are a host of practical, everyday applications that involve filters of one kind or another. Just to mention two, filtration systems are used to eliminate impurities from drinking water, and sunglasses are used to filter out eye-damaging ultraviolet radiation and to reduce the intensity of sunlight reaching the eyes. An analogous concept applies to electrical circuits: it is possible to *attenuate* (i.e., reduce in amplitude) or altogether eliminate signals of unwanted frequencies, such as those that may be caused by electrical noise or other forms of interference. This section will be devoted to the analysis of electrical filters.

Low-Pass Filters

Figure 5.6.2 depicts a simple **RC filter** and denotes its input and output voltages by $\mathbf{V}_i$ and $\mathbf{V}_o$. The frequency response for the filter may be obtained by considering the function.

$$H(j\omega) = \frac{\mathbf{V}_o}{\mathbf{V}_i}(j\omega) \tag{5.6.10}$$



RC low-pass filter. The circuit preserves lower frequencies while attenuating the frequencies above the cutoff frequency, $\omega_0 = 1/RC$. The voltages V_i and V_0 are the filter input and output voltages, respectively.

FIGURE 5.6.2 A simple RC filter.

and noting that the output voltage may be expressed as a function of the input voltage by means of a voltage divider, as follows:

$$\mathbf{V}_o(j\omega) = \mathbf{V}_i(j\omega)\frac{1/j\omega C}{R + 1/j\omega C} = \mathbf{V}_i(j\omega)\frac{1}{1 + j\omega RC} \tag{5.6.11}$$

$$H(j\omega) = \frac{\mathbf{V}_o}{\mathbf{V}_i}(j\omega) = \frac{1}{1 + j\omega CR}$$

or

$$H(j\omega) = |H(j\omega)|e^{j\phi_H(j\omega)} \tag{5.6.12}$$

with

$$|H(j\omega)| = \frac{1}{\sqrt{1 + (\omega CR)^2}} = \frac{1}{\sqrt{1 + (\omega/\omega_0)^2}} \tag{5.6.13}$$

and

$$\phi_H(j\omega) = -\arctan(\omega CR) = -\arctan\left(\frac{\omega}{\omega_0}\right) \tag{5.6.14}$$

with

$$\omega_0 = \frac{1}{RC} \tag{5.6.15}$$

The simplest way to envision the effect of the filter is to think of the phasor voltage $\mathbf{V}_i = V_i e^{j\phi_i}$ scaled by a factor of $|\mathbf{H}|$ and shifted by a phase angle ϕ_H by the filter *at each frequency*, so that the resultant output is given by the phasor $V_o e^{j\phi_o}$, with

$$V_o = |H| \cdot V_i$$
$$\phi_o = \phi_H + \phi_i \tag{5.6.16}$$

and where $|H|$ and ϕ_H are functions of frequency. The frequency ω_0 is called the **cutoff frequency** of the filter and, as will presently be shown, gives an indication of the filtering characteristics of the circuit.

It is customary to represent $H(j\omega)$ in two separate plots, representing $|H|$ and ϕ_H as functions of ω. These are shown in Figure 5.6.3 in normalized form — that is, with $|H|$ and ϕ_H plotted vs. ω/ω_0, corresponding to a cutoff frequency $\omega_0 = 1$ rad/sec. Note that, in the plot, the frequency axis has been scaled logarithmically. This is a common practice in electrical engineering, because it allows viewing a very broad range of frequencies on the same plot without excessively compressing the low-frequency end of the plot. The frequency response plots of Figure 5.6.3 are commonly employed to describe the frequency response of a circuit, since they can provide a clear idea at a glance of the effect of a filter on an excitation signal. For example, the *RC* filter of Figure 5.6.2 has the property of "passing" signals at low frequencies ($\omega \ll 1/RC$) and of filtering out signals at high frequencies ($\omega \gg 1/RC$). This type of filter is called a **low-pass filter**. The cutoff frequency $\omega = 1/RC$ has a special significance in that it represents — approximately — the point where the filter begins to filter out the higher-frequency signals. The value of $H(j\omega)$ at the cutoff frequency is $1/\sqrt{2} = 0.707$. Note how the cutoff frequency depends exclusively on the values of R and C. Therefore, one can adjust the filter response as desired simply by selecting appropriate values for C and R, and therefore choose the desired filtering characteristics.

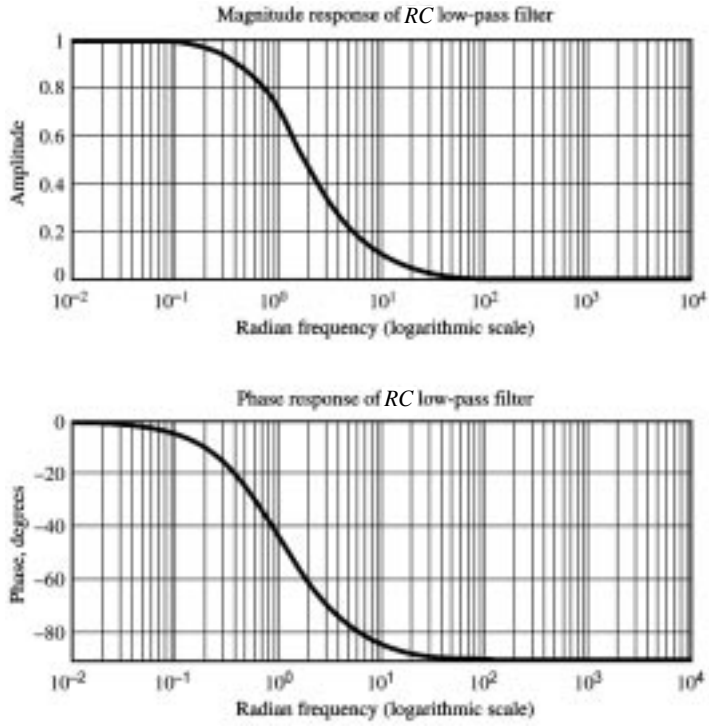


FIGURE 5.6.3 Magnitude and phase response plots for *RC* filter.

Example 5.6.1 Wheatstone Bridge Filter

The Wheatstone bridge circuit is used in a number of instrumentation applications, including the measurement of force (see Example 5.6.2, describing the strain gauge bridge). Figure 5.6.4 depicts the appearance of the bridge circuit. When undesired noise and interference are present in a measurement,

it is often appropriate to use a low-pass filter to reduce the effect of the noise. The capacitor that is connected to the output terminals of the bridge in Figure 5.6.4 constitutes an effective and simple low-pass filter, in conjunction with the bridge resistance. Assume that the average resistance of each leg of the bridge is 350 Ω (a standard value for strain gauges) and that we desire to measure a sinusoidal force at a frequency of 30 Hz. From prior measurements, it has been determined that a filter with a cutoff frequency of 300 Hz is sufficient to reduce the effects of noise. Choose a capacitor that matches this filtering requirement.

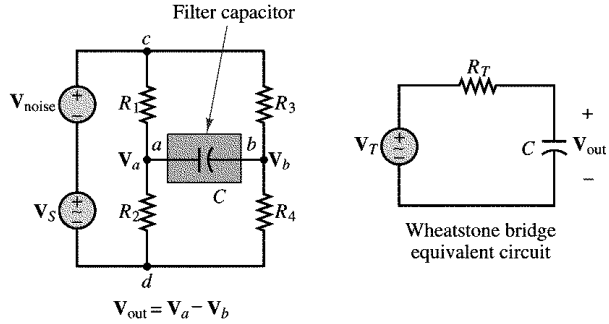


FIGURE 5.6.4 Wheatstone bridge with equivalent circuit and simple capacitive filter.

Solution. By evaluating the Thévenin equivalent circuit for the Wheatstone bridge, calculating the desired value for the filter capacitor becomes relatively simple, as illustrated at the bottom of Figure 5.6.4. The Thévenin resistance for the bridge circuit may be computed by short-circuiting the two voltage sources and removing the capacitor placed across the load terminals:

$$R_T = R_1 \| R_2 + R_3 \| R_4 = 350 \| 350 + 350 \| 350 = 350\ \Omega$$

Since the required cutoff frequency is 300 Hz, the capacitor value can be computed from the expression

$$\omega_0 = \frac{1}{R_T C} = 2\pi \times 300$$

or

$$C = \frac{1}{R_T \omega_0} = \frac{1}{350 \times 2\pi \times 300} = 1.51\ \mu\text{F}$$

The frequency response of the bridge circuit is of the form

$$\frac{\mathbf{V}_{out}}{\mathbf{V}_T}(j\omega) = \frac{1}{1 + j\omega C R_T}$$

This response can be evaluated at the frequency of 30 Hz to verify that the attenuation and phase shift at the desired signal frequency are minimal:

$$\frac{\mathbf{V}_{out}}{\mathbf{V}_T}(j\omega = j2\pi \times 30) = \frac{1}{1 + j2\pi \times 30 \times 1.51 \times 10^{-6} \times 350}$$

$$= 0.9951\angle -5.7°$$

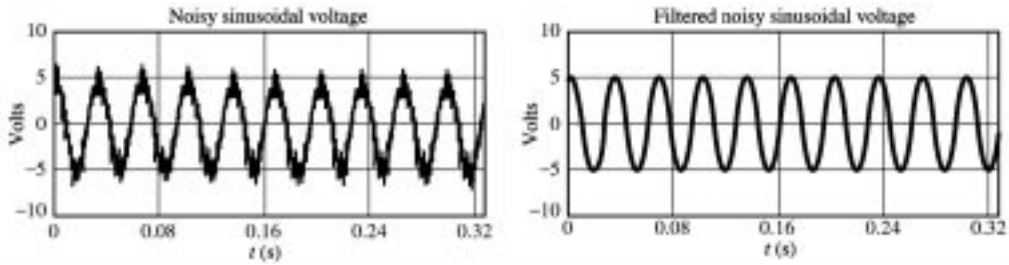


FIGURE 5.6.5 Unfiltered and filtered bridge output.

Figure 5.6.5 depicts the appearance of a 30-Hz sinusoidal signal before and after the addition of the capacitor to the circuit.

High-Pass Filters

Just as you can construct a simple filter that preserves low frequencies and attenuates higher frequencies, you can easily construct a **high-pass filter** that passes mainly those frequencies *above a certain cutoff frequency*. The analysis of a simple high-pass filter can be conducted by analogy with the preceding discussion of the low-pass filter. Consider the circuit shown in Figure 5.6.6.

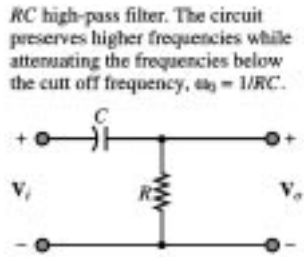


FIGURE 5.6.6 High-pass filter.

The frequency response of the filter is

$$\frac{\mathbf{V}_o}{\mathbf{V}_i}(j\omega) = \frac{j\omega CR}{1 + j\omega CR}$$

or

$$H(j\omega) = |H|e^{j\phi_H}$$

with

$$H(j\omega) = \frac{\omega CR}{\sqrt{1 + (\omega CR)^2}}$$

$$\phi_H(j\omega) = 90° - \arctan(\omega CR)$$

Amplitude-and-phase response curves for the high-pass filter are shown in Figure 5.6.7. These plots have been normalized to have the filter cutoff frequency $\omega_0 = 1$ rad/sec. Note that, once again, it is possible to define a cutoff frequency at $\omega_0 = 1/RC$ in the same way as was done for the low-pass filter.

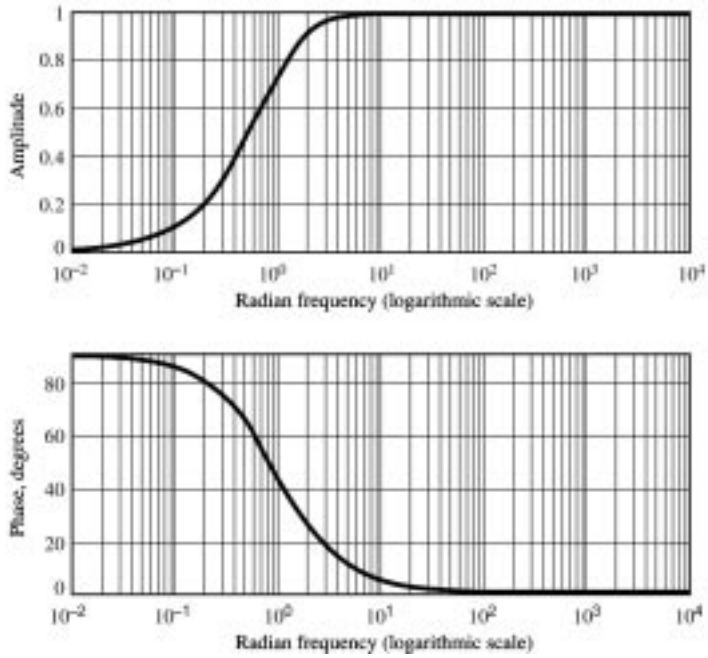


FIGURE 5.6.7 Frequency response of a high-pass filter.

Band-Pass Filters

Building on the principles developed in the preceding sections, we can also construct a circuit that acts as a **band-pass filter**, passing mainly those frequencies *within a certain frequency range*. The analysis of a simple *second-order* band-pass filter (i.e., a filter with two energy-storage elements) can be conducted by analogy with the preceding discussions of the low-pass and high-pass filters. Consider the circuit shown in Figure 5.6.8, and the related frequency response function for the filter $H(j\omega) = (\mathbf{V}_o/\mathbf{V}_i)(j\omega)$.

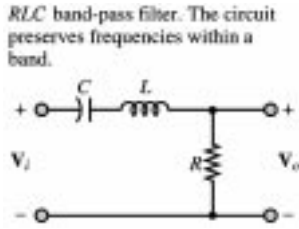


FIGURE 5.6.8 *RLC* band-pass filter.

We may write the frequency response of the filter as

$$\frac{\mathbf{V}_o}{\mathbf{V}_i}(j\omega) = \frac{j\omega CR}{1 + j\omega CR + (j\omega)^2 LC} \tag{5.6.17}$$

Equation (5.6.17) can often be factored into the following form:

$$\frac{\mathbf{V}_o}{\mathbf{V}_i}(j\omega) = \frac{jA\omega}{(j\omega/\omega_1 + 1)(j\omega/\omega_2 + 1)} \tag{5.6.18}$$

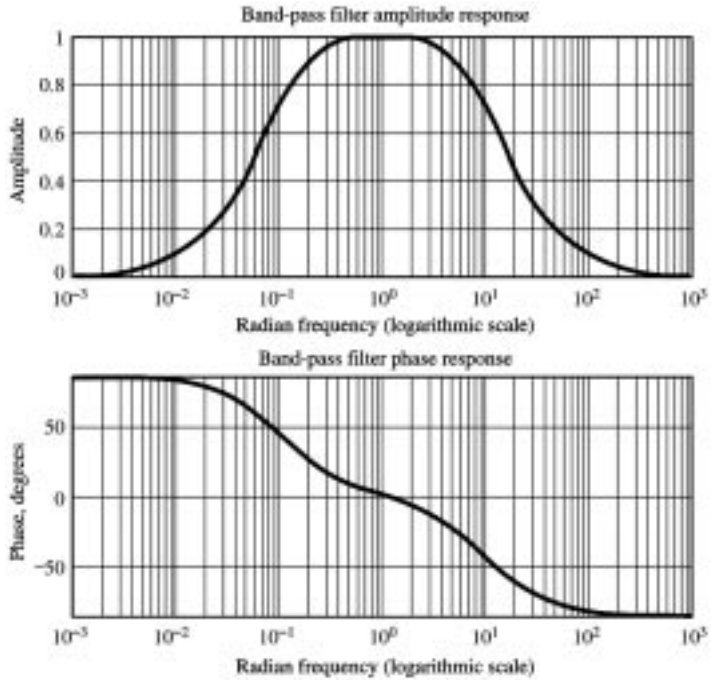


FIGURE 5.6.9 Frequency response of *RLC* band-pass filter.

where ω_1 and ω_2 are the two frequencies that determine the **pass-band** (or **band-width**) of the filter — that is, the frequency over which the filter "passes" the input signal. The magnitude and phase plots for the frequency response of the band-pass filter of Figure 5.6.8 are shown in Figure 5.6.9. These plots have been normalized to have the filter pass-band centered at the frequency $\omega = 1$ rad/sec.

The expression for the frequency response of a second-order band-pass filter (Equation (5.6.17)) can also be rearranged to illustrate two important features of this circuit: the **quality factor**, Q, and the **resonant frequency**, ω_0. Let

$$\omega_0 = \frac{1}{\sqrt{LC}} \quad \text{and} \quad Q = \omega_0 CR = \frac{R}{\omega_0 L} \tag{5.6.19}$$

Then we can write $\omega CR = \omega_0 CR(\omega/\omega_0) = Q(\omega/\omega_0)$ and rearrange Equation 5.6.18 as follows:

$$\frac{\mathbf{V}_o}{\mathbf{V}_i}(j\omega) = \frac{jQ\dfrac{\omega}{\omega_0}}{\left(\dfrac{j\omega}{\omega_0}\right)^2 + jQ\dfrac{\omega}{\omega_0} + 1} \tag{5.6.20}$$

In Equation (5.6.20), the resonant frequency, ω_0, corresponds to the center frequency of the filter, while Q, the quality factor, indicates the *sharpness* of the resonance, that is, how narrow or wide the shape of the pass-band of the filter is. The width of the pass-band is also referred to as the *bandwidth*, and it can easily be shown that the bandwidth of the filter is given by the expression

$$B = \frac{\omega_0}{Q} \tag{5.6.21}$$

Thus, a high-Q filter has a narrow bandwidth, while a low-Q filter has a large bandwidth and is therefore less selective. The quality factor of a filter provides an immediate indication of the nature of the filter.

Transient Analysis

In analyzing the frequency response of AC circuits earlier in this chapter, we made the assumption that the particular form of the voltages and currents in the circuit was sinusoidal.

There are many signals, however, for which the steady-state sinusoidal representation is not adequate. In particular, the sinusoidal, or AC method of analysis does not apply to **transient signals**, that is, voltages and currents that vary as a function of time as a consequence of a sudden change in the input. Figure 5.6.10 illustrates the appearance of the voltage across some hypothetical load when a DC and an AC source, respectively, are abruptly switched on at time $t = 2$ sec. The waveforms in Figure 5.6.10 can be subdivided into three regions: a *steady-state region* for $0 \le t \le 0.2$ sec; a *transient region*, for 0.2 sec $\le t \le 2$ sec (approximately); and a new steady-state region for $t > 2$ sec, where the waveform reaches a new steady-state DC or sinusoidal condition. The objective of **transient analysis** is to describe the behavior of a voltage or current during the transition that takes place between two different steady-state conditions.

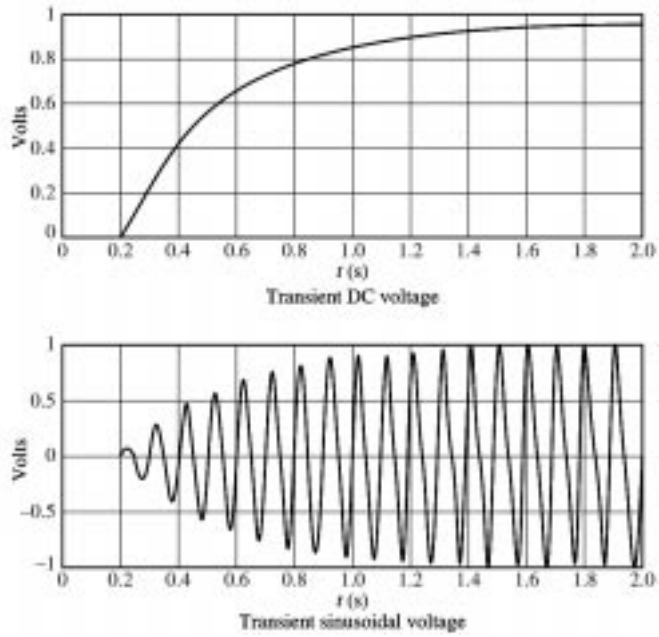


FIGURE 5.6.10 Examples of transient response.

You already know how to analyze circuits in a sinusoidal steady state by means of phasors. The material presented in the remainder of this chapter will provide the tools necessary to describe the *transient response* of circuits containing resistors, inductors, and capacitors. A general example of the type of circuit that will be discussed in this section is shown in Figure 5.6.11. The switch indicates that we turn the battery power on at time $t = 0$. Transient behavior may be expected whenever a source of electrical energy is switched on or off, whether it be AC or DC. A typical example of the transient response to a switched DC voltage would be what occurs when the ignition circuits in an automobile are turned on, so that a 12-V battery is suddenly connected to a large number of electrical circuits. The degree of complexity in transient analysis depends on the number of energy-stored elements in the circuit; the analysis can become quite involved for high-order circuits. In this chapter, we shall analyze

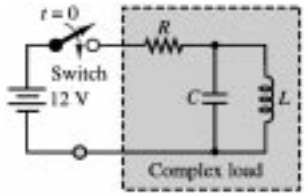


FIGURE 5.6.11 Circuit with switched DC excitation.

only first- and second-order circuits — that is, circuits containing one or two energy-storage elements, respectively. In electrical engineering practice, we would typically resort to computer-aided analysis for higher-order circuits.

A convenient starting point in approaching the transient response of electrical circuits is to consider the general model shown in Figure 5.6.12, where the circuits in the box consist of a combination of resistors connected to a *single energy-storage element*, either an inductor or a capacitor. Regardless of how many resistors the circuit contains, it is a **first-order circuit**. In general, the response of a first-order circuit to a switched DC source will appear in one of the two forms shown in Figure 5.6.13 which represent, in order, **a decaying exponential** and a **rising exponential** waveform. In the next sections, we will systematically analyze these responses by recognizing that they are exponential in nature and can be computed very easily once we have the proper form of the differential equation describing the circuit.

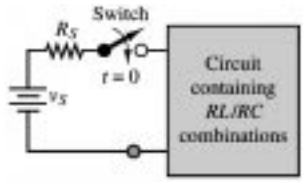


FIGURE 5.6.12 A general model of the transient analysis problem.

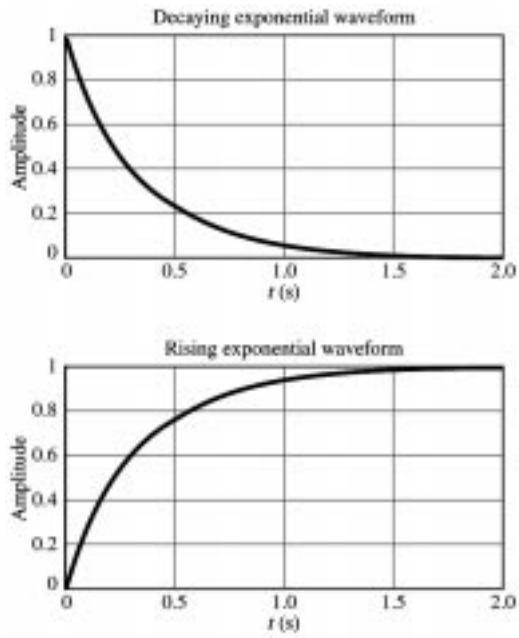


FIGURE 5.6.13 Decaying and rising exponential responses.

Example 5.6.2 Pulse Response

A problem of great practical importance is the transmission of voltage *pulses* along cables. Short voltage pulses are used to represent the two-level binary signals that are characteristic of digital computers; it is often necessary to transmit such voltage pulses over a long distance through **coaxial cables**, which are characterized by a finite resistance per unit length and by a certain capacitance per unit length, usually expressed in units of pF/m. A simplified model of a long coaxial cable is shown in Figure 5.6.14. It has the appearance of a low-pass filter. If a 100-m cable has a capacitance of 40 pF/m and a series resistance of 0.2 Ω/m, what will the output pulse look like after traveling the length of the cable? Assume the input pulse has a duration of 0.1 μsec and has an amplitude of 5 V. The load resistance is 150 Ω.

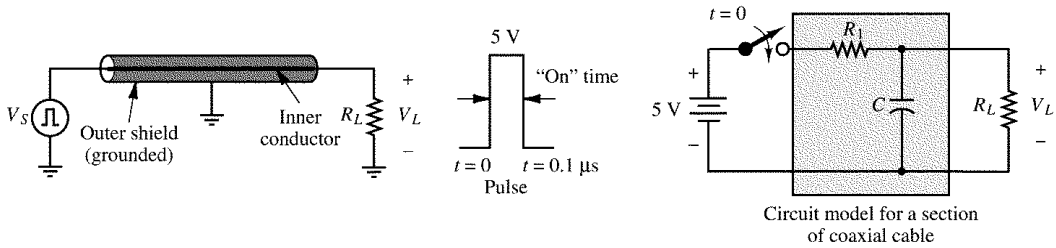


FIGURE 5.6.14 Pulse transmission in a coaxial cable.

Solution. The Thévenin equivalent circuit seen by the capacitor will vary, depending on whether the pulse is "on" or "off". In the former case, the equivalent resistance consists of the parallel combination of R_1 and R_L; in the latter case, since the switch is open, the capacitor is connected only to R_L. Thus, the effect of the pulse is to charge C through the parallel combination of R_1 and R_L during the "on" time ($0 \le t < 0.1$ μsec); the capacitor will then discharge through R_L during the "off" time. This behavior is depicted by the circuit model of Figure 5.6.14 in which the pulse signal is represented by a 5-V battery in series with a switch.

The charging time constant of the coaxial cable equivalent circuit when the switch is closed is therefore given by

$$\tau_{\text{on}} = \left(R_1 \| R_L\right) C = 17.65 \times 4000 \times 10^{-12} = 0.07 \ \mu\text{sec}$$

and the transient response of the cable during the "on" time is

$$v_L(t) = 4.41\left(1 - e^{-t/\tau}\right) = 4.41\left(1 - e^{-1.42 \times 10^7 t}\right) \qquad 0 \le t \le 0.1 \ \mu\text{sec}$$

where $V_T = R_L/(R_1 + R_L) \times 5 = 4.41$ V. At $t = 0.1$ μsec we calculate the load voltage to be

$$v_L(0.1 \ \mu\text{sec}) = 4.41\left(1 - e^{-1.42}\right) = 3.35 \text{ V}$$

For $t \ge 0.1$ μsec, the output will naturally decay to zero, starting from the initial condition, v_L (0.1 μsec), with a time constant τ_{off} equal to

$$\tau_{\text{off}} = R_L C = 150 \times 4000 \times 10^{-12} = 0.6 \ \mu\text{sec}$$

The load voltage will therefore decay according to the following expression:

$$v_L(t) = 3.35\left(e^{-1.67 \times 10^6 \left(t - 0.1 \times 10^{-6}\right)}\right) \qquad t > 0.1 \ \mu\text{sec}$$

The appearance of the response is shown in Figure 5.6.15. It should be apparent that as the cable becomes longer, both R_1 and C increase, and therefore the output voltage will respond more slowly to the input pulse; according to the simple model of a long coaxial cable given in this example, there is a limit to the maximum distance over which a voltage pulse can be transmitted by cable.

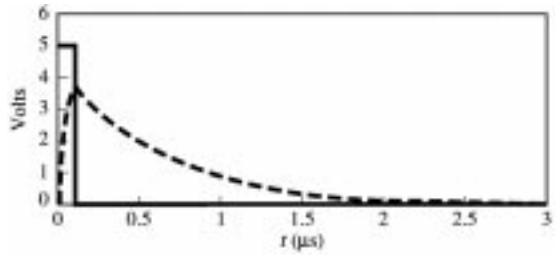


FIGURE 5.6.15 Pulse response of 100-m-long coaxial cable.

5.7 Electronics

This chapter introduces semiconductor-based electronic devices, and in so doing, it provides a transition between the fundamentals of electrical circuit analysis and the study of electronic circuits.

Semiconductors and *pn* Junctions

This section briefly introduces the mechanism of conduction in a class of materials called **semiconductors**. Semiconductors are typically materials consisting of elements from group IV of the periodic table and having electrical properties falling somewhere between those of conducting and of insulating materials. As an example, consider the conductivity of three common materials. Copper, a good conductor, has a conductivity of 0.59×16^6 S/cm; glass, a common insulator, may range between 10^{-16} and 10^{-13} S/cm; while silicon, a semiconductor, has a conductivity that varies from 10^{-8} to 10^{-1} S/cm. You see, then, that the name *semiconductor* is an appropriate one.

A conducting material is characterized by a large number of conduction-band electrons, which have a very weak bond with the basic structure of the material. Thus, an electric field easily imparts energy to the outer electrons in a conductor and enables the flow of electric current.

The free valence electrons are not the only mechanism of conduction in a semiconductor, however, Whenever a free electron leaves the lattice structure, it creates a corresponding positive charge within the lattice. The vacancy caused by the departure of a free electron is called a **hole**. Note that whenever a hole is present, we have, in effect, a positive charge. The positive charges also contribute to the conduction process, in the sense that if a valence-band electron "jumps" to fill a neighboring hole, thereby neutralizing a positive charge, it correspondingly creates a new hole at a different location. Thus, the effect is equivalent to that of a positive charge moving.

Semiconductor technology rarely employs pure, or intrinsic, semiconductors. To control the number of charge carriers in a semiconductor, the process of **doping** is usually employed. Doping consists of adding impurities to the crystalline structure of the semiconductor. The amount of these impurities is controlled, and the impurities can be of one of two types.

Semiconductors doped with donor elements conduct current predominantly by means of free electrons and are therefore called ***n*-type semiconductors**. When an acceptor element is used as the dopant, holes constitute the most common carrier, and the resulting semiconductor is said to be a ***p*-type semiconductor**. Doping usually takes place at such levels that the concentration of carriers due to the dopant is significantly greater than the intrinsic concentration of the original semiconductor. If n is the total number of free electrons and p that of holes, then in an n-type doped semiconductor, we have

$$n \gg n_i$$

and

$$p \ll p_i$$

Thus, free electrons are the **majority carriers** in an n-type material, while holes are the **minority carriers**. In a p-type material, the majority and minority carriers are reversed.

A simple section of semiconductor material does not in and of itself possess properties that make it useful for the construction of electronic circuits. However, when a section of p-type material and a section of n-type material are brought in contact to form a ***pn* junction**, a number of interesting properties arise. The pn junction forms the basis of the **semiconductor diode**, a widely used circuit element.

Figure 5.7.1 depicts an idealized pn junction, where on the p side, we see a dominance of positive charge carriers, or holes, and on the n side, the free electrons dominate. The charge separation causes a **contact potential** to exist at the junction. This potential is typically on the order of a few tenths of a

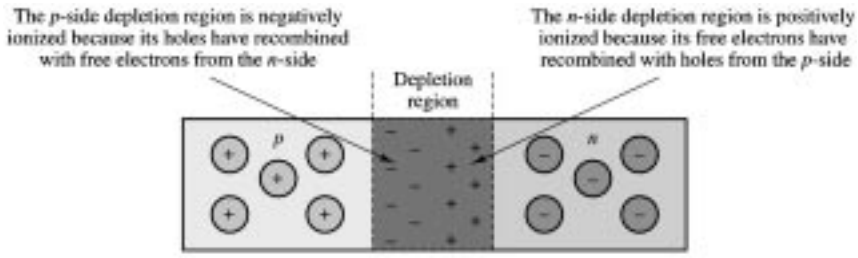


FIGURE 5.7.1 A *pn* junction.

volt and depends on the material (about 0.6 to 0.7 V for silicon). The contact potential is also called the *offset voltage*, V_γ.

Consider the diagrams of Figure 5.7.2, where a battery has been connected to a *pn* junction in the **reverse-biased** direction (Figure 5.7.2(a)), and in the **forward-biased** direction (Figure 5.7.2(b)). We assume that some suitable form of contact between the battery wires and the semiconductor material can be established (this is called an **ohmic contact**). The effect of a reverse bias is to increase the contact potential at the junction. Now, the majority carriers trying to diffuse across the junction need to overcome a greater barrier (a larger potential) and a wider depletion region. Thus, the diffusion current becomes negligible. The only current that flows under reverse bias is the very small reverse saturation current, so that the diode current, i_D (defined in the figure), is

$$i_D = -I_0$$

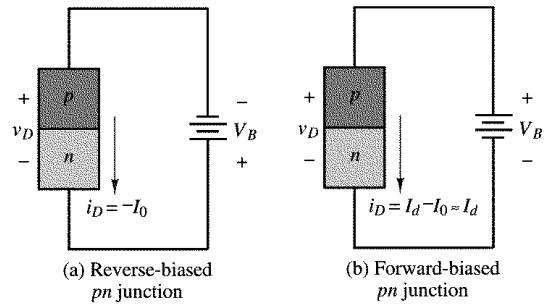


FIGURE 5.7.2 Forward- and reverse-biased *pn* junctions.

When the *pn* junction is forward-biased, the contact potential across the junction is lowered (note that V_B acts in opposition to the contact potential). Now, the diffusion of majority carriers is aided by the external voltage source; in fact, the diffusion current increases as a function of the applied voltage, according to the expression

$$I_d = I_0 e^{qv_D/kT} \tag{5.7.1}$$

where v_D is the voltage across the *pn* junction, $k = 1.381 \times 10^{-23}$ j/K is Boltzmann's constant, q the charge of one electron, and T the temperature of the material in kelvins (K). The quantity kT/q is constant at a given temperature and is approximately equal to 25 mV at room temperature. The net diode current under forward bias is given by the expression

$$i_D = I_d - I_0 = I_0\left(e^{qv_D/kT} - 1\right) \tag{5.7.2}$$

which is known as the **diode equation**. Figure 5.7.3 depicts the diode *i-v* characteristic described by the diode equation for a fairly typical silicon diode for positive diode voltages. Since the reverse saturation current, I_0, is typically very small (10^{-9} to 10^{-15} A), the expression

$$i_D = I_0 e^{qv_D/kT} \tag{5.7.3}$$

is a good approximation if the diode voltage, v_D, is greater than a few tenths of a volt.

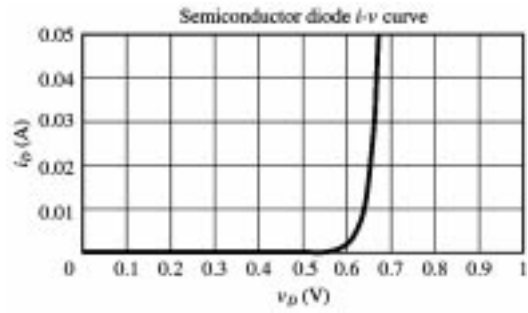


FIGURE 5.7.3 Semiconductor diode *i-v* characteristic.

The ability of the *pn* junction to essentially conduct current in only one direction — that is, to conduct only when the junction is forward-biased — makes it valuable in circuit applications. A device having a single *pn* junction and ohmic contacts at its terminals, as described in the preceding paragraphs, is called a *semiconductor diode,* or simply *diode*. As will be shown later in this chapter, it finds use in many practical circuits. The circuit symbol for the diode is shown in Figure 5.7.4, alongside with a sketch of the *pn* junction.

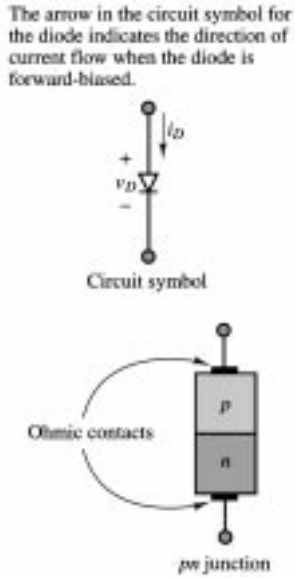


FIGURE 5.7.4 Semiconductor diode circuit symbol.

Circuit Models for the Semiconductor Diode

From the viewpoint of a *user* of electronic circuits (as opposed to a *designer*), it is often sufficient to characterize a device in terms of its *i-v* characteristic, using appropriate circuit models to determine the operating currents and voltages.

Ideal Diode Model

The large-signal model treats the diode as a simple on–off device (much like a check valve in hydraulic circuits). Figure 5.7.5 illustrates how, on a large scale, the *i-v* characteristic of a typical diode may be approximated by an open circuit when $v_D < 0$ and by a short circuit when $v_D \geq 0$ (recall the *i-v* curves of the ideal short and open circuits presented in Section 5.2). The analysis of a circuit containing a diode may be greatly simplified by using the short-circuit–open-circuit model. From here on, this diode model will be known as the **ideal diode model**. In spite of its simplicity, the ideal diode model (indicated by the symbol shown in Figure 5.7.5 can be very useful in analyzing diode circuits.

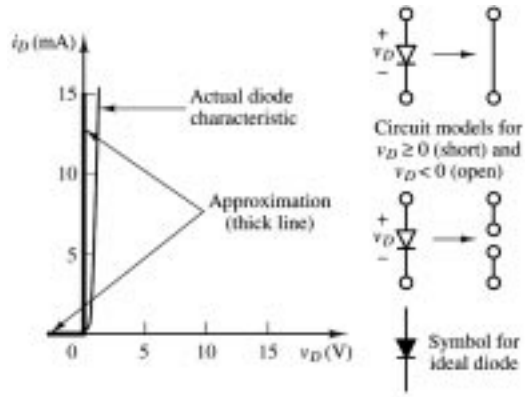


FIGURE 5.7.5 Large-signal on–off diode model.

One of the important applications of the semiconductor diode is **rectification** of AC signals, that is, the ability to convert an AC signal with zero average (DC) value to a signal with a nonzero DC value. The application of the semiconductor diode as a rectifier is very useful in obtaining DC voltage supplies from the readily available AC line voltage. Here we illustrate the basic principle of rectification using an ideal diode, for simplicity, and also because the large-signal model is appropriate when the diode is used in applications involving large AC voltage and current levels.

Consider the circuit of Figure 5.7.6, where an AC source, $v_i = 155.56 \cdot \sin \omega t$, is connected to a load by means of a series ideal diode. The diode will conduct only during the positive half-cycle of the sinusoidal voltage — that is, that the condition $v_D \geq 0$ will be satisfied only when the AC source voltage is positive — and that it will act as an open circuit during the negative half-cycle of the sinusoid ($v_D < 0$). Thus, the appearance of the load voltage will be as shown in Figure 5.7.7 with the negative portion of the sinusoidal waveform cut off. The rectified waveform clearly has a nonzero DC (average) voltage, whereas the average input waveform voltage was zero. When the diode is conducting, or $v_D \geq 0$, the unknowns v_L and i_D can be found by using the following equations:

$$i_D = \frac{v_i}{R_L} \quad \text{when} \quad v_i > 0 \tag{5.7.4}$$

and

$$v_L = i_D R_L \tag{5.7.5}$$

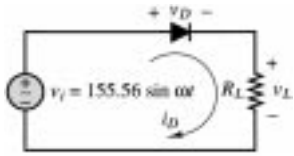


FIGURE 5.7.6

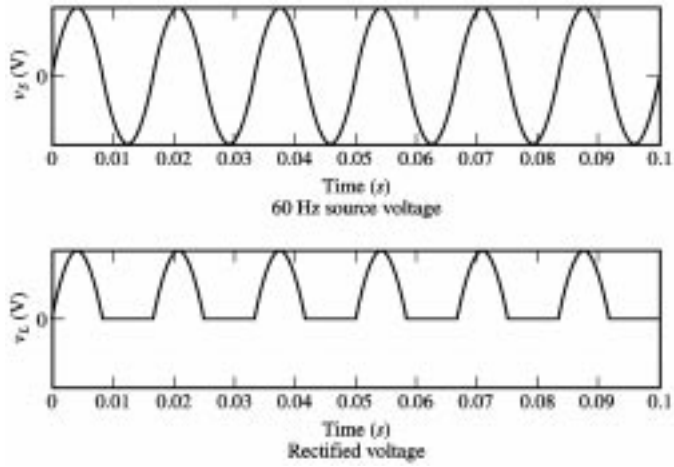


FIGURE 5.7.7 Ideal diode rectifier input and output voltages.

The load voltage, v_L and the input voltage, v_i, are sketched in Figure 5.7.7. From Equation (5.7.5) it is obvious that the current waveform has the same shape as the load voltage. The average value of the load voltage is obtained by integrating the load voltage over one period and dividing by the period:

$$v_{\text{load,DC}} = \frac{\omega}{2\pi}\int_0^{\pi/\omega} 155.56 \sin \omega t \, dt = \frac{155.56}{\pi} = 49.52 \text{ V} \tag{5.7.6}$$

The circuit of Figure 5.7.6 is called a **half-wave rectifier**, since it preserves only half of the waveform. This is not usually a very efficient way of rectifying an AC signal, since half the energy in the AC signal is not recovered. It will be shown in a later section that it is possible to recover also the negative half of the AC waveform by mans of a *full-wave rectifier.*

Offset Diode Model

While the ideal diode model is useful in approximating the large-scale characteristics of a physical diode, it does not account for the presence of an offset voltage, which is an unavoidable component in semiconductor diodes. The **offset diode model** consists of an ideal diode in series with a battery of strength equal to the offset voltage (we shall use the value $V_\gamma = 0.6$ V for silicon diodes, unless otherwise indicated). The effect of the battery is to shift the characteristic of the ideal diode to the right on the voltage axis, as shown in Figure 5.9.3. This model is a better approximation of the large-signal behavior of a semiconductor diode than the ideal diode model.

According to the offset diode model, the diode of Figure 5.7.8 acts as an open circuit for $v_D < 0.6$ V, and it behaves like a 0.6-V battery for $v_D \geq 0.6$ V. The equations describing the offset diode model are as follows:

$$\begin{aligned} v_D &\geq 0.6 \text{ V} \quad \text{Diode} \rightarrow \text{0.6-V battery} \\ v_D &< 0.6 \text{ V} \quad \text{Diode} \rightarrow \text{Open circuit} \end{aligned} \tag{5.7.7}$$

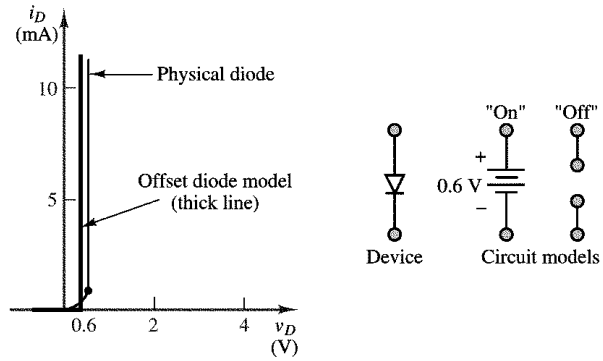


FIGURE 5.7.8

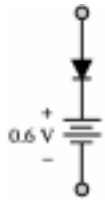


FIGURE 5.7.9 Offset diode as an extension of ideal diode model.

The diode offset model may be represented by an ideal diode in series with a 0.6-V ideal battery, as shown in Figure 5.7.9.

Practical Diode Circuits

The Full-Wave Rectifier

The half-wave rectifier discussed earlier is one simple method of converting AC energy to DC energy. The need for converting one form of electrical energy into the other arises frequently in practice. The most readily available form of electric power is AC (the standard 110- or 220-V rms AC line power), but one frequently needs a DC power supply, for applications ranging from the control of certain types of electric motors to the operation of electronic circuits.

The half-wave rectifier, however, is not a very efficient AC–DC conversion circuit, because it fails to utilize half the energy available in the AC waveform by not conducting current during the negative half-cycle of the AC waveform. The full-wave rectifier shown in Figure 5.7.10 offers a substantial improvement in efficiency over the half-wave rectifier. The first section of the full-wave rectifier circuit includes an AC source and a center-tapped transformer (see Section 5.5) with 1:2 N turns ratio. The purpose of the transformer is to obtain the desired voltage amplitude prior to rectification. Thus, if the peak amplitude of the AC source voltage is v_s, the amplitude of the voltage across each half of the output side of the transformer will be Nv_s; this scheme permits scaling the source voltage up or down (depending on whether N is greater or less than 1), according to the specific requirements of the application. In addition to scaling the source voltage, the transformer also isolates the rectifier circuit from the AC source voltage, since there is no direct electrical connection between the input and output of a transformer.

The Bridge Rectifier

Another rectifier circuit commonly available "off the shelf" as a single *integrated circuit package* is the *bridge rectifier*, which employs four diodes in a bridge configuration, similar to the Wheatstone bridge already explored in Section 5.2. Figure 5.7.11 depicts the bridge rectifier, along with the associated integrated circuit (IC) package.

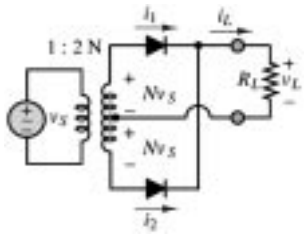
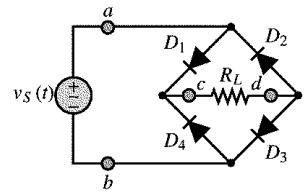
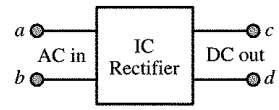


FIGURE 5.7.10 Full-wave rectifier.



Bridge rectifier



Corresponding IC package

FIGURE 5.7.11 Full-wave bridge rectifier.

The analysis of the bridge rectifier is simple to understand by visualizing the operation of the rectifier for the two half-cycles of the AC waveform separately. The key is that, as illustrated in Figure 5.7.12, diodes D_1 and D_3 conduct during the positive half-cycle, while diodes D_2 and D_4 conduct during the negative half-cycle. Because of the structure of the bridge, the flow of current through the load resistor is in the same direction (from *c* to *d*) during both halves of the cycle; hence, the full-wave rectification of the waveform.

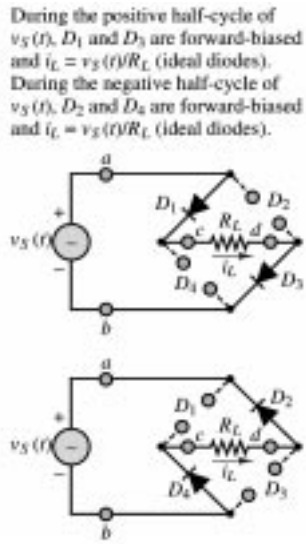


FIGURE 5.7.12 Diodes conduct.

DC Power Supplies and Voltage Regulation

The principal application of rectifier circuits is in the conversion of AC to DC power. A circuit that accomplishes this conversion is usually called a **DC power supply**. In power supply applications, transformers are employed to obtain an AC voltage that is reasonably close to the desired DC supply voltage. DC power supplies are very useful in practice: many familiar electrical and electronic appliances (e.g., radios, personal computers, TVs) require DC power to operate. For most applications, it is desirable

that the DC supply be as steady and ripple-free as possible. To ensure that the DC voltage generated by a DC supply is constant, the DC supply is made up of voltage regulators, that is, devices that can hold a DC load voltage relatively constant in spite of possible fluctuations in the DC supply. This section will describe the fundamentals of voltage regulators.

A typical DC power supply is made up of the components shown in Figure 5.7.13. In the figure, a transformer is shown connecting the AC source to the rectifier circuit to permit scaling of the AC voltage to the desired level. For example, one might wish to step the 110-V rms line voltage down to 24 V rms by means of a transformer prior to rectification and filtering, to eventually obtain a 12-VDC regulated supply (*regulated* here means that the output voltage is a DC voltage that is constant and independent of load and supply variations). Following the step-down transformer are a bridge rectifier, a filter capacitor, a voltage regulator, and, finally, the load.

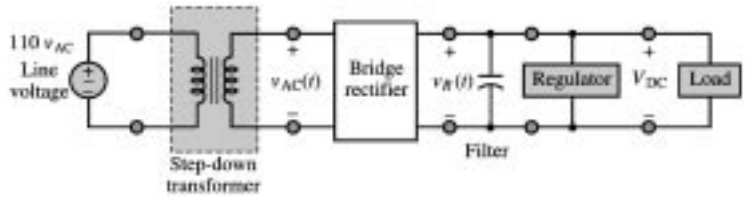


FIGURE 5.7.13 DC power supply.

The most common device employed in voltage regulation schemes is the Zener diode. Zener diodes function on the basis of the reverse portion of the *i-v* characteristic of the diode with forward offset voltage V_γ and **reverse Zener voltage** V_Z.

The operation of the Zener diode may be analyzed by considering three modes of operation:

1. For $v_D \geq V_\gamma$, the device acts as a conventional forward-biased diode (Figure 5.7.14).
2. For $V_Z < v_D < V_\gamma$, the diode is reverse-biased but Zener breakdown has not taken place yet. Thus, it acts as an open circuit.
3. For $v_D \leq V_Z$, Zener breakdown occurs and the device holds a nearly constant voltage, $-V_Z$ (Figure 5.7.15).

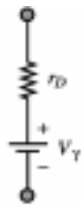


FIGURE 5.7.14 Zener diode model for forward bias.

FIGURE 5.7.15 Zener diode model for reverse bias.

To illustrate the operation of a Zener diode as a voltage regulator, consider the circuit of Figure 5.7.16, where the unregulated DC source, v_S, is regulated to the value of the Zener voltage, V_Z. Note how the diode must be connected "upside down" to obtain a positive regulated voltage. Note also that if v_S is greater than V_Z, it follows that the Zener diode is in its reverse-breakdown mode. Thus, one need not

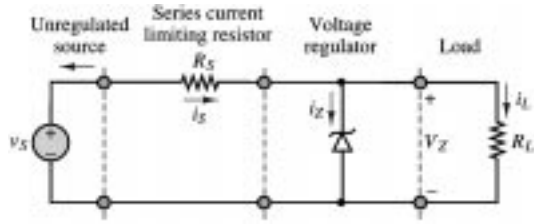


FIGURE 5.7.16 A Zener diode voltage regulator.

worry whether the diode is conducting or not in simple voltage regulator problems, provided that the unregulated supply voltage is guaranteed to stay above V_Z (a problem arises, however, if the unregulated supply can drop below the Zener voltage). Assuming that the resistance r_Z is negligible with respect to R_S and R_L, we replace the Zener diode with the simplified circuit mode of Figure 5.7.17 consisting of a battery of strength V_Z.

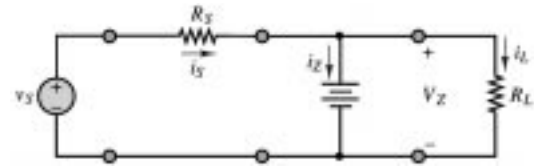


FIGURE 5.7.17 Simplified circuit for Zener regulator.

Three simple observations are sufficient to explain the operation of this voltage regulator:

1. The load voltage must equal V_Z, as long as the Zener diode is in the reverse-breakdown mode. Then,

$$i_L = \frac{V_Z}{R_L} \tag{5.7.10}$$

2. The load current (which should be constant if the load voltage is to be regulated to sustain V_Z) is the difference between the unregulated supply current, i_s, and the diode current i_z:

$$i_L = i_S - i_Z \tag{5.7.11}$$

   This second point explains intuitively how a Zener diode operates: any current in excess of that required to keep the load at the constant voltage V_Z is "dumped" to ground through the diode. Thus, the Zener diode acts as a sink to the undesired source current.
3. The source current is given by

$$i_S \frac{v_S - V_Z}{R_S} \tag{5.7.12}$$

The Zener diode is usually rated in terms of its maximum allowable power dissipation. The power dissipated by the diode, P_Z, may be computed from

$$P_Z = i_Z V_Z \tag{5.7.13}$$

Thus, one needs to worry about the possibility that i_Z will become too large. This may occur either if the supply current is very large (perhaps because of an unexpected upward fluctuation of the unregulated

supply), or if the load is suddenly removed and all of the supply current sinks through the diode. The latter case, of an open-circuit load, is an important design consideration.

Another significant limitation occurs when the load resistance is small, thus requiring large amounts of current from the unregulated supply. In this case, the Zener diode is hardly taxed at all in terms of power dissipation, but the unregulated supply may not be able to provide the current required to sustain the load voltage. In this case, regulation fails to take place. Thus, in practice, the range of load resistances for which load voltage regulation may be attained is constrained to a finite interval:

$$R_{L\,\min} \le R_L \le R_{L\,\max} \tag{5.7.14}$$

where $R_{L\max}$ is typically limited by the Zener diode power dissipation and $R_{L\min}$ by the maximum supply current.

Example 5.7.1

This example illustrates the calculation of the range of allowable load resistances for a Zener regulator design. For the Zener regulator shown in Figure 5.7.18, we want to maintain the load voltage at 14 V. Find the range of load resistances for which regulation can be obtained if the Zener diode is rated at 14 V, 5 W.

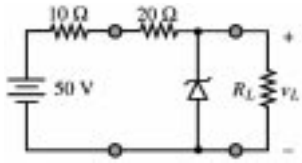


FIGURE 5.7.18 Circuit for example.

Solution. The minimum load resistance for which a regulated load voltage of 14 V may be attained is found by requiring that the load voltage be 14 V and applying KVL subject to this constraint:

$$50\left(\frac{R_{L\,\min}}{R_{L\,\min} + 30}\right) = 14$$

$$R_{L\,\min} = \frac{14}{50}\left(R_{L\,\min} + 30\right)$$

$$= 11.7\ \Omega$$

The maximum current through the Zener diode that does not exceed the diode power rating may be computed by considering the 5-W power rating:

$$i_{Z\,\max} = \frac{5}{14} = 0.357\ \text{A}$$

The current through the 20-Ω resistor will be

$$\frac{50 - 14}{30} = 1.2\ \text{A}$$

so that the maximum load resistance for which regulation occurs is

$$R_{L\max} = \frac{14}{1.2 - 0.357}$$

$$= 16.6\ \Omega$$

Finally, the range of allowable load resistances is:

$$16.6\ \Omega \geq R_L \geq 11.7\ \Omega$$

Photodiodes

Another property of semiconductor materials that finds common application in measurement systems is their response to light energy. In appropriately fabricated diodes, called **photodiodes**, when light reaches the depletion region of a *pn* junction, photons cause hole-electron pairs to be generated by a process called *photo-ionization*. This effect can be achieved by using a surface material that is transparent to light. As a consequence, the reverse saturation current depends on the light intensity (i.e., on the number of incident photons), in addition to the other factors mentioned earlier. In a photodiode, the reverse current is given by $-(I_0 + I_p)$, where I_p is the additional current generated by photo-ionization. The result is depicted in the family of curves of Figure 5.7.19, where the diode characteristic is shifted downward by an amount related to the additional current generated by photo-ionization. Figure 5.7.19 depicts the appearance of the *i-v* characteristic of a photodiode for various values of I_p, where the *i-v* curve is shifted to lower values for progressively larger values of I_p. The circuit symbol is depicted in Figure 5.7.20.

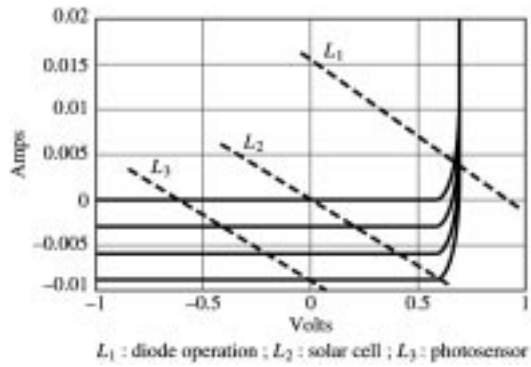


FIGURE 5.7.19 Photodiode *i-v* curves.

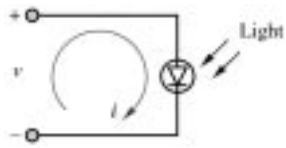


FIGURE 5.7.20 Photodiode circuit symbol.

As displayed in Figure 5.7.19 are three load lines, which depict the three modes of operation of a photodiode. Curve L_1 represents normal diode operation, under forward bias. Note that the operating point of the device is in the positive *i*, positive *v* (first) quadrant of the *i-v* plane; thus, the diode dissipates positive power in this mode, and is therefore a passive device, as we already know. On the other hand, load line L_2 represents operation of the photodiode as a **solar cell**; in this mode, the operating point is in the negative *i*, positive *v*, or fourth, quadrant, and therefore the power dissipated by the diode is *negative*. In other words, the photodiode is generating power by converting light energy to electrical energy. Note further that the load line intersects the voltage axis at zero, meaning that no supply voltage is required to bias the photodiode in the solar-cell mode. Finally, load line L_3 represents the operation

of the diode as a light sensor: when the diode is reverse-biased, the current flowing through the diode is determined by the light intensity; thus, the diode current changes in response to changes in the incident light intensity.

The operation of the photodiode can also be reversed by forward-biasing the diode and causing a significant level of recombination to take place in the depletion region. Some of the energy released is converted to light energy by emission of photons. Thus, a diode operating in this mode emits light when forward-biased. Photodiodes used in this way are called **light-emitting diodes** (LEDs); they exhibit a forward (offset) voltage of 1 to 2 volts. The circuit symbol for the LED is shown in Figure 5.7.21.

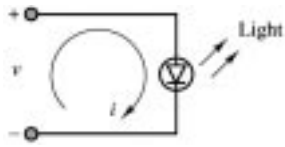


FIGURE 5.7.21 Light-emitting diode (LED) circuit symbol.

Gallium arsenide (GaAs) is one of the more popular substrates for creating LEDs; gallium phosphide (GaP) and the alloy $GaAs_{1-x}P_x$ are also quite common. Table 5.7.1 lists combinations of materials and dopants used for common LEDs and the colors they emit. The dopants are used to create the necessary *pn* junction.

TABLE 5.7.1 LED Materials and Wavelengths

Material	Dopant	Wavelength (nm)	Color
GaAs	Zn	900	Infrared
GaAs	Si	910–1020	Infrared
GaP	N	570	Green
GaP	N	590	Yellow
GaP	Zn,O	700	Red
$GaAs_{0.6}P_{0.4}$		650	Red
$GaAs_{0.35}P_{0.65}$	N	632	Orange
$GaAs_{0.15}P_{0.85}$	N	589	Yellow

Example 5.7.2 Opto-Isolators

One of the common applications of photodiodes and LEDs is the **opto-coupler**, or **opto-isolator**. This device, which is usually enclosed in a sealed package, uses the light-to-current and current-to-light conversion property of photodiodes and LEDs to provide signal connection between two circuits without any need for electrical connections. Figure 5.7.22 depicts the circuit symbol for the opto-isolator.

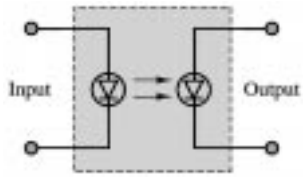


FIGURE 5.7.22 Opto isolator.

Because diodes are nonlinear services, the opto-isolator is not used in transmitting analog signals: the signals would be distorted because of the nonlinear diode *i-v* characteristic. However, opto-isolators find a very important application when on–off signals need to be transmitted from high-power machinery to delicate computer control circuitry. The optical interface ensures that potentially damaging large currents cannot reach delicate instrumentation and computer circuits.

Transistors

A transistor is a three-terminal semiconductor device that can perform two functions that are fundamental to the design of electronic circuits: **amplification** and **switching**. Put simply, amplification consists of magnifying a signal by transferring energy to it from an external source; whereas a transistor switch is a device for controlling a relatively large current between or voltage across two terminals by means of a small control current or voltage applied at a third terminal. In this chapter, we provide an introduction to the two major families of transistors: *bipolar junction transistors*, or *BJTs*; and *field-effect transistors*, or *FETs*.

The operation of the transistor as a linear amplifier can be explained qualitatively by the sketch of Figure 5.7.23 in which the four possible modes of operation of a transistor are illustrated by means of circuit models employing controlled sources. In Figure 5.7.23 controlled voltage and current sources are shown to generate an output proportional to an input current or voltage; the proportionality constant, μ, is called the internal *gain* of the transistor. As will be shown, the BJT acts essentially as a current-controlled device, while the FET behaves as a voltage-controlled device.

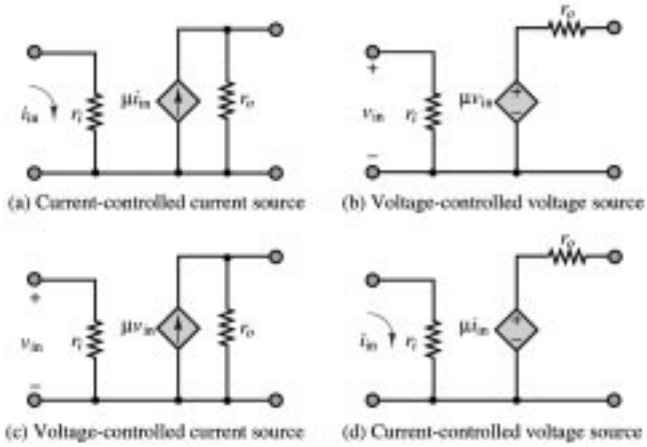


FIGURE 5.7.23 Controlled-source models of linear amplifier transistor operation.

Transistors can also act in a nonlinear mode, as voltage- or current-controlled switches. When a transistor operates as a switch, a small voltage or current is used to control the flow of current between two of the transistor terminals in an on–off fashion. Figure 5.7.24 depicts the idealized operation of the transistor as a switch, suggesting that the switch is closed (on) whenever a control voltage or current is greater than zero and is open (off) otherwise. It will later become apparent that the conditions for the switch to be on or off need not necessarily be those depicted in Figure 5.7.24.

The Bipolar Junction Transistor (BJT)

The *pn* junction forms the basis of a large number of semiconductor devices. The semiconductor diode, a two-terminal device, is the most direct application of the *pn* junction. In this section, we introduce the **bipolar junction transistor** (BJT). As we did in analyzing the diode, we will introduce the physics of transistor devices as intuitively as possible, resorting to an analysis of their *i-v* characteristics to discover important properties and applications.

A BJT is formed by joining three sections of semiconductor material, each with a different doping concentration. The three sections can be either a thin *n* region sandwiched between p^+ and *p* layers, or a *p* region between *n* and n^+ layers, where the superscript "plus" indicates more heavily doped material.

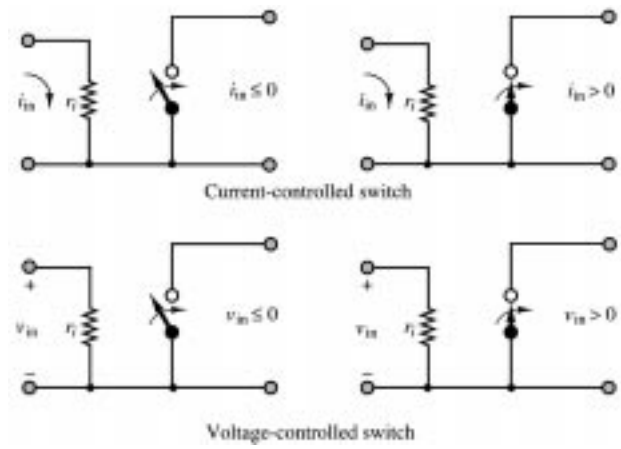


FIGURE 5.7.24 Models of ideal transistor switches.

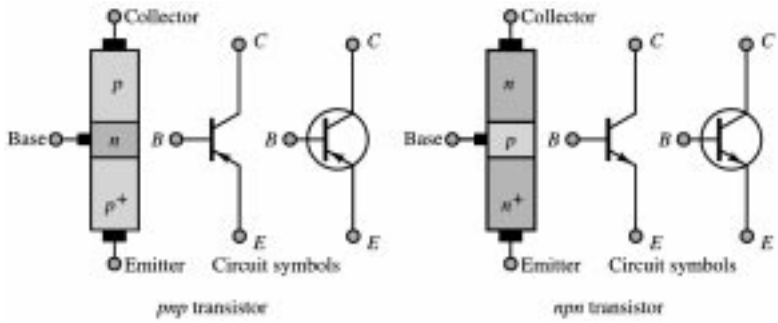


FIGURE 5.7.25 Bipolar junction transistors.

The resulting BJTs are called *pnp* and *npn* transistors, respectively; we shall discuss only the latter in this section. Figure 5.7.25 illustrates the approximate construction, symbols, and nomenclature for the two types of BJTs.

The most important property of the bipolar transistor is that the small base current controls the amount of the much larger collector current

$$I_C = \beta I_B \tag{5.7.15}$$

where β is a current amplification factor dependent on the physical properties of the transistor. Typical values of β range from 20 to 200.

The detailed operation of bipolar transistors can be explained by resorting to a detailed physical analysis of the *npn* or *pnp* structure of these devices. The reader interested in such a discussion of transistors is referred to any one of a number of excellent books on semiconductor electronics.

The focus of this section will be on the analysis of the *i-v* characteristic of the *npn* BJT, based on the circuit notation defined in Figure 5.7.26. The device *i-v* characteristics will be presented qualitatively, without deriving the underlying equations, and will be utilized in constructing circuit models for the device.

The number of independent variables required to uniquely define the operation of the transistor may be determined by applying KVL and KCL to the circuit of Figure 5.7.26. It should be apparent that two voltages and two currents are sufficient to specify the operation of the device. Note that, since the BJT is a three-terminal device, it will not be sufficient to deal with a single *i-v* characteristic; it will soon

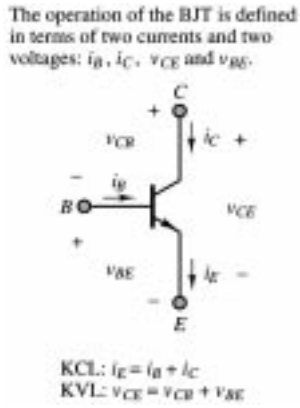


FIGURE 5.7.26 Definition of BJT voltages and currents.

become apparent that two such characteristics are required to explain the operation of this device. One of these characteristics relates the base current, i_B, to the base-emitter voltage, v_{BE}; the other relates the collector current, i_C, to the collector-emitter voltage, v_{CE}. As will be shown, the latter characteristic actually consists of a *family* of curves. To determine these *i-v* characteristics, consider the *i-v* curves of Figures 5.7.27 and 5.7.28 using the circuit notation of Figure 5.7.26. In Figure 5.7.27, the collector is open and the *BE* junction is shown to be very similar to a diode. The ideal current source, I_{BB}, injects a base current, which causes the junction to be forward-biased. By varying I_{BB}, one can obtain the open-collector *BE* junction *i-v* curve shown in the figure.

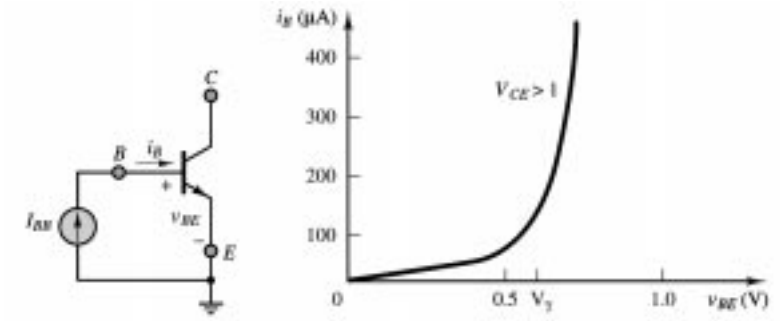


FIGURE 5.7.27 Determining the *BE* junction open-collector *i-v* characteristic.

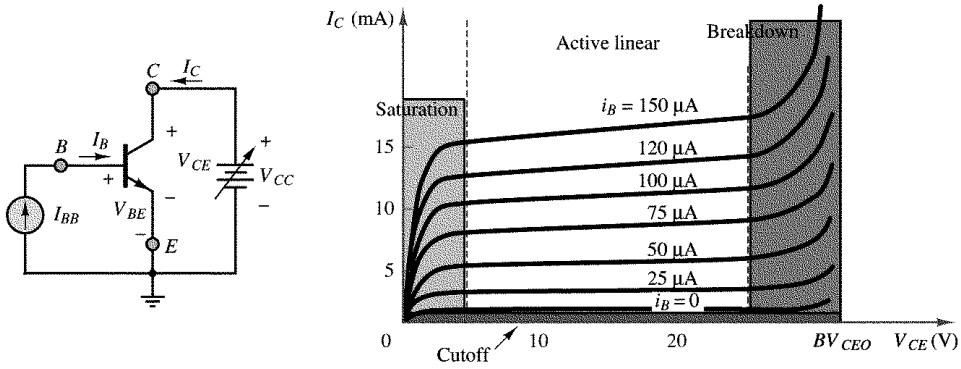


FIGURE 5.7.28 (a) Ideal test circuit to determine the *i-v* characteristic of a BJT; (b) the collector-emitter output characteristics of a BJT.

If a voltage source were now to be connected to the collector circuit, the voltage v_{CE} and, therefore, the collector current, i_C, could be varied, in addition to the base current, i_B. The resulting circuit is depicted in Figure 5.7.28(a). By varying both the base current and the collector-emitter voltage, one could then generate a plot of the device **collector characteristic**. This is also shown in Figure 5.7.28(b). Note that this figure depicts not just a single i_C-v_{CE} curve, but an entire family, since for each value of the base current, i_B, an i_C-v_{CE} curve can be generated. Four regions are identified in the collector characteristic:

1. The **cutoff region**, where both junctions are reverse-biased, the base current is very small, and essentially no collector current flows
2. The **active linear region**, in which the transistor can act as a linear amplifier, where the *BE* junction is forward-biased and the *CB* junction is reverse-biased
3. The **saturation region**, in which both junctions are forward-biased
4. The **breakdown region**, which determines the physical limit of operation of the device.

Large-Signal Model of the *npn* BJT

The large-signal model for the BJT recognizes three basic operating modes of the transistor (Figure 5.7.29). When the *BE* junction is reverse-biased, no base current (and therefore no forward collector current) flows, and the transistor acts (virtually as an open circuit; the transistor is said to be in the *cutoff region*. In practice, there is always a leakage current flowing through the collector, even when $V_{BE} = 0$ and $I_B = 0$. This leakage current is denoted by I_{CEO}. When the *BE* junction becomes forward-biased, the transistor is said to be in the *active region,* and the base current is amplified by a factor of β at the collector:

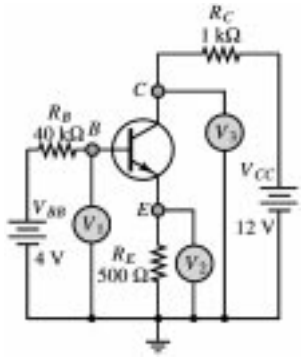


FIGURE 5.7.29 Determination of the operating state of a BJT.

$$I_C = \beta I_B \tag{5.7.16}$$

Since the collector current is controlled by the base current, the controlled-source symbol is used to represent the collector current. Finally, when the base current becomes sufficiently large, the collector-emitter voltage, V_{CE}, reaches its saturation limit, and the collector current is no longer proportional to the base current; this is called the *saturation region*. The three conditions are described in Figure 5.7.30 in terms of simple circuit models.

Example 5.7.3 illustrates the application of this large-signal model in a practical circuit and illustrates how to determine which of the three states is applicable, using relatively simple analysis.

Example 5.7.3 LED Driver

The circuit shown in Figure 5.7.31 is being used to "drive" (i.e., provide power for) a light-emitting diode (LED) from a desktop computer. The signal available from the computer consists of a 5-V low-current output. The reason for using a BJT (rather than driving the LED directly with the signal available

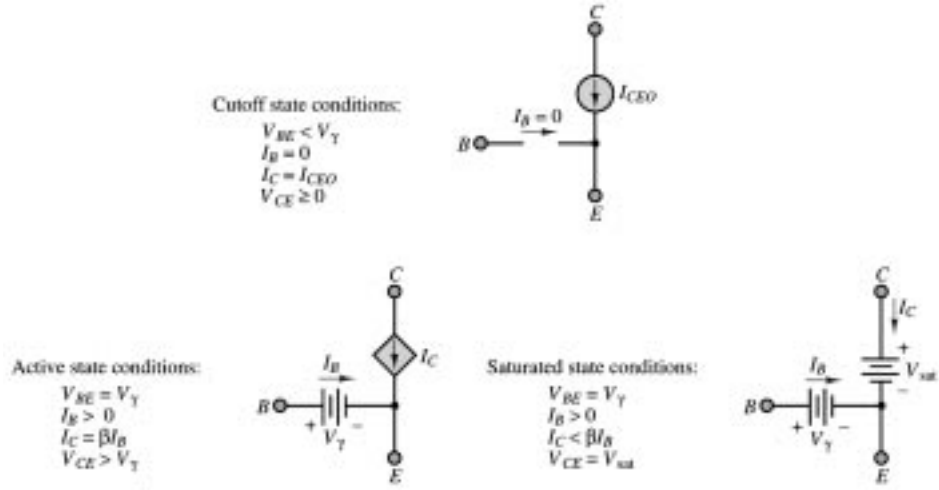


FIGURE 5.7.30 *npn* BJT large-signal model.

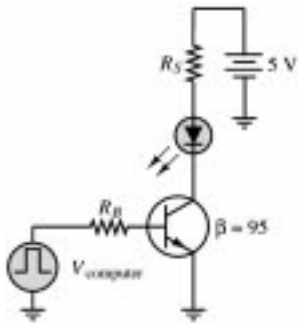


FIGURE 5.7.31 LED driver circuit.

from the computer) is that the LED requires a substantial amount of current (at least 15 mA for the device used in this example) and the computer output is limited to 5mA. Thus, some current amplification is required. The LED has an offset voltage of 1.4 V and a maximum power dissipation rating of 280 mW. The circuit has been designed so that the transistor will be either in cutoff, when the LED is to be turned off, or in saturation, when the LED is to be turned on.

Assume that the base-emitter junction of the transistor has an offset voltage of 0.7 V and that R_B = 1 kΩ and R_S = 42.5 Ω.

1. When the computer is supplying 5 V to the circuit, is the transistor in cutoff, in saturation, or in the linear region of operation?
2. Determine the current in the LED when "turned on" and, thus, state whether the LED will exceed its power rating.

Solution. The base-emitter circuit is considered first, to determine whether the *BE* junction is forward-biased. The equivalent circuit is shown in Figure 5.7.32

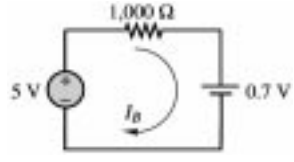


FIGURE 5.7.32 *BE* circuit for LED driver.

Writing KVL around the base circuit, we obtain

$$(5 - 0.7) = I_B(1000)$$

or

$$I_B = 4.3 \text{ mA}$$

Since this current is greater than zero (i.e., since positive current flows into the base), the transistor is not in cutoff.

Next, we need to determine whether the transistor is in the linear active or in the saturation region. One method that can be employed to determine whether the device is in the active region is to assume that it is and to solve for the circuit that results from the assumed condition. If the resulting equations are consistent, then the assumption is justified. Assuming that the transistor is in its active region, the following equations apply:

$$I_C = \beta I_B$$

or

$$I_C = 95(4.3 \text{ mA}) = 408.5 \text{ mA}$$

With reference to the circuit of Figure 5.7.33, KVL may be applied around the collector circuit, to obtain

$$5 = 1.4 + V_{CE} + I_C R_S$$

or

$$(5 - 1.4) - 408.5(42.5) = V_{CE}$$

or

$$V_{CE} = -13.76 \text{ V}$$

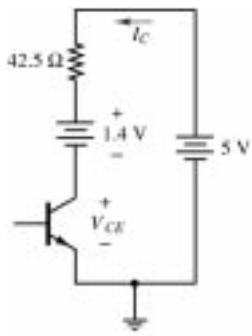


FIGURE 5.7.33 Equivalent collector circuit of LED driver, assuming that the BJT is in the linear active region.

This result is clearly not possible, since the supply voltage is only 5V! It must therefore be concluded that the transistor is not in the linear region and must be in saturation. To test this hypothesis, we can substitute the nominal saturation voltage for the BJT in place of V_{CE} (a typical saturation voltage would be V_{CE} sat = 0.2 V) and verify that in this case we obtain a consistent solution:

$$I_C = \frac{(5 - 1.4 - 0.2)}{42.5} = 80 \text{ mA}$$

This is a reasonable result, stating that the collector current in the saturation region for the given circuit is of the order of 80 mA.

In the above part, it was determined that the transistor was in saturation. Using the circuit model for saturation, we may draw the equivalent circuit of Figure 5.7.34. Since $I_C = 80$ mA, the power dissipated by the LED may be calculated as follows:

$$P_{\text{LED}} = I_C V_{\text{LED}} = 80 \text{ mA} \times 1.4 \text{ V} = 112 \text{ mW}$$

Thus, the power limitation of the LED will not be exceeded.

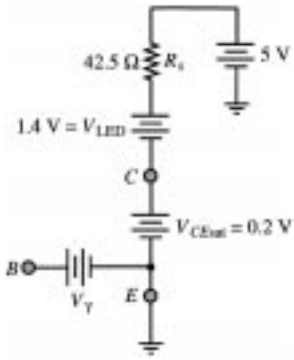


FIGURE 5.7.34 LED driver equivalent collector circuit, assuming that the BJT is in the saturation region.

Field-Effect Transistors

The second transistor family discussed in this section operates on the basis of a principle that is quite different from that of the *pn* junction devices. The concept that forms the basis of the operation of the **field-effect transistor**, or FET, is that the width of a conducting channel in a semiconductor may be varied by the external application of an electric field. Thus, FETs behave as *voltage-controlled resistors*. This family of electronic devices can be subdivided into three groups, all of which will be introduced in the remainder of this chapter. Figure 5.7.35 depicts the classification of field-effect transistors, as well as the more commonly used symbols for these devices. These devices can be grouped into three major categories. The first two categories are both types of **metal-oxide-semiconductor field-effect transistors**, or MOSFETs: **enhancement-mode MOSFETs** and **depletion-mode MOSFETs**. The third category consists of **junction field-effect transistors**, or JFETs. In addition, each of these devices can be fabricated either as an ***n*-channel** device or as a ***p*-channel** device, where the *n* or *p* designation indicates the nature of the doping in the semiconductor channel.

The construction of the MOSFET is shown in Figure 5.7.36, along with its circuit symbol. The device consists of a metal **gate**, insulated from the bulk of the *p*-type semiconductor material by an oxide layer (thus the terminology *metal-oxide-semiconductor*). Two n^+ regions on either side of the gate are called **source** and **drain**, respectively. An electric field can be applied between the gate and the *bulk* of the semiconductor by connecting the gate to a voltage source. The effect of the electric field, the intensity of which depends on the strength of the gate voltage, is to push positive charge carriers away from the surface of the bulk material. As a consequence of this effect, the *p*-type material has a lower concentration of positive charge carriers near its surface, and it behaves progressively more like intrinsic semiconductor material and then, finally, like *n*-type material as the electric field strength increases. Thus, in a narrow layer between n^+ regions, the *p*-type material is *inverted* and *n*-type charge carriers become available for conduction. This layer is called the *channel*. The device of Figure 5.7.36 takes the name *n-channel*

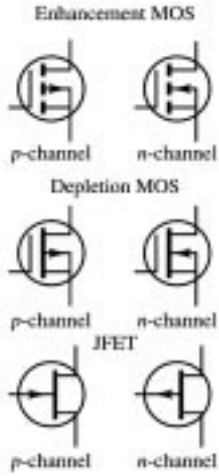


FIGURE 5.7.35 Classification of field-effect transistors.

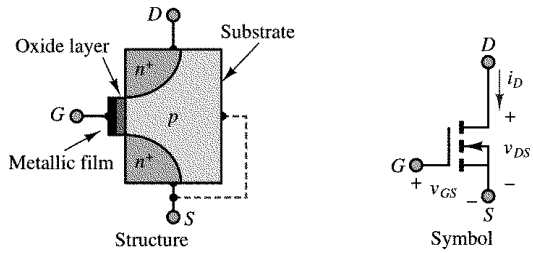


FIGURE 5.7.36 *n*-Channel enhancement MOSFET.

MOSFET. It is also called an *enhancement-mode* device, because the applied electric field "enhances" the conduction in the channel by attracting *n*-type charge carriers. There are also *depletion-mode* devices, in which the application of an electric field "depletes" the channel of charge carriers, reducing the effective channel width. It is useful to think of enhancement-mode devices as being normally off: current cannot flow from drain to source unless a gate voltage is applied. On the other hand, depletion-mode devices are normally on; that is, the gate voltage is used to reduce the conduction of current from drain to source.

An analogous discussion could be carried out for *p-channel MOSFETs*. In a *p*-channel device, conduction in the channel occurs through positive charge carriers. The correspondence between *n*-channel and *p*-channel devices is akin to that between *npn* and *pnp* bipolar devices.

We are now ready to describe qualitatively the *i-v* characteristic of this enhancement-mode MOSFET, for small values of drain-to-source voltage, v_{DS}, and for constant v_{GS}, the channel has essentially constant width and therefore acts as a constant resistance. As the gate voltage is changed, this resistance can be varied over a certain range. This mode of operation, for small drain voltages, is called the **ohmic state**, and, as depicted in Figure 5.7.37, it corresponds to a linear *i-v* curve for fixed v_{GS}, as would be expected of a resistor. Thus, in the ohmic state the MOSFET acts as a **voltage-controlled resistor**.

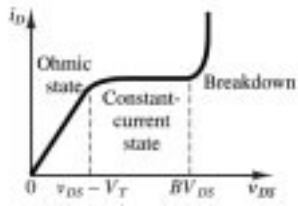


FIGURE 5.7.37 MOSFET *i-v* curve.

As the drain voltage is increased, the gate-to-drain voltage, v_{GD}, decreases, reducing the electric field strength at the drain end of the device.

When $v_{GD} = v_{GS} - v_{DS} < v_T$ and $v_{DS} > v_{GS} - v_T$, the channel is *pinched down*, and the electron flow into the drain is physically limited, so that the drain current becomes essentially constant. This phenomenon is clearly reflected in the curve of Figure 5.7.37, where it is shown that for drain-to-source voltages above $v_{DS} > v_{GS} - v_T$, the drain current becomes constant, independent of v_{DS}. This mode of operation is the **constant-current state**. If v_{DS} exceeds a given **drain breakdown voltage**, BV_{DS} (usually between 20 and 50 V), avalanche breakdown occurs and the drain current increases substantially. Operation in this **breakdown state** can lead to permanent damage because of the excessive heat caused by the large drain current. Finally, if v_{GS} exceeds the **gate breakdown voltage** (around 50 V), permanent damage to the oxide layer can occur. It is important to know that it is possible to generate *static* voltages of magnitude sufficient to exceed this breakdown voltage just by handling the device; thus, some attention must usually be paid to static voltage buildup in handling MOS circuits.

Transistor Gates and Switches

Transistor switching circuits form the basis of digital logic circuits. The objective of this section is to discuss the internal operation of these circuits and to provide the reader interested in the internal workings of digital circuits with an adequate understanding of the basic principles.

An **electronic gate** is a device that, on the basis of one or more input signals, produces one of two or more prescribed outputs; as will be seen shortly, one can construct both digital and analog gates. A word of explanation is required, first, regarding the meaning of the words *analog* and *digital*. An analog voltage or current — or, more generally, an analog signal — is one that varies in a continuous fashion over time, in *analogy* (hence the expression *analog*) with a physical quantity. An example of an analog signal is a sensor voltage corresponding to ambient temperature on any given day, which may fluctuate between, say, 30° and 50°F. A digital signal, on the other hand, is a signal that can take only a finite number of values; in particular, a commonly encountered class of digital signals consists of **binary signals**, which can take only one of two values (for example, 1 and 0). A typical example of a binary signal would be the control signal for the furnace in a home heating system controlled by a conventional thermostat, where one can think of this signal as being "on" (or 1) if the temperature of the house has dropped below the thermostat setting (desired value), or "off" (or 0) if the house temperature is greater than or equal to the set temperature (say, 68°F). Figure 5.7.38 illustrates the appearance of the analog and digital signals in this furnace example.

Analog Gates

A common form of analog gate — probably the most important, in practice — employs an FET and takes advantage of the fact that current can flow in either direction in an FET biased in the ohmic region. Recall that the drain characteristic of the MOSFET consists of three regions: ohmic, active, and breakdown. A MOSFET amplifier is operated in the active region, where the drain current is nearly constant for any given value of v_{GS}. On the other hand, a MOSFET biased in the ohmic state acts very much as a linear resistor. For example, for an *n*-channel enhancement MOSFET, the conditions for the transistor to be in the ohmic region are

$$v_{GS} > V_T \quad \text{and} \quad |v_{DS}| \le \frac{1}{4}(v_{GS} - V_T) \tag{5.7.17}$$

As long as the FET is biased within these conditions, it acts simply as a linear resistor, and it can conduct current in either direction (provided that v_{DS} does not exceed the limits stated in Equation (5.7.17). In particular, the resistance of the channel in the ohmic region is found to be

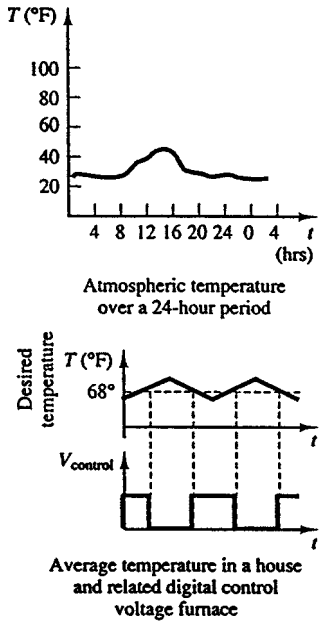


FIGURE 5.7.38 Illustration of analog and digital signals.

$$R_{DS} = \frac{V_T^2}{2I_{DSS}\left(v_{GS} - V_T\right)} \tag{5.7.18}$$

so that the drain current is equal to

$$i_D \approx \frac{v_{DS}}{r_{DS}} \quad \text{for} \quad \left|v_{DS}\right| \le \frac{1}{4}\left(v_{GS} - V_T\right) \quad \text{and} \quad v_{GS} > V_T \tag{5.7.19}$$

The most important feature of the MOSFET operating in the ohmic region, then, is that it acts as a voltage-controlled resistor, with the gate-source voltage, v_{GS}, controlling the channel resistance, R_{DS}. The use of the MOSFET as a switch in the ohmic region, then, consists of providing a gate-source voltage that can either hold the MOSFET in the cutoff region ($v_{GS} < V_T$) or bring it into the ohmic region. In this fashion, v_{GS} acts as a control voltage for the transistor.

Consider the circuit shown in Figure 5.7.39, where we presume that v_C can be varied externally and that v_{in} is some analog signal source that we may wish to connect to the load R_L at some appropriate time. The operation of the switch is as follows. When $v_C \le V_T$, the FET is in the cutoff region and acts as an open circuit. When $v_C > V_T$ (with a value of v_{GS} such that the MOSFET is in the ohmic region), the transistor acts as a linear resistance, R_{DS}. If $R_{DS} \ll R_L$, then $v_{\text{out}} \approx v_{\text{in}}$. By using a pair of MOSFETs, it is possible to improve the dynamic range of signals one can transmit through this analog gate.

MOSFET analog switches are usually produced in integrated circuit (IC) form and denoted by the symbol shown in Figure 5.7.40.

Digital Gates

BJT Gates. In discussing large-signal models for the BJT, we observed that the *i-v* characteristic of this family of devices includes a *cutoff* region, where virtually no current flows through the transistor. On the other hand, when a sufficient amount of current is injected into the base of the transistor, a bipolar transistor will reach *saturation*, and a substantial amount of collector current will flow. This behavior

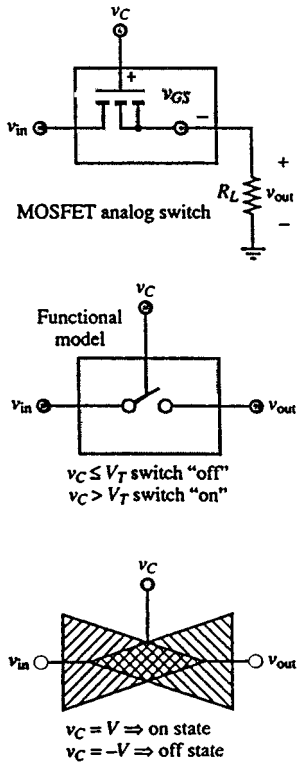


FIGURE 5.7.39 MOSFET analog switch.

FIGURE 5.7.40 Symbol for bilateral FET analog gate.

is quite well suited to the design of electronic gates and switches and can be visualized by superimposing a load line on the collector characteristic, as shown in Figure 5.7.41.

The operation of the simple **BJT switch** is illustrated in Figure 5.7.41 by means of load-line analysis. Writing the load-line equation at the collector circuit, we have

$$v_{CE} = V_{CC} - i_C R_C \tag{5.7.20}$$

and

$$v_{\text{out}} = v_{CE} \tag{5.7.21}$$

Thus, when the input voltage, v_{in}, is low (say, 0 V, for example) the transistor is in the cutoff region and little or no current flows, and

$$v_{\text{out}} = v_{CE} = V_{CC} \tag{5.7.22}$$

so that the output is "logic high."

When v_{in} is large enough to drive the transistor into the saturation region, a substantial amount of collector current will flow and the collector-emitter voltage will be reduced to the small saturation value, $V_{CE\,\text{sat}}$, which is typically a fraction of a volt. This corresponds to the point labeled B on the load line. For the input voltage v_{in} to drive the BJT of Figure 5.7.41 into saturation, a base current of approximately 50 mA will be required. Suppose, then, that the voltage v_{in} could take the values 0 or 5 V. Then, if $v_{\text{in}} = 0$ V, v_{out} will be nearly equal to V_{CC} or, again, 5 V. If, on the other hand, $v_{\text{in}} = 5$ V and R_B is, say, equal to 85 kΩ [so that the base current required for saturation flows into the base: $i_B \approx (5\text{ V} - 0.7\text{ V})/R_B = 50.6\ \mu\text{A}$], we have the BJT in saturation, and $v_{\text{out}} = V_{CE\,\text{sat}} \approx 0.2$ V.

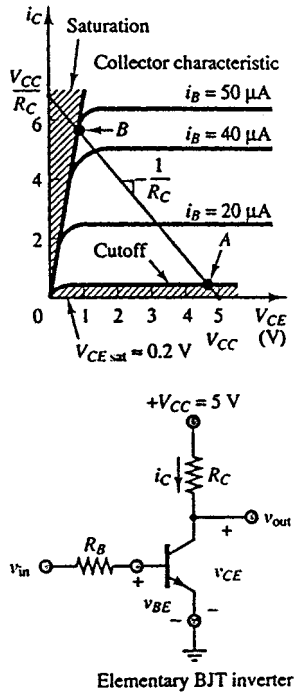


FIGURE 5.7.41 BJT switching characteristic.

Thus, you see that whenever v_{in} corresponds to a logic high (or logic 1), v_{out} takes a value close to 0 V, or logic low (or 0); conversely, v_{in} ="0" (logic "low") leads to v_{out} = "1". The values of 5 and 0 V for the two logic levels 1 and 0 are quite common in practice and are the standard values used in a family of logic circuits denoted by the acronym TTL, which stands for **transistor-transistor logic**. One of the more common TTL blocks is the **inverter** shown in Figure 5.7.41, so called because it "inverts" the input by providing a low output for a high input, and vice versa. This type of inverting, or "negative", logic behavior is quite typical of BJT gates (and of transistor gates in general).

MOSFET Logic Gates. Having discussed the BJT as a switching element, we might quite naturally suspect that FETs may similarly serve as logic gates. In fact, in some respects, FETs are better suited to be employed as logic stages than BJTs. The *n*-channel enhancement MOSFET serves as an excellent illustration: because of its physical construction, it is normally off (that is, it is off until a sufficient gate voltage is provided), and therefore it does not require much current from the input signal source. Further, MOS devices offer the additional advantage of easy fabrication into integrated circuit form, making production economical in large volume. On the other hand, MOS devices cannot provide as much current as BJTs, and their switching speeds are not quite as fast — although these last statements may not hold true for long, because great improvements are taking place in MOS technology. Overall, it is certainly true that in recent years it has become increasingly common to design logic circuits based on MOS technology. In particular, a successful family of logic gates called *CMOS* (for *complementary metal-oxide-semiconductor*) takes advantage of both *p*- and *n*-channel enchancement-mode MOSFETs to exploit the best features of both types of transistors. CMOS logic gates (and many other types of digital circuits constructed using the same technology) consume very little supply power and have become the mainstay in pocket calculators, wristwatches, portable computers, and many other consumer electronics products. Without delving into the details of CMOS technology, we shall briefly illustrate the properties of MOSFET logic gates and of CMOS gates in the remainder of this section.

Figure 5.7.42 depicts a MOSFET switch with its drain *i-v* characteristic. Note the general similarity with the switching characteristic of the BJT shown in the previous section. When the input voltage, v_{in},

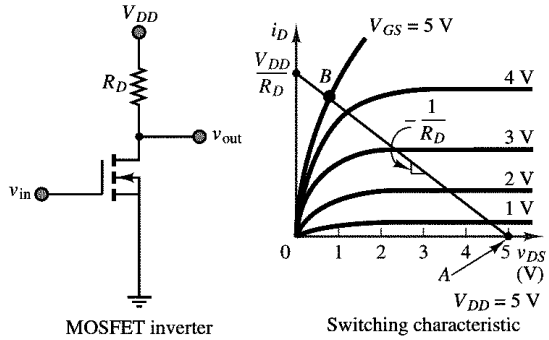


FIGURE 5.7.42 MOSFET switching characteristic.

is zero, the MOSFET conducts virtually no current and the output voltage, v_{out}, is equal to V_{DD}. When v_{in} is equal to 5 V, the MOSFET Q point moves from point A to point B along with the load line, with v_{DS} = 0.5 V. Thus, the circuit acts as an inverter. Much as in the case of the BJT, the inverter forms the basis of all MOS logic gates.

An elementary CMOS inverter is shown in Figure 5.7.43. Note first the simplicity of this configuration, which simply employs two enhancement-mode MOSFETs: *p*-channel at the top, denoted by the symbol Q_p, and *n*-channel at the bottom, denoted by Q_n. Recall from Chapter 8 that when v_{in} is low, transistor Q_n is off. However, transistor Q_p sees a gate-to-source voltage $v_{GS} = v_{in} - V_{DD} = -V_{DD}$; in a *p*-channel device, this condition is the counterpart of having $v_{GS} = V_{DD}$ for an *n*-channel MOSFET. Thus, when Q_n is off, Q_p is on and acts very much as a small resistance. In summary, when v_{in} is low, the output is $v_{out} \approx V_{DD}$. When v_{in} is high, the situation is reversed: Qn is now on and acts nearly as a short circuit while Q_p is open (since $v_{GS} = 0$ for Q_p). Thus, $v_{in} \approx 0$. The complementary MOS operation is depicted in Figure 5.7.43 in simplified form by showing each transistor as either a short or an open circuit, depending on it state. This simplified analysis is sufficient for the purpose of a qualitative analysis.

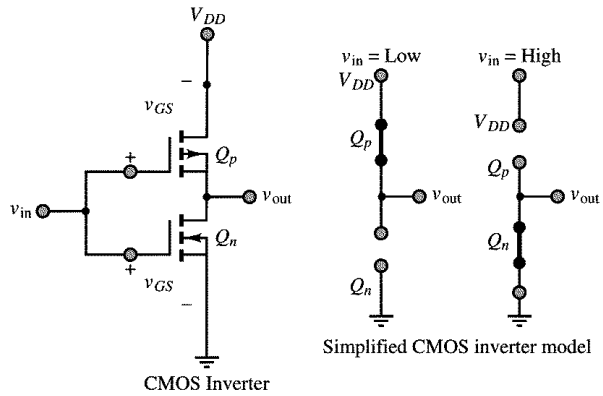


FIGURE 5.7.43 CMOS inverter.

5.8 Power Electronics

The objective of this section is to present a survey of power electronic devices and systems. Power electronic devices form the "muscle" of many electromechanical systems. For example, one finds such devices in many appliances, in industrial machinery, and virtually wherever an electric motor is found, since one of the foremost applications of power electronic devices is to supply and control the currents and voltages required to power electric machines, such as those introduced in Section 5.12.

Classification of Power Electronic Devices

Power semiconductors can be broadly subdivided into five groups: (1) power diodes, (2) thyristors, (3) power bipolar junction transistors (BJTs), (4) insulated-gate bipolar transistors (IGBTs), and (5) static induction transistors (SITs). Figure 5.8.1 depicts the symbols for the most common power electronic devices.

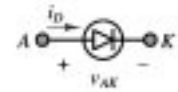
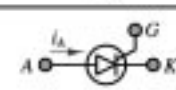
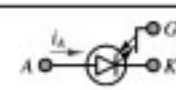
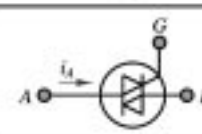
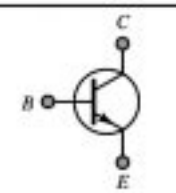
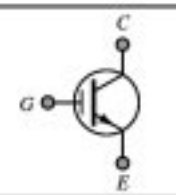
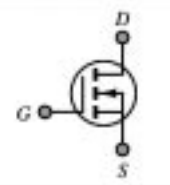
Device	Device symbol
Diode	
Thyristor	
Gate Turn-off Thyristor (GTO)	
Triac	
npn BJT	
IGBT	
n-channel MOSFET	

FIGURE 5.8.1 Classification of power electronic devices.

Power diodes are functionally identical to the diodes introduced in Section 5.7, except for their ability to carry much larger currents. A diode conducts in the forward-biased mode when the anode voltage ($V_{A)}$ is higher than the cathode voltage (V_K). Three types of power diodes exist: general-purpose, high-speed (fast-recovery), and Schottky. Typical ranges of voltage and current are 3000 V and 3500 A for general-purpose diodes and 3000 V and 1000 A for high-speed devices. The latter have switching times as low as a fraction of a microsecond. Schottky diodes can switch much faster (in the nanosecond range) but are limited to around 100 V and 300 A. The forward voltage drop of power diodes is not much higher than that of low-power diodes, being between 0.5 and 1.2 V. Since power diodes are used with rather large voltages, the forward bias voltage is usually considered negligible relative to other voltages in the circuit, and the switching characteristics of power diodes may be considered near ideal. The principal consideration in choosing power diodes is their power rating.

Thyristors function like power diodes with an additional gate terminal that controls the time when the device begins conducting; a thyristor starts to conduct when a small gate current is injected into the gate terminal, provided that the anode voltage is greater than the cathode voltage (or $V_{AK} > 0$ V). The forward voltage drop of a thyristor is of the order of 0.5 to 2 V. Once conduction is initiated, the gate current has no further control. To stop conduction, the device must be reverse-biased; that is, one must ensure that $V_{AK} \leq 0$ V. Thyristors can be rated at up to 6000 V and 3500 A. The **turn-off time** is an important characteristic of thyristors; it represents the time required for the device current to return to zero after external switching of V_{AK}. The fastest turn-off times available are in the range of 10 μsec; however, such turn-off times are achieved only in devices with slightly lower power ratings (1200 V, 1000 A). Thyristors can be subclassified into the following groups: force-commutated and line-commutated thyristors, gate turn-off thyristors (GTOs), reverse-conducting thyristors (RCTs), static induction thyristors (SITHs), gate-assisted turn-off thyristors (GATTs), light-activated silicon-controlled rectifiers (LASCSRs), and MOS-controlled thyristors (MCTs). It is beyond the scope of this chapter to go into a detailed description of each of these types of devices; their operation is typically a slight modification of the basic operation of the thyristor. The reader who wishes to gain greater insight into this topic may refer to one of a number of excellent books specifically devoted to the subject of power electronics.

Two types of thyristor-based device deserve some more attention. The **triac**, as can be seen in Figure 5.8.1, consists of a pair of thyristors connected back to back, with a single gate; this allows for current control in either direction. Thus, a triac may be thought of as bidirectional thyristor. The **gate turn-off** thyristor (GTO), on the other hand, can be turned on by applying a short positive pulse to the gate, like a thyristor, and can also be turned off by application of a short negative pulse. Thus, GTOs are very convenient in that they do not require separate commutation circuits to be turned on and off.

Power BJTs can reach ratings up to 1200 V and 400 A, and they operate in much the same way as a conventional BJT. Power BJTs are used in power converter applications at frequencies up to around 10 kHz. **Power MOSFETs** can operate at somewhat higher frequencies (a few to several tens of kHz), but are limited in power (typically up to 1000 V, 50 A). **Insulated-gate bipolar transistors** (IGBTs) are voltage-controlled (because of their insulated gate, reminiscent of insulated-gate FETs) power transistors that offer superior speed with respect to BJTs but are not quite as fast as power MOSFETs.

Classification of Power Electronic Circuits

The devices that will be discussed in the present chapter find application in a variety of **power electronic circuits**. This section will briefly summarize the principal types of power electronic circuits and will qualitatively describe their operation. The following sections will describe the devices and their operation in these circuits in more detail.

One possible classification of power electronic circuits is given in Table 5.8.1. Many of the types of circuits are similar to circuits that were introduced in earlier chapters. Voltage regulators were introduced in Section 5.7. Power electronic switches function exactly like the transistor switches described in Section 5.7; their function is to act as voltage- or current-controlled switches to turn AC or DC supplies on and

TABLE 5.8.1 Power Electronic Circuits

Circuit Type	Essential Features
Voltage regulators	Regulate a DC supply to a fixed voltage output
Power amplifiers	Large-signal amplification of voltages and currents
Switches	Electronic switches (for example, transistor switches)
Diode rectifier	Converts fixed AC voltage (single- or multiphase) to fixed DC voltage
AC-DC converter (controlled rectifier)	Converts fixed AC voltage (single- or multiphase) to variable DC voltage
AC-AC converter (AC voltage controller)	Converts fixed AC voltage to variable AC voltage (single- or multiphase)
DC-DC converter (chopper)	Converts fixed DC voltage to variable DC voltage
DC-AC converter (inverter)	Converts fixed DC voltage to variable AC voltage (single- or multiphase)

off. Transistor power amplifiers are the high-power version of the BJT and MOSFET amplifiers mentioned in Section 5.7.

Diode rectifiers were discussed in Section 5.7 in their single-phase form; similar rectifiers can also be designed to operate with three-phase sources. The operation of a single-phase full-wave rectifier was summarized in Figure 5.8.2. AC-DC converters are also rectifiers, but they take advantage of the controlled properties of thyristors. The thyristor gate current can be timed to "fire" conduction at variable times, resulting in a variable DC output, as illustrated in Figure 5.8.2, which shows the circuit and behavior of a single-phase AC-DC converter. This type of converter is very commonly used as a supply for DC electric motors. In Figure 5.8.2 α is the firing angle of thyristor T_1, where the device starts to conduct.

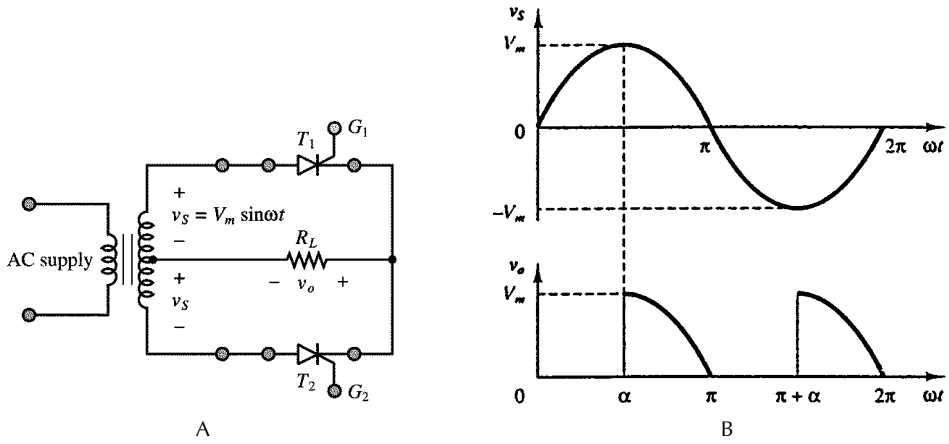


FIGURE 5.8.2 AC-DC converter circuit and waveform.

AC-AC converters are used to obtain a variable AC voltage from a fixed AC source. Figure 5.8.3 shows a triac-based AC-AC converter, which takes advantage of the bidirectional capability of triacs to control the rms value of an alternating voltage. Note in particular that the resulting AC waveform is no longer a pure sinusoid even though its fundamental period (frequency) is unchanged. A DC-DC converter, also known as a *chopper*, or *switching regulator*, permits conversion of a fixed DC source to a variable DC supply. Figure 5.8.4 shows how such an effect may be obtained by controlling the base-emitter voltage of a bipolar transistor, enabling conduction at the desired time. This results in the conversion of the DC input voltage to a variable-duty-cycle output voltage, whose average value can be controlled by selecting the "on" time of the transistor. DC-DC converters find application as variable voltage supplies for DC electric motors used in electric vehicles.

Finally, DC-AC supplies, or inverters, are used to convert a fixed DC supply to a variable AC supply; they find application in AC motor control. The operation of these circuits is rather complex; it is illustrated

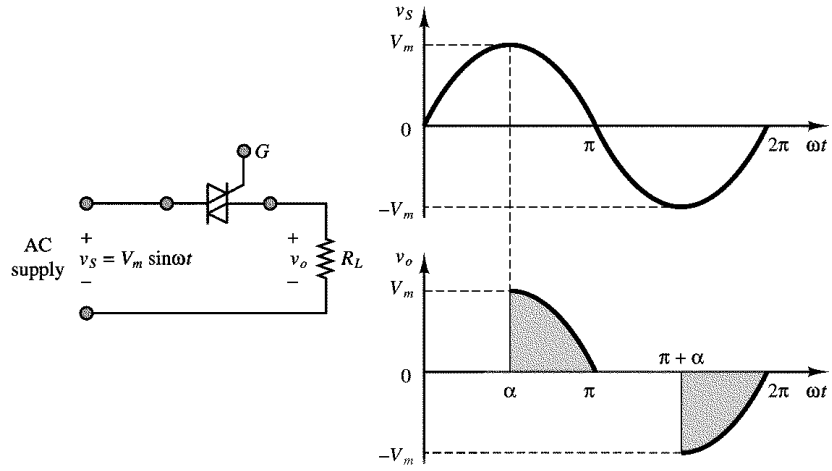


FIGURE 5.8.3 AC-AC converter circuit and waveform.

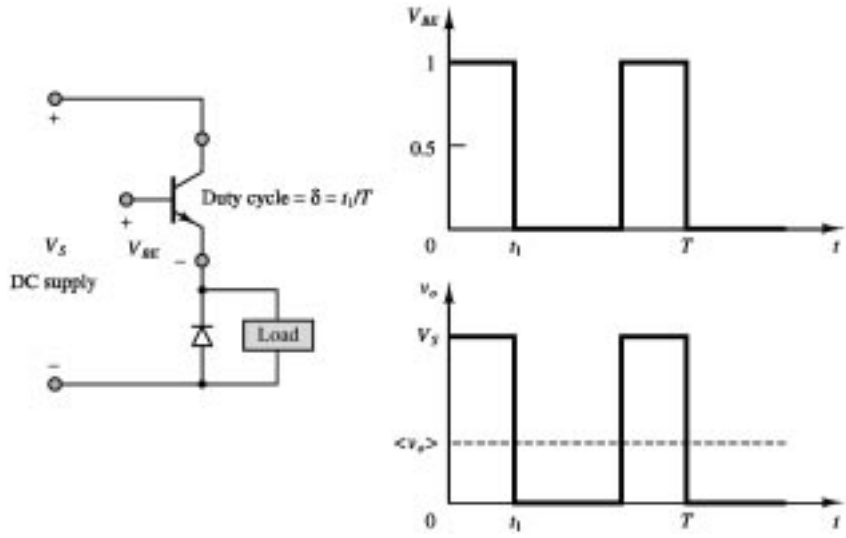


FIGURE 5.8.4 DC-DC converter circuit and waveform.

conceptually in the waveforms of Figure 5.8.5, where it is shown that by appropriately switching two pairs of transistors it is possible to generate an alternating current waveform (square wave).

Each of the circuits of Table 5.8.1 will now be analyzed in greater detail.

Rectifiers and Controlled Rectifiers (AC-DC Converters)

Thyristors and Controlled Rectifiers

In a number of applications, it is useful to be able to externally control the amount of current flowing from an AC source to the load. A family of power semiconductor devices called **controlled rectifiers** allows for control of the rectifiers state by means of a third input, called the gate. Figure 5.8.6 depicts the appearance of a **thyristor**, or **silicon-controlled rectifier** (SCR), illustrating how the physical structure of this device consists of four layers, alternating *p*-type and *n*-type material. Note that the circuit symbol for the thyristor suggests that this device acts as a diode, with provision for an additional external control signal.

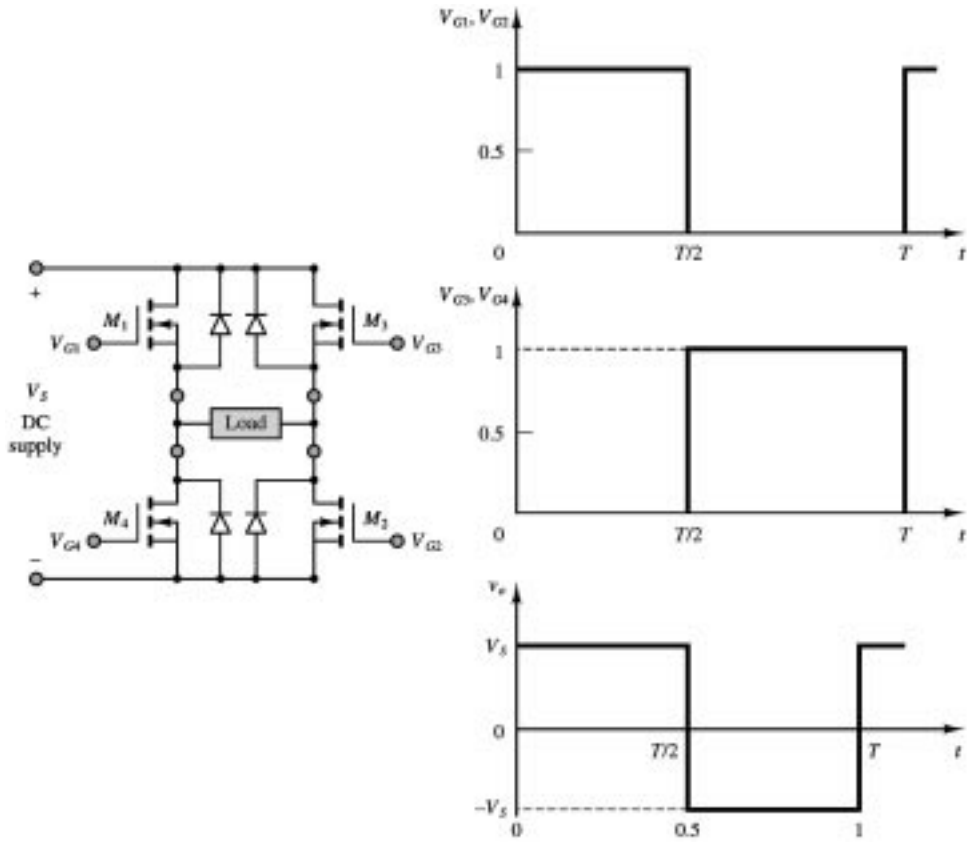


FIGURE 5.8.5 DC-AC converter circuit and waveform.

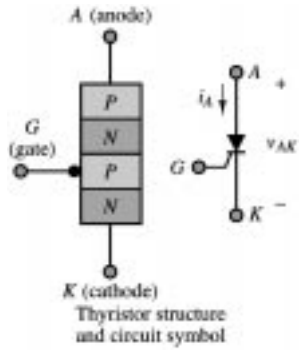


FIGURE 5.8.6 Thyristor structure and circuit symbol.

The operation of the thyristor can be explained in an intuitive fashion as follows. When the voltage v_{AK} is negative (i.e., providing reverse bias), the thyristor acts just like a conventional *pn* junction in the off state. When v_{AK} is forward-biased *and* a small amount of current is injected into the gate, the thyristor conducts forward current. The thyristor then continues to conduct (even in the absence of gate current), provided that v_{AK} remains positive. Figure 5.8.7 depicts the *i-v* curve for the thyristor. Note that the thyristor has two stable states, determined by the bias v_{AK} and by the gate current. In summary, the thyristor acts as a diode with a control gate that determines the time when conduction begins.

A somewhat more accurate description of thyristor operation may be provided if we realize that the four-layer *pnpn* device can be modeled as a *pnp* transistor connected to an *npn* transistor. Figure 5.8.8 clearly shows that, physically, this is a realistic representation. Note that the anode current, i_A, is equal

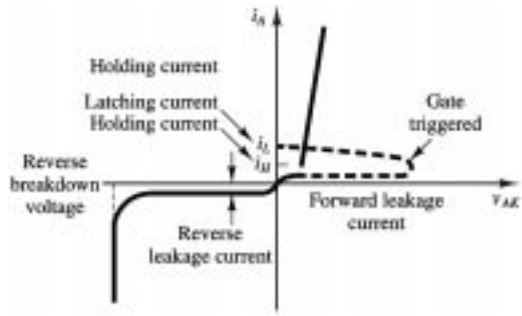


FIGURE 5.8.7 Thyristor *i-v* characteristic.

to the emitter current of the *pnp* transistor (labeled Q_p) and the base current of Q_p is equal to the collector current of the *npn* transistor, Q_n. Likewise, the base current of Q_n is the sum of the gate current and the collector current of Q_p. The behavior of this transistor model is explained as follows. Suppose, initially, i_G and i_{B_n} are both zero. Then it follows that Q_n is in cutoff, and therefore $i_{C_n} = 0$. But if $i_{C_n} = 0$, then the base current going into Q_p is also zero and Q_p is also in cutoff, and $i_{C_n} = 0$, consistent with our initial assumption. Thus, this is a stable state, in the sense that unless an external condition perturbs the thyristor, it will remain off.

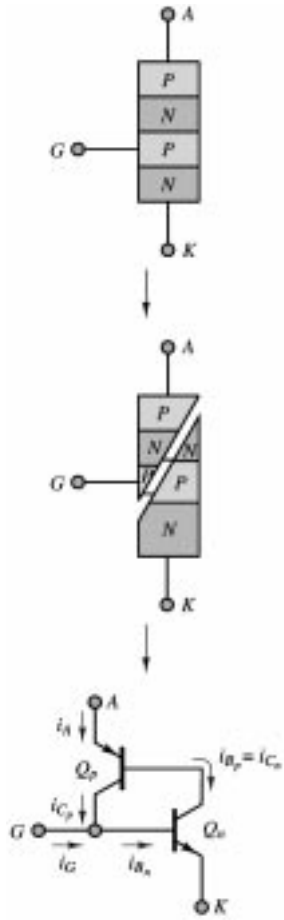


FIGURE 5.8.8 Thyristor two-transistor model.

Now, suppose a small pulse of current is injected at the gate. Then $i_{B_n} > 0$ and Q_n starts to conduct, provided, of course, that $v_{AK} > 0$. At this point, i_{C_n}, and therefore i_{B_p}, must be greater than zero, so that Q_p conducts. It is important to note that once the gate current has turned Q_n on, Q_p also conducts, so that $i_{C_p} > 0$. Thus, even though i_G may cease, once this "on" state is reached, $i_{C_p} = i_{B_n}$ continues to drive Q_n so that the on state is also self-sustaining. The only condition that will cause the thyristor to revert to the off state is the condition in which v_{AK} becomes negative; in this case, both transistors return to the cutoff state.

In a typical controlled rectifier application, the device is used as a half-wave rectifier that conducts only after a trigger pulse is applied to the gate. Without concerning ourselves with how the trigger pulse is generated, we can analyze the general waveforms for the circuit of Figure 5.8.9 as follows. Let the voltage v_{trigger} be applied to the gate of the thyristor at $t = \tau$. The voltage v_{trigger} can be a short pulse, provided by a suitable trigger-timing circuit (Section 5.11 will discuss timing and switching circuits). At $t = \tau$, the thyristor begins to conduct, and it continues to do so until the AC source enters its negative cycle. Figure 5.8.10 depicts the relevant waveforms.

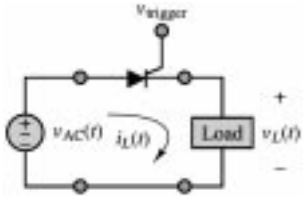


FIGURE 5.8.9 Controlled rectifier circuit.

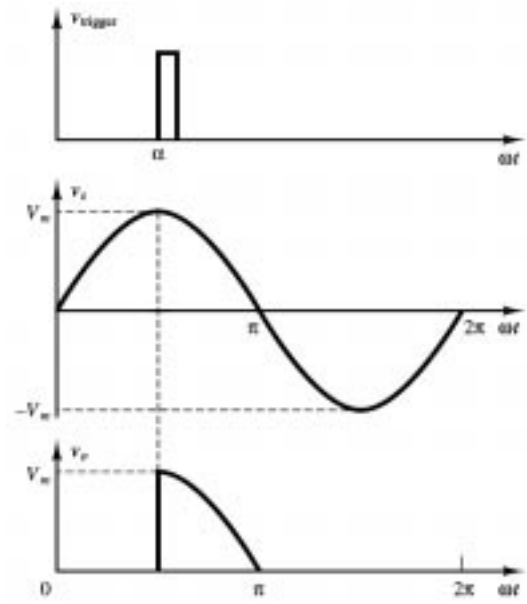


FIGURE 5.8.10 Half-wave controlled rectifier waveforms.

Note how the DC load voltage is controlled by the firing time τ, according to the following expression:

$$\langle v_L \rangle = V_L = \frac{1}{T}\int_{\tau}^{T/2} v_{\text{AC}}(t)\, dt \tag{5.8.1}$$

where T is the period of $v_{AC}(t)$, Now, if we let

$$v_{AC}(t) = A\sin\omega t \tag{5.8.2}$$

we can express the average (DC) value of the load voltage

$$V_L = \frac{1}{T}\int_{\tau}^{T/2} A\sin\omega t\, dt = (1+\cos\omega t)\frac{A}{2\pi} \tag{5.8.3}$$

in terms of the **firing angle**, α, defined as

$$\alpha = \omega t \tag{5.8.4}$$

By evaluating the integral of Equation (5.8.3), we can see that the (DC) load voltage amplitude depends on the firing angle, α:

$$V_L = (1+\cos\alpha)\frac{A}{2\pi} \tag{5.8.5}$$

Electric Motor Drives

The advent of high-power semiconductor devices has made it possible to design effective and relatively low-cost electronic supplies that take full advantage of the capabilities of the devices introduced in this chapter. Electronic power supplies for DC and AC motors have become one of the major fields of application of power electronic devices. The last section of this chapter is devoted to an introduction to two families of power supplies, or **electric drives: choppers**, or **DC-DC converters**; and **inverters**, or **DC-AC converters**. These circuits find widespread application in the control of AC and DC motors in a variety of applications and power ranges.

Before we delve into the discussion of the electronic supplies, it will be helpful to introduce the concept of quadrants of operation of a drive. Depending on the direction of current flow and on the polarity of the voltage, an electronic drive can operate in one of four possible modes, as indicated in Figure 5.8.11.

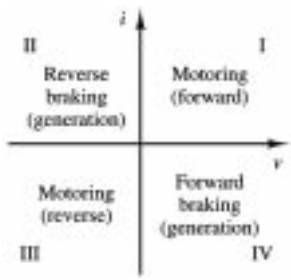


FIGURE 5.8.11 The four quadrants of an electric drive.

Choppers (DC-DC Converters)

As the name suggests, a DC-DC converter is capable of converting a fixed DC supply to a variable DC supply. This feature is particularly useful in the control of the speed of a DC motor (described in greater detail in Section 5.12). In a DC motor, shown schematically in Figure 5.8.12, the developed torque, T_m, is proportional to the current supplied to the motor **armature**, I_a, while the **electromotive force** (emf), E_a, which is the voltage developed across the armature, is proportional to the speed of rotation of the motor, ω_m. A DC motor is an electromechanical energy-conversion system; that is, it converts electrical to mechanical energy (or vice versa if it is used as a generator); if we recall that the product of torque and speed is equal to power in the mechanical domain, and that current times voltage is equal to power in the electrical domain, we conclude that in the ideal case of complete energy conversion, we have

$$E_a \times I_a = T_m \times \omega_m \tag{5.8.6}$$

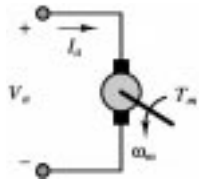


FIGURE 5.8.12 DC motor.

Naturally, such ideal energy conversion cannot take place; however, we can see that there is a correspondence between the four-electrical quadrants of Figure 5.8.11 and the mechanical power output of the motor: namely, if the voltage and current are both positive or both negative, the electrical power will be positive, and so will the mechanical power. This corresponds to the **forward** (*i, v* both positive) and **reverse** (*i,v* both negative) **motoring** operation. Forward motoring corresponds to quadrant I, and reverse motoring to quadrant III in Figure 5.8.12. If the voltage and current are of opposite polarity (quadrants II and IV), electrical energy is flowing back to the electric drive; in mechanical terms this corresponds to a braking condition. Operation in the fourth quadrant can lead to **regenerative braking**, so called because power is generated by making current flow back to the source. This mode could be useful, for example, to recharge a battery supply, because the braking energy can be regenerated by returning it to the electric supply.

A simple circuit that can accomplish the task of providing a variable DC supply from a fixed DC source is the **step-down chopper (buck converter)**, shown in Figure 5.8.13. The circuit consists of a "chopper" switch, denoted by the symbol *S*, and a free-wheeling diode. The switch can be any of the power switches described in this chapter, for example, a power BJT or MOSFET, or a thyristor; see, for example, the BJT switch of Figure 5.8.4. The circuit to the right of the diode is a model of a DC motor, including the inductance and resistance of the armature windings and the effect of the back emf E_a. When the switch is turned on (say, at $t = 0$), the supply V_s is connected to the load and $v_o = V_s$. The load current, i_o, is determined by the motor parameters. When the switch is turned off, the load current continues to flow through the free-wheeling diode, but the output voltage is now $v_o = 0$. At time T, the switch is turned on again, and the cycle repeats.

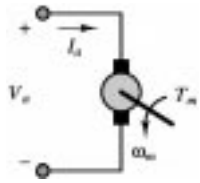


FIGURE 5.8.13 Step-down chopper (buck converter).

Figure 5.8.14 depicts the v_o and i_o waveforms. The average value of the output voltage, $\langle v_o \rangle$, is given by the expression

$$\langle v_o \rangle = \frac{t_1}{T} V_S = \delta V_S \tag{5.8.7}$$

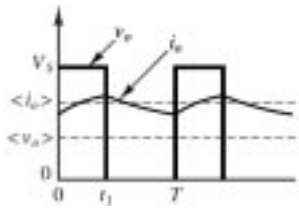


FIGURE 5.8.14 Step-down chopper waveforms.

where δ is the **duty cycle** of the chopper. The step-down chopper has a useful range,

$$0 \le \langle v_o \rangle \le V_S \tag{5.8.8}$$

It is also possible to increase the range of a DC-DC converter to above the supply voltage by making use of the energy-storage properties of an inductor; the resulting circuit is shown in Figure 5.8.15. When the chopper switch, *S*, is on, the supply current flows through the inductor and the closed switch, storing energy in the inductor; the output voltage, v_o, is zero, since the switch is a short circuit. When the switch is open, the supply current will flow through the load via the diode; but the inductor voltage is negative during the transient following the opening of the switch and therefore adds to the source voltage: the energy stored in the inductor while the switch was closed is now released and transferred to the load. This stored energy makes it possible for the output voltage to be higher than the supply voltage for a finite period of time.

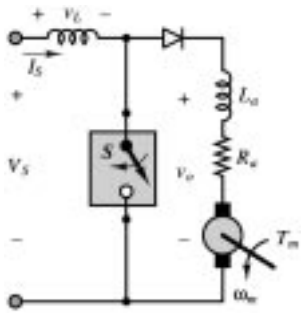


FIGURE 5.8.15 Step-up chopper (boost converter).

Let t_1 once again be the time during which the chopper conducts; neglecting for the moment the ripple in the supply current, the energy stored in the inductor during this period is

$$W_i = V_S I_S t_1 \tag{5.8.9}$$

When the chopper is off, the energy released to the load is

$$W_i = \left(\langle v_o \rangle - V_S\right) I_S \left(T - t_1\right) \tag{5.8.10}$$

If the system is lossless, the two energy expressions must be equal:

$$V_S I_S t_1 = \left(\langle v_o \rangle - V_S\right) I_S \left(T - t_1\right) \tag{5.8.11}$$

and we can determine the average output voltage to be

$$\langle v_o \rangle = V_S \frac{t_1 + T - t_1}{T - t_1} = V_S \frac{T}{T - t_1} = V_S \frac{1}{1 - \delta} \tag{5.8.12}$$

Since the duty cycle, δ, is always less than 1, the theoretical range of the supply is

$$V_S \le \langle v_o \rangle < \infty \tag{5.8.13}$$

The waveforms for the boost converter are shown in Figure 5.8.16.

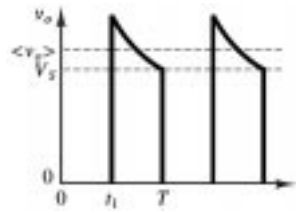


FIGURE 5.8.16 Step-up chopper output voltage waveform.

A step-up chopper can also be used to provide regenerative braking: if the "supply" voltage is the motor armature voltage and the output voltage is the fixed DC supply (battery) voltage, then power can be made to flow from the motor to the DC supply (i.e., recharging the battery). This configuration is shown in Figure 5.8.17.

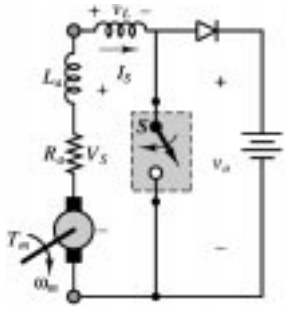


FIGURE 5.8.17 Step-up chopper used for regenerative braking.

Finally, the operation of the step-down and step-up choppers can be combined into a **two-quadrant chopper**, shown in Figure 5.8.18. The circuit shown schematically in Figure 5.8.18 can provide both regenerative braking and motoring operation in a DC motor. When switch S_2 is open and switch S_1 serves as a chopper, the circuit operates as a step-down chopper, precisely as was described earlier in this section (convince yourself of this by redrawing the circuit with S_2 and D_2 replaced by open circuits). Thus, the drive and motor operate in the first quadrant (motoring operation). The output voltage, v_o, will switch between V_S and zero, as shown in Figure 5.8.14, and the load current will flow in the direction indicated by the arrow in Figure 5.8.18 diode D_1 free-wheels whenever S_1 is open. Since both output voltage and current are positive, the system operates in the first quadrant.

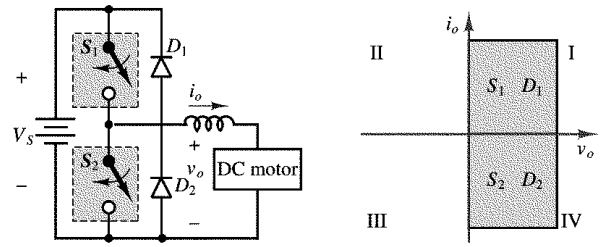


FIGURE 5.8.18 Two-quadrant chopper.

When switch S_1 is open and switch S_2 serves as a chopper, the circuit resembles a step-up chopper. The source is the motor emf, E_a, and the load is the battery; this is the situation depicted in Figure 5.8.17. The current will now be negative, since the sum of the motor emf and the voltage across the inductor (corresponding to the energy stored during the "on" cycle of S_2) is greater than the battery voltage. Thus, the drive operates in the fourth quadrant.

Inverters (DC-AC Converters)

As will be explained in Section 5.12, variable-speed drives for AC motors require a multiphase variable-frequency, variable-voltage supply. Such drives are called *DC-AC converters*, or *inverters*. Inverter circuits can be quite complex, so the objective of this section is to present a brief introduction to the subject, with the aim of illustrating the basic principles. A **voltage source inverter** (VSI) converts the output of a fixed DC supply (e.g., a battery) to a variable-frequency AC supply. Figure 5.8.19 depicts a **half-bridge VSI**; once again, the switches can be either bipolar or MOS transistors, or thyristors. The operation of this circuit is as follows. When switch S_1 is turned on, the output voltage is in the positive half-cycle, and $v_o = V_S/2$. The switching sequence of S_1 and S_2 is shown in Figure 5.8.20. It is important that each switch be turned off before the other is turned on; otherwise, the DC supply would be short-circuited. Since the load is always going to be inductive in the case of a motor drive, it is important to observe that the load current, i_o, will lag the voltage waveform, as shown in Figure 5.8.20. As shown in this figure, there will be some portions of the cycle in which the voltage is positive but the current is negative. The function of diodes D_1 and D_2 is precisely to conduct the load current whenever it is of direction opposite to the polarity of the voltage. Without these diodes, there would be no load current in this case. Figure 5.8.20 also shows which element is conducting in each portion of the cycle.

A full-bridge version of the VSI can also be designed as shown in Figure 5.8.21; the associated output voltage waveform is shown in Figure 5.8.22. A 3-phase VSI is shown in Figure 5.8.23.

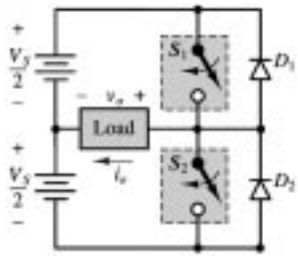


FIGURE 5.8.19 Half-bridge voltage source inverter.

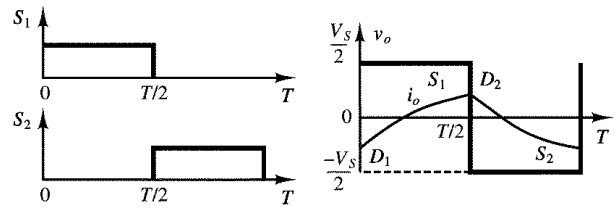


FIGURE 5.8.20 Half-bridge voltage source inverter waveforms.

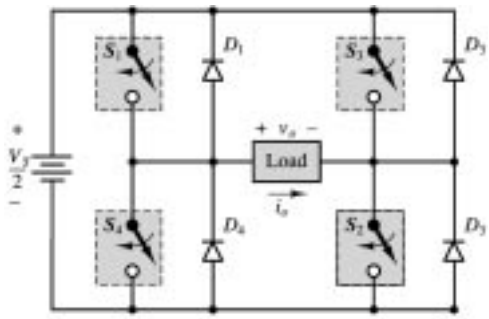


FIGURE 5.8.21 Full-bridge voltage source inverter.

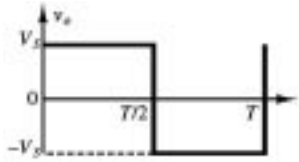


FIGURE 5.8.22 Half-bridge voltage source inverter output voltage waveform.

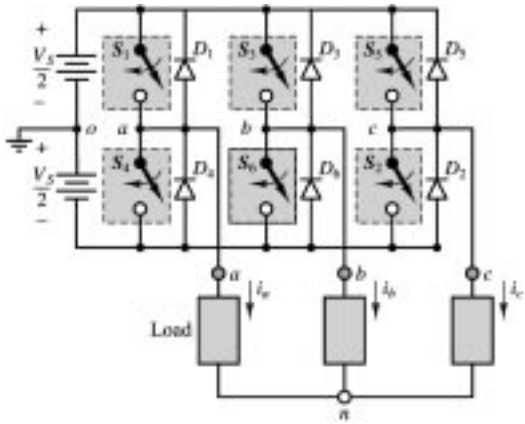


FIGURE 5.8.23 Three-phase voltage source inverter.

5.9 Operational Amplifiers

In this section we will analyze the properties of a general-purpose amplifier circuit known as the *operational amplifier.*

The Operational Amplifier

An **operational amplifier** is an **integrated circuit**, that is, a large collection of individual electrical and electronic circuits integrated on a single silicon wafer. An operational amplifier — or op-amp — can perform a great number of operations, such as addition, filtering, or integration, which are all based on the properties of ideal amplifiers and of ideal circuit elements. The introduction of the operational amplifier in integrated circuit form marked the beginning of a new era in modern electronics. Since the introduction of the first IC op-amp, the trend in electronic instrumentation has been to move away from the discrete (individual-component) design of electronic circuits, toward the use of integrated circuits for a large number of applications. This statement is particularly true for applications of the type the nonelectrical engineer is likely to encounter: op-amps are found in most measurement and instrumentation applications, serving as extremely versatile building blocks for any application that requires the processing of electrical signals.

In the following pages, simple circuit models of the op-amp will be introduced. The simplicity of the models will permit the use of the op-amp as a circuit element, or building block, without the need to describe its internal workings in detail. Integrated circuit technology has today reached such an advanced stage of development that it can be safely stated that for the purpose of many instrumentation applications, the op-amp can be treated as an ideal device. Following the introductory material presented in this chapter, more advanced instrumentation applications will be explored in Section 5.11.

The Open-Loop Model

The ideal operational amplifier behaves very much as an ideal **differential amplifier**, that is, a device that amplifies the difference between two input voltages. Operational amplifiers are characterized by near-infinite input resistance and very small output resistance. As shown in Figure 5.9.1 the output of the op-amp is an amplified version of the difference between the voltages present at the two inputs:

$$v_{\text{out}} = A_{V(OL)}\left(v^{+} - v^{-}\right) \tag{5.9.1}$$

The input denoted by a positive sign is called the **noninverting input** (or terminal), while that represented with a negative sign is termed the **inverting input** (or terminal). The amplification factor, or gain, $A_{V(OL)}$, is called the **open-loop voltage gain** and is quite large by design, typically of the order of 10^5 to 10^7; it will soon become apparent why a large open-loop gain is a desirable characteristic. Together with the high input resistance and low output resistance, the effect of a large amplifier open-loop voltage gain, $A_{V(OL)}$, is such that op-amp circuits can be designed to perform very nearly as ideal voltage or current amplifiers. In effect, to analyze the performance of an op-amp circuit, only one assumption will be needed: that the current flowing into the input circuit of the amplifier is zero, or

$$i_{\text{in}} = 0 \tag{5.9.2}$$

This assumption is justified by the large input resistance and large open-loop gain of the operational amplifier. The model just introduced will be used to analyze three amplifier circuits in the next part of this section.

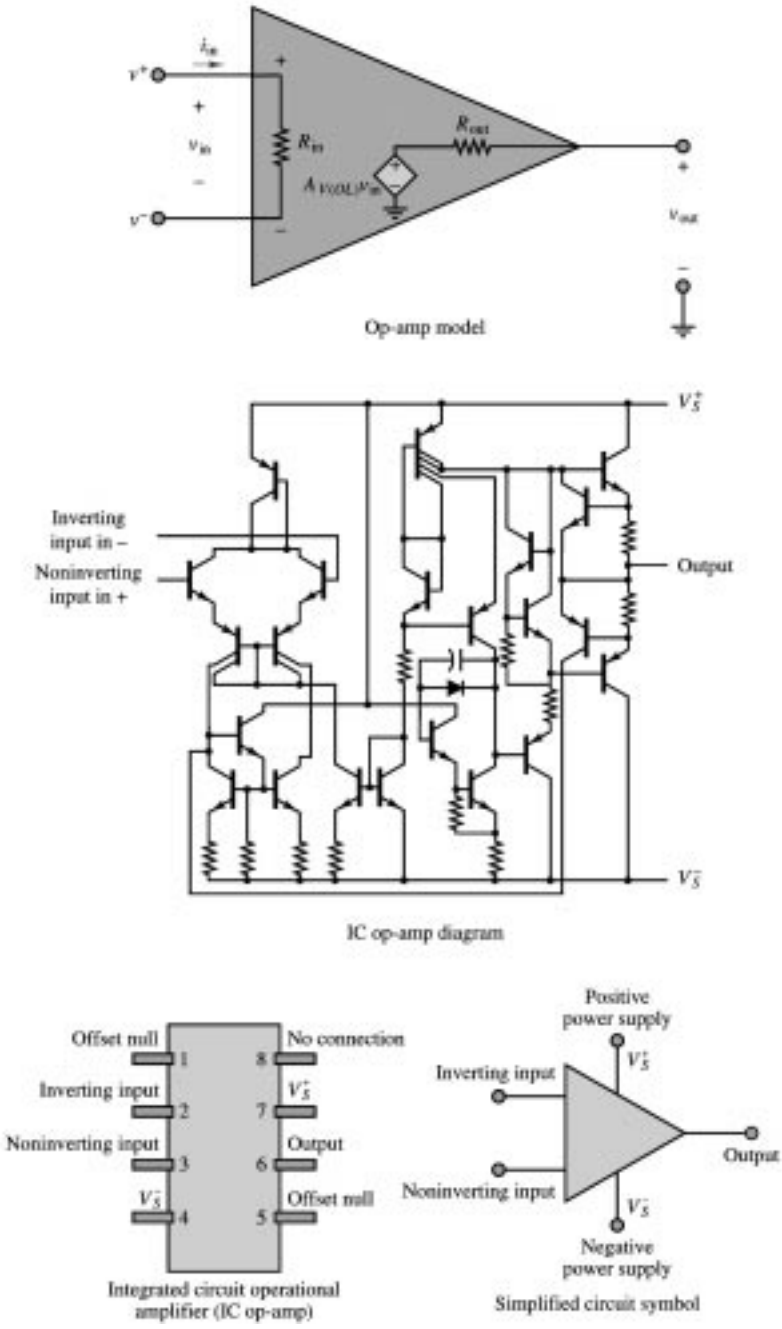


FIGURE 5.9.1 Operational amplifier model symbols, and circuit diagram.

The Operational Amplifier in the Closed-Loop Mode

The Inverting Amplifier. One of the more popular circuit configurations of the op-amp, because of its simplicity, is the so-called inverting amplifier, shown in Figure 5.9.2.

$$v_S = -v_{\text{out}}\left(\frac{1}{R_F/R_S} + \frac{1}{A_{V(OL)}R_F/R_S} + \frac{1}{A_{V(OL)}}\right) \tag{5.9.3}$$

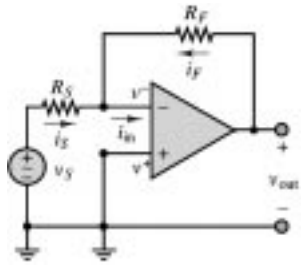


FIGURE 5.9.2 Inverting amplifier.

If the open-loop gain of the amplifier, $A_{V(OL)}$, is sufficiently large, the terms $1/(A_{V(OL)}R_F/R_S)$ and $1/A_{V(OL)}$, are essentially negligible, compared with $1/(R_F/R_S)$. As stated earlier, typical values of $A_{V(OL)}$ range from 10^5 to 10^7, and thus it is reasonable to conclude that, to a close approximation, the following expression describes the closed-loop gain of the inverting amplifier:

$$v_{\text{out}} = -\frac{R_F}{R_S}v_S \quad \text{Inverting amplifier closed-loop gain} \tag{5.9.4}$$

Next, we show that by making an additional assumption it is possible to simplify the analysis considerably. Consider that, as was shown for the inverting amplifier, the inverting terminal voltage is given by

$$v^- = -\frac{v_{\text{out}}}{A_{V(OL)}} \tag{5.9.5}$$

Clearly, as $A_{V(OL)}$ approaches infinity, the inverting-terminal voltage is going to be very small (practically, of the order of microvolts). It may then be assumed that *in the inverting amplifier*, v^- is virtually zero:

$$v^- \approx 0 \tag{5.9.6}$$

This assumption prompts an interesting observation (which may not yet appear obvious at this point): the effect of the feedback connection from output to inverting input is to force the voltage at the inverting input to be equal to that at the noninverting input.

This is equivalent to stating that for an op-amp *with negative feedback*,

$$v^- \approx v^+ \tag{5.9.7}$$

The analysis of the operational amplifier can now be greatly simplified if the following two assumptions are made:

1. $i_{in} = 0$
2. $v^- = v^+$ (5.9.8)

This technique will be tested in the next subsection by analyzing a noninverting amplifier configuration.

A useful op-amp circuit that is based on the inverting amplifier is the **op-amp summer**, or **summing amplifier**. This circuit, shown in Figure 5.9.3, is used to add signal sources. The primary advantage of using the op-amp as a summer is that the summation occurs independently of load and source impedances, so that sources with different internal impedances will not interact with each other.

$$v_{out} = -\sum_{n=1}^{N} \frac{R_F}{R_{S_n}} v_{S_n} \tag{5.9.9}$$

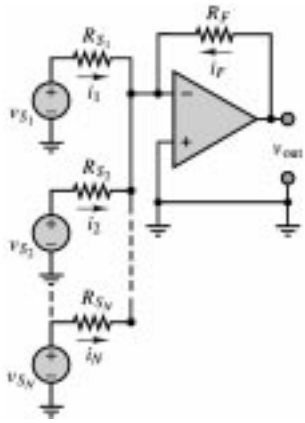


FIGURE 5.9.3 Summing amplifier.

That is, the output consists of the weighted sum of N input signal sources, with the weighting factor for each source equal to the ratio of the feedback resistance to the source resistance.

The Noninverting Amplifier. To avoid the negative gain (i.e., phase inversion) introduced by the inverting amplifier, a **noninverting amplifier** configuration is often employed. A typical noninverting amplifier is shown in Figure 5.9.4; note that the input signal is applied to the noninverting terminal this time.

$$\frac{v_{out}}{v_S} = 1 + \frac{R_F}{R_S} \quad \text{Noninverting amplifier closed-loop gain}$$

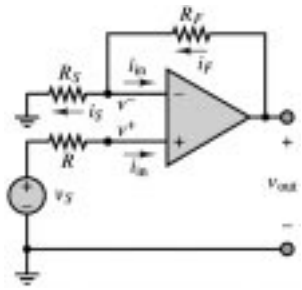


FIGURE 5.9.4 Noninverting amplifier.

The Differential Amplifier. The third closed-loop model examined in this chapter is a combination of the inverting and noninverting amplifiers; it finds frequent use in situations where the difference between two signals needs to be amplified. The basic **differential amplifier** circuit is shown in Figure 5.9.5,

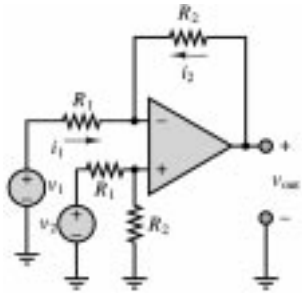


FIGURE 5.9.5 Differential amplifier.

where the two sources, v_1 and v_2, may be independent of each other, or may originate from the same process, as they do in Example 5.9.1.

The following expression for the output voltage is obtained:

$$v_{\text{out}} = R_2\left[\frac{-v_1}{R_1} + \frac{1}{R_1 + R_2}v_2 + \frac{R_2}{R_1\left(R_1 + R_2\right)}v_2\right]$$

$$= \frac{R_2}{R_1}\left(v_2 - v_1\right)$$

In practice, it is often necessary to amplify the difference between two signals that are both corrupted by noise or some other form of interference. In such cases, the differential amplifier provides an invaluable tool in amplifying the desired signal while rejecting the noise. Example 5.9.1 provides a realistic look at a very common application of the differential amplifier.

Example 5.9.1 An EKG Amplifier

This example illustrates the principle behind a two-lead electrocardiogram (EKG) measurement. The desired cardiac waveform is given by the difference between the potentials measured by two electrodes suitably placed on the patient's chest, as shown in Figure 5.9.6. A healthy, noise-free EKG waveform, $v_1 - v_2$, is shown in Figure 5.9.7.

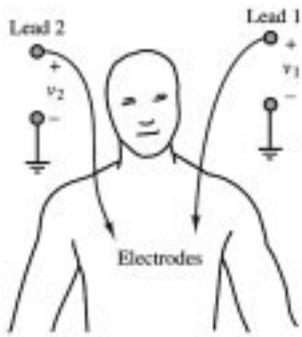


FIGURE 5.9.6 Two-lead electrocardiogram.

FIGURE 5.9.7 EKG waveform.

Unfortunately, the presence of electrical equipment powered by the 60-Hz, 110-VAC line current causes undesired interference at the electrode leads: the lead wires act as antennas and pick up some of the 60-Hz signal in addition to the desired EKG voltage. In effect, instead of recording the desired EKG signals, v_1 and v_2, the two electrodes provide the following inputs to the EKG amplifier, shown in Figure 5.9.8.

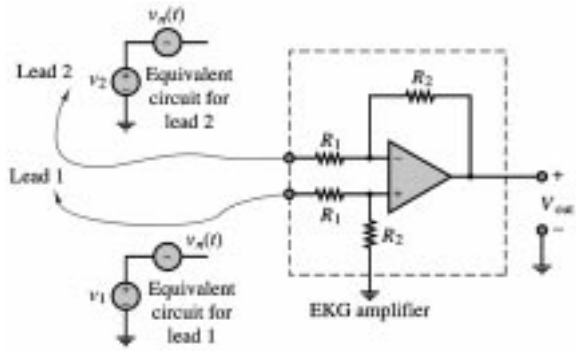


FIGURE 5.9.8 EKG amplifier.

Lead 1:

$$v_1(t) + v_n(t) = v_1(t) + V_n \cos\left(377t + \phi_n\right)$$

Lead 2:

$$v_2(t) + v_n(t) = v_2(t) + V_n \cos\left(377t + \phi_n\right)$$

The interference signal, $V_n \cos(377t + \phi n)$, is approximately the same at both leads, because the electrodes are chosen to be identical (e.g., they have the same lead lengths) and are in close proximity to each other. Further, the nature of the interference signal is such that it is common to both leads, since it is a property of the environment the EKG instrument is embedded in. On the basis of the analysis presented earlier, then,

$$v_{\text{out}} = \frac{R_2}{R_1}\left[\left(v_1 + v_n(t)\right) - \left(v_2 + v_n(t)\right)\right]$$

or

$$v_{\text{out}} = \frac{R_2}{R_1}\left(v_1 - v_2\right)$$

Thus, the differential amplifier nullifies the effect of the 60-Hz interference, while amplifying the desired EKG waveform.

The preceding example introduced the concept of so-called **common-mode** and **differential-mode signals**. In the EKG example, the desired differential-mode EKG signal was amplified by the op-amp while the common-mode disturbance was canceled. Thus, the differential amplifier provides the ability to reject common-mode signal components (such as noise or undesired DC offsets) while amplifying the differential-mode components. This is a very desirable feature in instrumentation systems. In practice, rejection of the common-mode signal is not always complete: some of the common-mode signal

component will always appear in the output. This fact gives rise to a figure of merit called the *common-mode rejection ratio*, which is discussed later in this section.

Often, to provide impedance isolation between bridge transducers and the differential amplifier stage, the signals v_1 and v_2 are amplified separately. This technique gives rise to the so-called **instrumentation amplifier** (IA), shown in Figure 5.9.9. Example 5.9.2 illustrates the calculation of the closed-loop gain for a typical instrumentation amplifier.

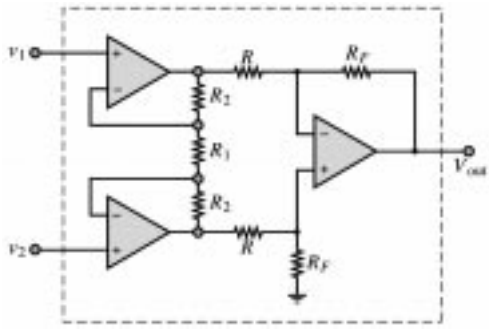


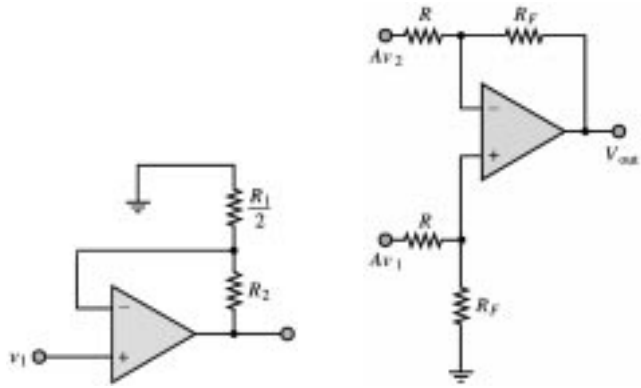
FIGURE 5.9.9 Instrumentation amplifier.

Example 5.9.2 Instrumentation Amplifier

In this example, we compute the closed-loop gain of the instrumentation amplifier of Figure 5.9.9.

Solution. To carry out the desired analysis as easily as possible, it is helpful to observe that resistor R_1 is shared by the two input amplifiers. This corresponds to having each amplifier connected to a resistor equal to $R_1/2$, as shown in Figure 5.9.10(a). Because of the symmetry of the circuit, one can view the shared resistor as being connected to ground in the center, so that the circuit takes the form of a noninverting amplifier, with closed-loop gain given by

$$A = 1 + \frac{2R_2}{R_1} \tag{5.9.10}$$



a b

FIGURE 5.9.10 (a) and (b).

Thus, each of the input voltages is amplified by this gain, and the overall gain of the instrumentation amplifier can then be computed by considering that the voltage difference $(Av_1 - Av_2)$ is then amplified by the differential amplifier stage, with gain R_F/R, as shown in Figure 5.9.10(b):

$$v_{out} = \frac{R_F}{R}\left(Av_1 - v_2\right) = \frac{R_F}{R}\left(1 + \frac{2R_2}{R_1}\right)\left(v_1 - v_2\right) \tag{5.9.11}$$

Active Filters

The class of filters one can obtain by means of op-amp designs is called **active filters**, because op-amps can provide amplification (gain) in addition to the filtering effects already studied in Section 5.6 for passive circuits (i.e., circuits comprising exclusively resistors, capacitors, and inductors).

The easiest way to see how the frequency response of an op-amp can be shaped (almost) arbitrarily is to replace the resistors R_F and R_S in Figures 5.9.2 and 5.9.4 with impedances Z_F and Z_S, as shown in Figure 5.9.11. It is a straightforward matter to show that in the case of the inverting amplifier, the expression for the closed loop gain is given by

$$\frac{\mathbf{V}_{\text{out}}}{\mathbf{V}_S}(j\omega) = -\frac{Z_F}{Z_S} \tag{5.9.12}$$

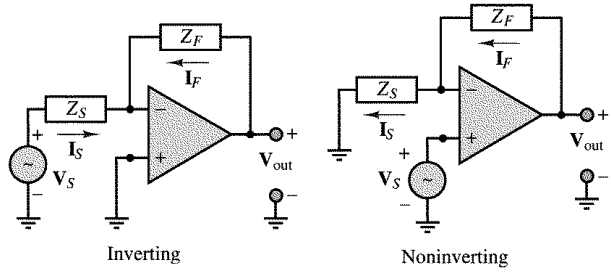


FIGURE 5.9.11 Op-amp circuits employing complex impedances.

whereas for the noninverting case, the gain is

$$\frac{\mathbf{V}_{\text{out}}}{\mathbf{V}_S}(j\omega) = 1 + \frac{Z_F}{Z_S} \tag{5.9.13}$$

where Z_F and Z_S can be arbitrarily complex impedance functions and where $\mathbf{V}_S$, $\mathbf{V}_{\text{out}}$, $\mathbf{I}_F$, and $\mathbf{I}_S$ are all phasors. Thus, it is possible to shape the frequency response of an ideal op-amp filter simply by selecting suitable ratios of feedback impedance to source impedance. The simplest op-amp low-pass filter is shown in Figure 5.9.12. The closed-loop gain $A_{LP}(j\omega)$ is then computed to be

$$A_{LP}(j\omega) = -\frac{Z_F}{Z_S} = -\frac{R_F/R_S}{1 + j\omega C_F R_F}$$

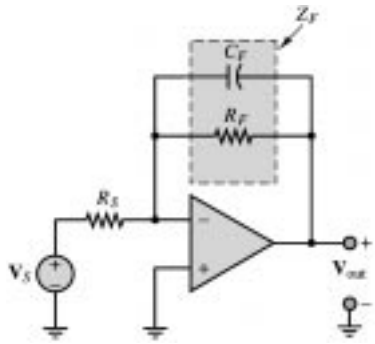


FIGURE 5.9.12 Active low-pass filter.

It should be apparent that the response of this op-amp filter is just an amplified version of that of the passive filter. Figure 5.9.13 depicts the amplitude response of the active low-pass filter (in the figure, $R_F/R_S = 10$ and $1/R_FC_F = 1$) in two different graphs; the first two plots the amplitude ratio $\mathbf{V}_{out}/\mathbf{V}_S$ vs. radian frequency, ω, on a logarithmic scale, while the second plots the amplitude ratio $20 \log_{10} (\mathbf{V}_{out}/\mathbf{V}_S)$ (in units of dB), also vs. ω on a logarithmic scale.

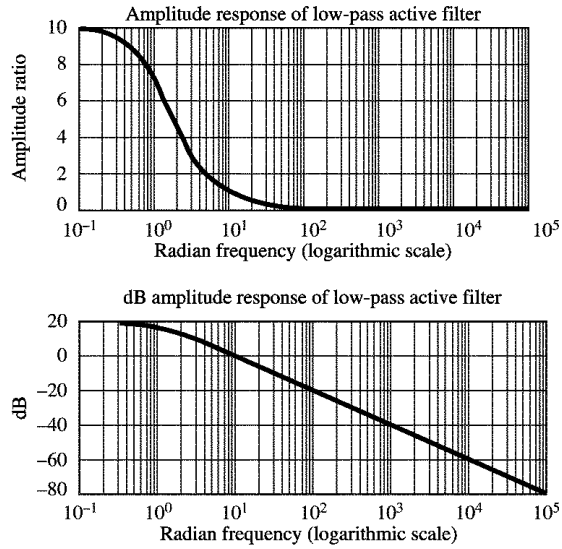


FIGURE 5.9.13 Normalized response of active low-pass filter.

A high-pass active filter can easily be obtained by using the circuit shown in Figure 5.9.14. The following gain function for the op-amp circuit can be derived:

$$A_{HP}(j\omega) = -\frac{Z_F}{Z_S} = -\frac{R_F}{R_S + 1/j\omega C_S} = -\frac{j\omega C_S R_F}{1 + j\omega R_S C_S} \tag{5.9.14}$$

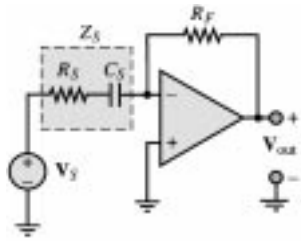


FIGURE 5.9.14 Active high-pass filter.

The high-pass response is depicted in Figure 5.9.15 in both linear and dB plots (in the figure, $R_F/R_S = 10$, $1/R_SC = 1$).

The band-pass filter of Figure 5.9.16 can be shown to have the following frequency response:

$$A_{BP}(j\omega) = -\frac{Z_F}{Z_S} = -\frac{j\omega C_S R_F}{(1 + j\omega C_F R_F)(1 + j\omega C_S R_S)} \tag{5.9.15}$$

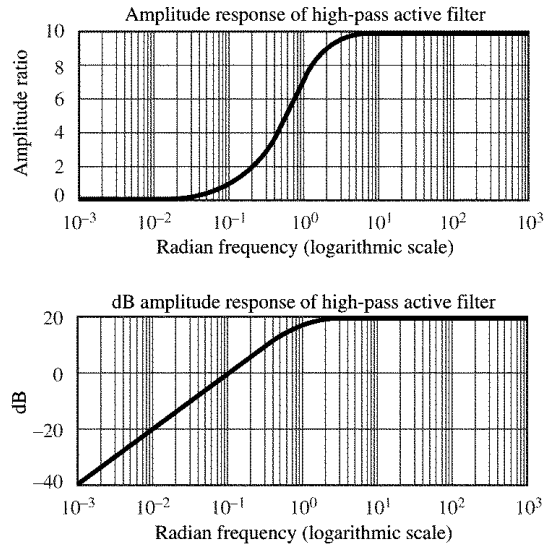


FIGURE 5.9.15 Normalized response of active high-pass filter.

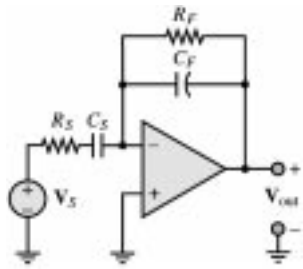


FIGURE 5.9.16 Active band-pass filter.

The form of the op-amp response we just obtained should not appear as a surprise. It is very similar (although not identical) to the product of the low-pass and high-pass responses:

$$A_{LP}(j\omega) = -\frac{R_F/R_S}{1 + j\omega C_F R_F} \tag{5.9.16}$$

$$A_{HP}(j\omega) = -\frac{j\omega C_S R_F}{1 + j\omega R_S C_S} \tag{5.9.17}$$

In particular, the denominator of $A_{BP}(j\omega)$ is exactly the product of the denominators of $A_{LP}(j\omega)$ and $A_{HP}(j\omega)$. It is particularly enlightening to rewrite $A_{LP}(j\omega)$ in a slightly different form, after making the observation that each *RC* product corresponds to some "critical" frequency:

$$\omega_1 = \frac{1}{R_F C_S} \quad \omega_{LP} = \frac{1}{R_F C_F} \quad \omega_{HP} = \frac{1}{R_S C_S} \tag{5.9.18}$$

It is easy to verify that for the case where

$$\omega_{HP} > \omega_{LP} \tag{5.9.19}$$

the response of the op-amp filter may be represented as shown in Figure 5.9.17 in both linear and dB plots (in the figure, $\omega_1 = 1$, $\omega_{HP} = 1000$, $\omega_{HP} = 10$). The dB plot is very revealing, for its shows that, in effect, the band-pass response is the graphical superposition of the low-pass and high-pass responses shown earlier. The two 3-dB (or cutoff) frequencies are the same as in $A_{LP}(j\omega)$, $1/R_FC_F$; and in $A_{HP}(j\omega)$, $1/R_SC_S$. The third frequency, $\omega_1 = 1/R_FC_S$, represents the point where the response of the filter crosses the 0-dB axis (rising slope). Since 0 dB corresponds to a gain of 1, this frequency is called the **unity gain frequency**.

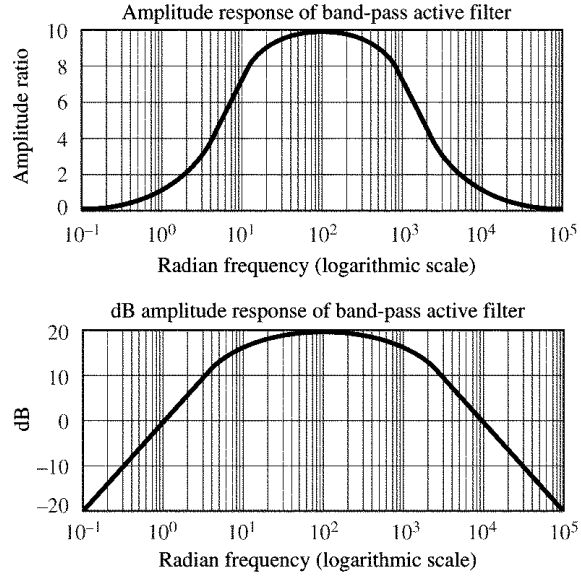


FIGURE 5.9.17 Normalized response of active band-pass filter.

The ideas developed thus far can be employed to construct more complex functions of frequency. In fact, most active filters one encounters in practical applications are based on circuits involving more than one or two energy-storage elements. By constructing suitable functions for Z_F and Z_S, it is possible to realize filters with greater frequency selectivity (i.e., sharpness of cutoff), as well as flatter band-pass or band-rejection functions (that is, filters that either allow or reject signals in a limited band of frequencies). A few simple applications are investigated in the homework problems. One remark that should be made in passing, though, pertains to the exclusive use of capacitors in the circuits analyzed thus far.

Integrator and Differentiator Circuits

The Ideal Integrator

Consider the circuit of Figure 5.9.18, where $v_s(t)$ is an arbitrary function of time (e.g., a pulse train, a triangular wave, or a square wave). The op-amp circuit shown provides an output that is proportional to the integral of $v_s(t)$.

$$v_{\text{out}} = -\frac{1}{R_SC_F}\int_{-\infty}^{t} v_S(t')\,dt' \tag{5.9.20}$$

This equation states that the output voltage is the integral of the input voltage.

There are numerous applications of the op-amp integrator, most notably the **analog computer**. The following example illustrates the operation of the op-amp integrator.

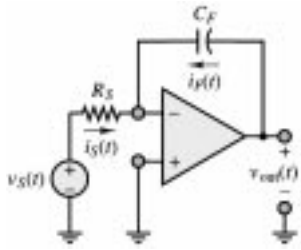


FIGURE 5.9.18 Op-amp integrator.

Example 5.9.3 Charge Amplifiers

One of the most common families of transducers for the measurement of force, pressure, and acceleration is that of **piezoelectric transducers**. These transducers contain a piezoelectric crystal, a crystal that generates an electric charge in response to deformation. Thus, if a force is applied to the crystal (leading to a displacement), a charge is generated within the crystal. If the external force generates a displacement x_i, then the transducer will generate a charge q according to the expression

$$q = K_p x_i$$

Figure 5.9.19 depicts the basic structure of the piezoelectric transducer and a simple circuit model. The model consists of a current source in parallel with a capacitor, where the current source represents the rate of change of the charge generated in response to an external force and the capacitance is a consequence of the structure of the transducer, which consists of a piezoelectric crystal (e.g., quartz or Rochelle salt) sandwiched between conducting electrodes (in effect, this is a parallel-plate capacitor).

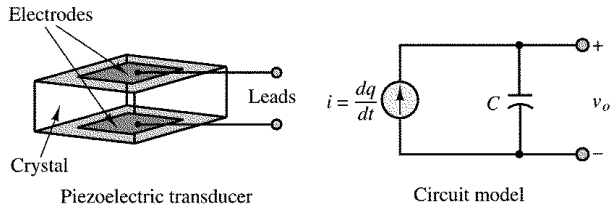


FIGURE 5.9.19 Piezolectric transducer.

Although it is possible, in principle, to employ a conventional voltage amplifier to amplify the transducer output voltage, v_o, given by

$$v_o = \frac{1}{C}\int i\,dt = \frac{1}{C}\int \frac{dq}{dt}\,dt = \frac{q}{C} = \frac{K_p x_i}{C}$$

it is often advantageous to use a **charge amplifier**. The charge amplifier is essentially an integrator circuit, as shown in Figure 5.9.20 characterized by an extremely high input impedance. The high impedance is essential; otherwise, the charge generated by the transducer would leak to ground through the input resistance of the amplifier.

Because of the high input impedance, the input current into the amplifier is negligible; further, because of the high open-loop gain of the amplifier, the inverting-terminal voltage is essentially at ground potential. Thus, *the voltage across the transducer is effectively zero*. As a consequence, to satisfy KCL, the feedback current, $i_F(t)$ must be equal and opposite to the transducer current, i:

$$i_F(t) = -i$$

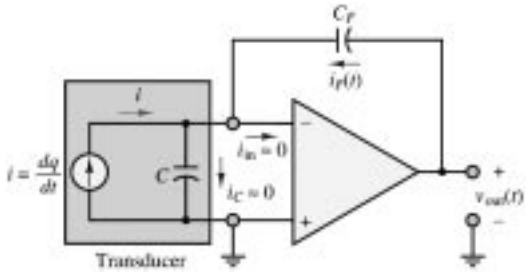


FIGURE 5.9.20 Charge amplifier.

and since

$$v_{\text{out}}(t) = \frac{1}{C_F} i_F(t)\, dt$$

it follows that the output voltage is proportional to the charge generated by the transducer, and therefore to the displacement:

$$v_{\text{out}}(t) = \frac{1}{C_F}\int -i\, dt = \frac{1}{C_F}\int -\frac{dq}{dt}\, dt = -\frac{q}{C_F} = -\frac{K_p x_i}{C_F}$$

Since the displacement is caused by an external force or pressure, this sensing principle is widely adopted in the measurement of force and pressure.

The Ideal Differentiator

Using an argument similar to that employed for the integrator, we can derive a result for the ideal differentiator circuit of Figure 5.9.21. The relationship between input and output is obtained by observing that

$$i_S(t) = C_S \frac{dv_S(t)}{dt} \tag{5.9.21}$$

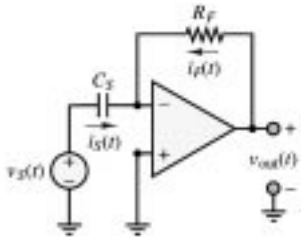


FIGURE 5.9.21 Op-amp differentiators.

and

$$i_F(t) = \frac{v_{\text{out}}(t)}{R_F} \tag{5.9.22}$$

so that the output of the differentiator circuit is proportional to the derivative of the input:

$$v_{\text{out}}(t) = -R_F C_S \frac{dv_S(t)}{dt} \tag{5.9.23}$$

Although mathematically attractive, the differentiation property of this op-amp circuit is seldom used in practice, because differentiation tends to simplify any noise that may be present in a signal.

Physical Limitations of Op-Amps

Thus far, the operational amplifier has been treated an ideal device, characterized by infinite input resistance, zero output resistance, and infinite open-loop voltage gain. Although this model is adequate to represent the behavior of the op-amp in a large number of applications, it is important to realize that practical operational amplifiers are not ideal devices, but exhibit a number of limitations that should be considered in the design of instrumentation. In particular, in dealing with relatively large voltages and currents, and in the presence of high-frequency signals, it is important to be aware of the nonideal properties of the op-amp. In the present section, we examine the principal limitations of the operational amplifier.

Voltage Supply Limits

The effect of limiting supply voltages is that amplifiers are capable of amplifying signals *only within the range of their supply voltages*;

$$V_S^- < v_{\text{out}} < V_S^+ \tag{5.9.24}$$

For most op-amps, the limit is actually approximately 1.5 V less than the supply voltages.

Frequency Response Limits

Another property of all amplifiers that may pose severe limitations to the op-amp is their finite bandwidth. We have so far assumed, in our ideal op-amp model, that the open-loop gain is a very large constant. In reality, $A_{V(OL)}$ is a function of frequency and is characterized by a low-pass response. For a typical op-amp,

$$A_{V(OL)}(j\omega) = \frac{A_0}{1 + j\omega/\omega_0} \tag{5.9.25}$$

The finite bandwidth of the practical op-amp results in a fixed **gain-bandwidth product** for any given amplifier. The effect of a constant gain-bandwidth product is that as the closed-loop gain of the amplifier is increased, its 3-dB bandwidth is proportionally reduced, until, in the limit, if the amplifier were used in the open-loop mode, its gain would be equal to A_0 and its 3-dB bandwidth would be equal to ω_0. Thus, *the product of gain and bandwidth in any given op-amp is constant*. That is,

$$A_0 \times \omega_0 = A_1 \times \omega_1 = A_2 \times \omega_2 = K \tag{5.9.26}$$

as is shown in Figure 5.9.22.

Input Offset Voltage

Another limitation of practical op-amps results because even in the absence of any external inputs, it is possible that an **offset voltage** will be present at the input of an op-amp. This voltage is usually denoted by $\pm V_{os}$ and it is caused by mismatches in the internal circuitry of the op-amp. The offset voltage appears as a differential input voltage between the inverting and noninverting input terminals. The presence of an additional input voltage will cause a DC bias error in the amplifier output, which can be modeled as shown in Figure 5.9.23.

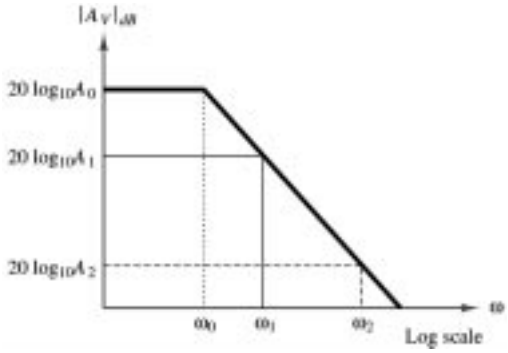


FIGURE 5.9.22

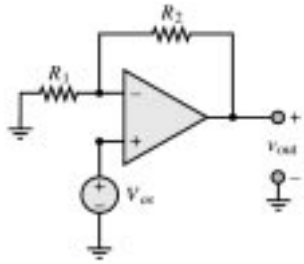


FIGURE 5.9.23 Op-amp input offset voltage.

Input Bias Currents

Another nonideal characteristic of op-amps results from the presence of small input bias currents at the inverting and noninverting terminals. Once again, these are due to the internal construction of the input stage of an operational amplifier. Figure 5.9.24 illustrates the presence of nonzero input bias currents (I_B) going into an op-amp.

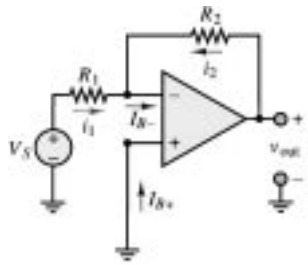


FIGURE 5.9.24

Typical values of I_B depend on the semiconductor technology employed in the construction of the op-amp. Op-amps with bipolar transistor input stages may see input bias currents as large as 1 μA, while for FET input devices, the input bias currents are less than 1 nA. Since these currents depend on the internal design of the op-amp, they are not necessarily equal. One often designates the **input offset current** I_{os} as

$$I_{os} = I_{B+} - I_{B-} \tag{5.9.27}$$

The latter parameter is sometimes more convenient from the standpoint of analysis.

Output Offset Adjustment

Both the offset voltage and the input offset current contribute to an output offset voltage $V_{\text{out},os}$. Some op-amps provide a means for minimizing $V_{\text{out},os}$. For example, the μA741 op-amp provides a connection for this procedure. Figure 5.9.25 shows a typical pin configuration for an op-amp in an eight-pin dual-

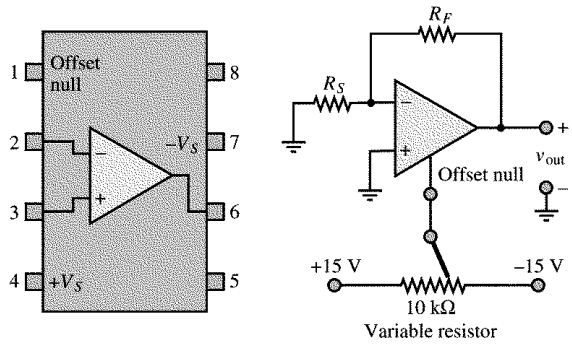


FIGURE 5.9.25 Output off-set voltage adjustment.

in-line package (DIP) and the circuit used for nulling the output offset voltage. The variable resistor is adjusted until v_{out} reaches a minimum (ideally, 0 V). Nulling the output voltage in this manner removes the effect of both input offset voltage and current on the output.

Slew Rate Limit

Another important restriction in the performance of a practical op-amp is associated with rapid changes in voltage. The op-amp can produce only a finite rate of change at its output. This limit rate is called the **slew rate**. Consider an ideal step input, where at $t = 0$ the input voltage is switched from zero to V volts. Then we would expect the output to switch from 0 to $A \cdot V$ volts, where A is the amplifier gain. However, $v_{out}(t)$ can change at only a finite rate; thus,

$$\left|\frac{dv_{out}(t)}{dt}\right|_{max} = S_0 = \text{Slew rate} \tag{5.9.28}$$

Figure 5.9.26 shows the response of an op-amp to an ideal step change in input voltage, Here S_0, the slope of $v_{out}(t)$, represents the slew rate.

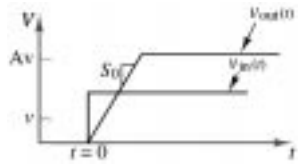


FIGURE 5.9.26 Slew rate limit in op-amps.

Short-Circuit Output Current

Recall the model for the op-amp, which represented the internal circuits of the op-amp in terms of an equivalent input resistance, R_{in}, and a controlled voltage source, $A_V\, v_{in}$. In practice, the internal source is not ideal, because it cannot provide an infinite amount of current (either to the load or to the feedback connection, or both). The immediate consequence of this nonideal op-amp characteristic is that the maximum output current of the amplifier is limited by the so-called short-circuit output current, I_{SC}:

$$\left|I_{out}\right| < I_{SC} \tag{5.9.29}$$

To further explain this point, consider that the op-amp needs to provide current to the feedback path (in order to "zero" the voltage differential at the input) and to whatever load resistance, R_L, may be connected to the output. Figure 5.9.27 illustrates this idea for the case of an inverting amplifier, where I_{SC} is the load current that would be provided to a short-circuit load ($R_L = 0$).

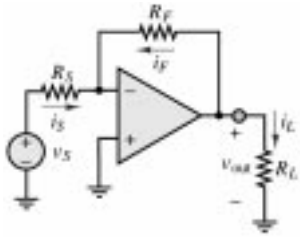


FIGURE 5.9.27

Common-Mode Rejection Ratio (CMRR)

Example 5.9.2 introduced the notion of differential-mode and common-mode signals. If we define A_{dm} as the **differential-mode gain** and A_{cm} as the **common-mode gain** of the op-amp, the output of an op-amp can then be expressed as follows:

$$v_{\text{out}} = A_{dm}\left(v_2 - v_1\right) + A_{cm}\left(\frac{v_2 - v_1}{2}\right) \tag{5.9.30}$$

Under ideal conditions, A_{cm} should be exactly zero, since the differential amplifier should completely reject common-mode signals. The departure from this ideal condition is a figure of merit for a differential amplifier and is measured by defining a quantity called the **common-mode rejection ratio** (CMRR). The CMRR is defined as the ratio of the differential-mode gain to the common-mode gain and should ideally be infinite:

$$\text{CMRR} = \frac{A_{dm}}{A_{cm}}$$

The CMRR is often expressed in units of decibels (dB).

5.10 Digital Circuits

The objective of this section is to discuss the essential features of digital logic circuits, which are at the heart of digital computers.

Analog and Digital Signals

One of the fundamental distinctions in the study of electronic circuits (and in the analysis of any signals derived from physical measurements) is that between analog and digital signals. As discussed in the preceding chapter, an **analog signal** is an electrical signal whose value varies in analogy with a physical quantity (e.g., temperature, force, or acceleration). For example, a voltage proportional to a measured variable pressure or to a vibration naturally varies in an analog fashion. Figure 5.10.1 depicts an arbitrary analog function of time, $f(t)$. We note immediately that for each value of time, t, $f(t)$ can take one value among any of the values in a given range. For example, in the case of the output voltage of an op-amp, we expect the signal to take any value between $+V_{sat}$ and $-V_{sat}$, where V_{sat} is the supply-imposed saturation voltage.

FIGURE 5.10.1 Analog signal.

A **digital signal**, on the other hand, can take only a *finite number of values*. This is an extremely important distinction, as will be shown shortly. An example of a digital signal is a signal that allows display of a temperature measurement on a digital readout. Let us hypothesize that the digital readout is three digits long and can display numbers from 0 to 100, and let us assume that the temperature sensor is correctly calibrated to measure temperatures from 0 to 100°F. Further, the output of the sensor ranges from 0 to 5 V, where 0 V corresponds to 0°F and 5 V to 100°F. Therefore, the calibration constant of the sensor is $k_T = (100° - 0°)/(5 - 0) = 20°/V$. Clearly, the output of the sensor is an analog signal; however, the display can show only a finite number of readouts (101, to be precise). Because the display itself can only take a value out of a discrete set of states — the integers from 0 to 100 — we call it a digital display, indicating that the variable displayed is expressed in digital form.

Now, each temperature on the display corresponds to a *range of voltages:* each digit on the display represents one hundredth of the 5-V range of the sensor, or 0.05 V = 50 m V. Thus, the display will read 0 if the sensor voltage is between 0 and 49 mV, 1 if it is between 50 and 99 mV, and so on. Figure 5.10.2 depicts the staircase function relationship between the analog voltage and the digital readout. This **quantization** of the sensor output voltage is, in effect, an approximation. If one wished to know the temperature with greater precision, a greater number of display digits could be employed.

The most common digital signals are binary signals. A **binary signal** is a signal that can take only one of two discrete values and is therefore characterized by transitions between two states. Figure 5.10.3 displays a typical binary signal. In binary arithmetic (which we discuss in the next section), the two discrete values f_1 and f_0 are represented by the numbers 1 and 0. In binary voltage waveforms, these values are represented by two voltage levels. For example, in the TTL convention, these values are (nominally) 5 and 0 V, respectively; in CMOS circuits, these values can vary substantially. Other conventions are also used, including reversing the assignment — for example, by letting a 0-V level represent a logic 1 and a 5-V level represent a logic 0. Note that in a binary waveform, knowledge of the transition between one stage and another (e.g., from f_0 to f_1 at $t = t_2$) is equivalent to knowledge of

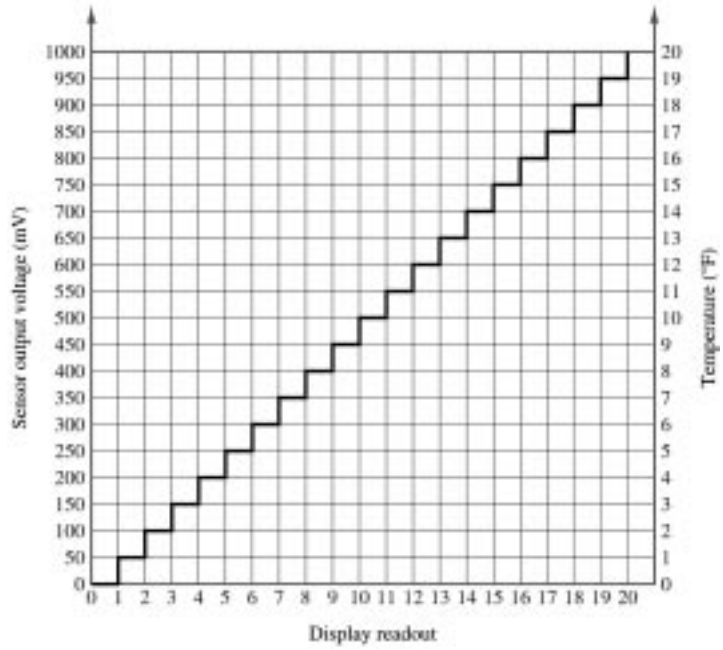


FIGURE 5.10.2 Digital representation of an analog signal.

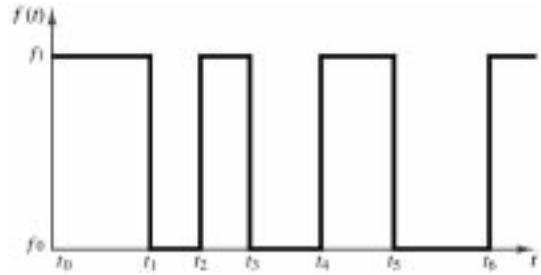


FIGURE 5.10.3 A binary signal.

the state. Thus, digital logic circuits can operate by detecting transitions between voltage levels. The transitions are often called **edges** and can be positive (f_0 to f_1) or negative (f_1 to f_0). Virtually all of the signals handled by a computer are binary. From here on, whenever we speak of digital signals, you may assume that the text is referring to signals of the binary type, unless otherwise indicated.

The Binary Number System

The binary number system is a natural choice for representing the behavior of circuits that operate in one of two states (on or off, 1 or 0, or the like). The diode and transistor gates and switches studied in Section 5.7 fall in this category. Table 5.10.1 shows the correspondence between decimal and binary number systems for decimal numbers up to 16.

Table 5.10.1 shows that it takes four binary digits, also called **bits**, to represent the decimal numbers up to 15. Usually, the rightmost bit is called the **least significant bit**, or LSB, and the leftmost bit is called the **most significant bit**, or MSB. Since binary numbers clearly require a larger number of digits than decimal numbers, the digits are usually grouped in sets of four, eight, or sixteen. Four bits are usually termed a **nibble**, eight bits are called a **byte**, and sixteen bits (or two bytes) form a **word**.

TABLE 5.10.1 Conversion from Decimal to Binary

Decimal Number, n_{10}	Binary Number, n_2
0	0
1	1
2	10
3	11
4	100
5	101
6	110
7	111
8	1000
9	1001
10	1010
11	1011
12	1100
13	1101
14	1110
15	1111
16	10000

Addition and Subtraction

The operations of addition and subtraction are based on the simple rules shown in Table 5.10.2. Figure 5.10.4 provides three examples.

TABLE 5.10.2 Rules for Addition

0 + 0 = 0
0 + 1 = 1
1 + 0 = 1
1 + 1 = 0 (with a carry of 1)

Decimal	Binary	Decimal	Binary	Decimal	Binary
5	101	15	1111	3.25	11.01
+6	+110	+20	+10100	+5.75	+101.11
11	1011	35	100011	9.00	1001.00

FIGURE 5.10.4 Examples of binary addition.

The procedure for subtracting binary numbers is based on the rules of Table 5.10.3. A few examples of binary subtraction are given in Figure 5.10.5, with their decimal counterparts.

TABLE 5.10.3 Rules for Subtraction

0 – 0 = 0
1 – 0 = 1
1 – 1 = 0
0 – 1 = 1 (with a borrow of 1)

Multiplication and Division

Whereas in the decimal system the multiplication table consists of $10^2 = 100$ entries, in the binary system we only have $2^2 = 4$ entries. Table 5.10.4 represents the complete multiplication table for the binary number system.

Decimal	Binary	Decimal	Binary	Decimal	Binary
9	1001	16	10000	6.25	110.01
−5	−101	−3	−11	−4.50	−100.10
4	0100	13	01101	1.75	001.11

FIGURE 5.10.5 Examples of binary subtraction.

TABLE 5.10.4 Rules for Multiplication

$0 \times 0 = 0$
$0 \times 1 = 0$
$1 \times 0 = 0$
$1 \times 1 = 1$

Division in the binary system is also based on rules analogous to those of the decimal system, with the two basic laws given in Table 5.10.5. Once again, we need be concerned with only two cases, and just as in the decimal system, division by zero is not contemplated.

TABLE 5.10.5 Rules for Division

$0 \div 1 = 0$
$1 \div 1 = 1$

Conversion from Decimal to Binary

The conversion of a decimal number to its binary equivalent is performed by successive division of the decimal number by 2, checking for the remainder each time. Figure 5.10.6 illustrates this idea with an example.

Remainder
$49 \div 2 = 24 + 1$
$24 \div 2 = 12 + 0$
$12 \div 2 = 6 + 0$
$6 \div 2 = 3 + 0$
$3 \div 2 = 1 + 1$
$1 \div 2 = 0 + 1$
$49_{10} = 110001_2$

FIGURE 5.10.6 Example of conversion from decimal to binary.

The same technique can be used for converting decimal fractional numbers to their binary form, provided that the whole number is separated from the fractional part and each is converted to binary form (separately), with the results added at the end. Figure 5.10.7 outlines this procedure by converting the number 37.53 to binary form.

Complements and Negative Numbers

To simplify the operation of subtraction in digital computers, **complements** are used almost exclusively. In practice, this corresponds to replacing the operation $X - Y$ with operation $X + (-Y)$. This procedure results in considerable simplification, since the computer hardware need include only adding circuitry. Two types of complements are used with binary numbers: the **one's complement** and the **two's complement**.

The one's complement of an n-bit binary number is obtained by subtracting the number itself from $(2^n - 1)$. Two examples are as follows:

Remainder
$37 \div 2 = 18 + 1$
$18 \div 2 = 9 + 0$
$9 \div 2 = 4 + 1$
$4 \div 2 = 2 + 0$
$2 \div 2 = 1 + 0$
$1 \div 2 = 0 + 1$
$37_{10} = 100101_2$

$2 \times 0.53 = 1.06 \rightarrow 1$
$2 \times 0.06 = 0.12 \rightarrow 0$
$2 \times 0.12 = 0.24 \rightarrow 0$
$2 \times 0.24 = 0.48 \rightarrow 0$
$2 \times 0.48 = 0.96 \rightarrow 0$
$2 \times 0.96 = 1.92 \rightarrow 1$
$2 \times 0.92 = 1.84 \rightarrow 1$
$2 \times 0.84 = 1.68 \rightarrow 1$
$2 \times 0.68 = 1.36 \rightarrow 1$
$2 \times 0.36 = 0.72 \rightarrow 0$
$2 \times 0.72 = 1.44 \rightarrow 1$
$0.53_{10} = 0.10000111101$

FIGURE 5.10.7 Conversion from decimal to binary.

$$a = 0101$$

$$\text{One's complement of } a = \left(2^4 - 1\right) - a$$

$$= (1111) - (0101)$$

$$= 1010$$

$$b = 101101$$

$$\text{One's complement of } b = \left(2^6 - 1\right) - b$$

$$= (111111) - (101101)$$

$$= 010010$$

The two's complement of n-bit binary number is obtained by subtracting the number itself from 2^n. Two's complements of the same numbers a and b used in the preceding illustration are computed as follows:

$$a = 0101$$

$$\text{Two's complement of } a = 2^4 - a$$

$$= (10000) - (0101)$$

$$= 1011$$

$$b = 101101$$

$$\text{Two's complement of } b = 2^6 - b$$

$$= (1000000) - (101101)$$

$$= 010011$$

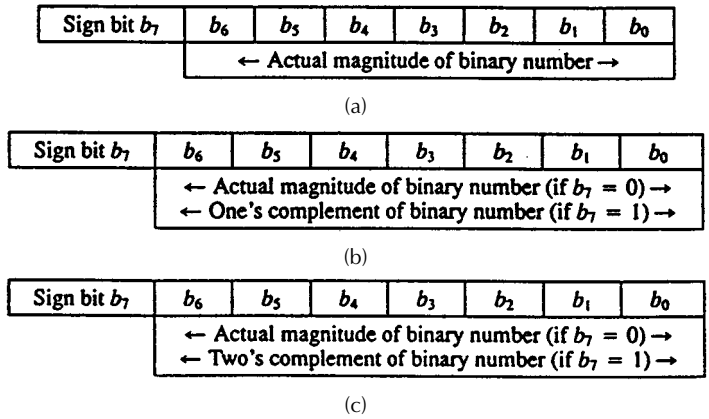


FIGURE 5.10.8 (a) Eight-bit sign-magnitude binary number; (b) eight-bit one's complement binary number; (c) eight-bit two's complement binary number.

Different conventions exist in the binary system to represent whether a number is negative or positive. These are summarized in Figure 5.10.8.

The Hexadecimal System

It should be apparent by now that representing numbers in base 2 and base 10 systems is purely a matter of convenience, given a specific application. Another base frequently used is the **hexadecimal system**, a direct derivation of the binary number system. In the hexadecimal (or hex) code, the bits in a binary number are subdivided into groups of four. Since there are 16 possible combinations for a four-bit number, the natural digits in the decimal system (0 through 9) are insufficient to represent a hex digit. To solve this problem, the first six letters of the alphabet are used, as shown in Table 5.10.6. Thus, in hex code, an eight-bit word corresponds to just two digits; for example:

$$1010\ 0111_2 = A7_{16}$$

$$0010\ 1001_2 = 29_{16}$$

TABLE 5.10.6 Hexadecimal Code

0	0000
1	0001
2	0010
3	0011
4	0100
5	0101
6	0110
7	0111
8	1000
9	1001
A	1010
B	1011
C	1100
D	1101
E	1110
F	1111

Binary Codes

In this subsection, we describe two common binary codes that are often used for practical reasons. The first is a method of representing decimal numbers in digital logic circuits that is referred to as **binary-coded decimal**, or BCD, **representation**. In effect, the simplest BCD representation is just a sequence of four-bit binary numbers that stops after the first ten entries, as shown in Table 5.10.7. There are also other BCD codes, all reflecting the same principle: that each decimal digit is represented by a fixed-length binary word. One should realize that although this method is attractive because of its direct correspondence with the decimal system, it is not efficient. Consider, for example, the decimal number 68. Its binary representation by direct conversion is the seven-bit number 1000100. On the other hand, the corresponding BCD representation would require eight bits:

$$68_{10} = 01101000_{\text{BCD}}$$

TABLE 5.10.7 BCD Code

0	0000
1	0001
2	0010
3	0011
4	0100
5	0101
6	0110
7	0111
8	1000
9	1001

Another code that finds many applications is the **Gray code**. This is simply a reshuffling of the binary code with the property that any two consecutive numbers differ only by one bit. Table 5.10.8 illustrates the three-bit Gray code. The Gray code can be very useful in practical applications, because in counting up or down according to this code, the binary representation of a number changes only one bit at a time.

TABLE 5.10.8 Three-Bit Gray Code

Binary	Gray
000	000
001	001
010	011
011	010
100	110
101	111
110	101
111	100

Boolean Algebra

The mathematics associated with the binary number system (and with the more general field of logic) is called *Boolean algebra*. The variables in a Boolean, or logic, expression can take only one of two values, usually represented by the numbers 0 and 1. These variables are sometimes referred to as true (1) and false (0). This convention is normally referred to as **positive logic**. There is also a **negative logic** convention in which the roles of logic 1 and logic 0 are reversed. In this book we shall employ only positive logic.

Analysis of **logic functions**, that is, functions of logical (Boolean) variables, can be carried out in terms of truth tables. A truth table is a listing of all the possible values each of the Boolean variables

can take, and of the corresponding value of the desired function. In the following paragraphs we shall define the basic logic functions upon which Boolean algebra is founded, and we shall describe each in terms of a set of rules and a truth table; in addition, we shall also introduce **logic gates**. Logic gates are physical devices that can be used to implement logic functions. Elementary logic gates were introduced in Section 5.7.

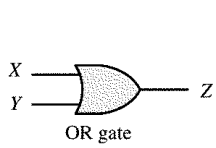
AND and OR Gates

The basis of **Boolean algebra** lies in the operations of **logical addition**, or the **OR** operation; and **logical multiplication**, or the **AND** operation. Both of these find a correspondence in simple logic gates, as we shall presently illustrate. Logical addition, although represented by the symbol +, differs from conventional algebraic addition, as shown in the last rule listed in Table 5.10.9. Note that this rule also differs from the last rule of binary addition studied in the previous section. Logical addition can be represented by the logic gate called an **OR gate**, whose symbol and whose inputs and outputs are shown in Figure 5.10.9. The OR gate represents the following logical statement:

TABLE 5.10.9 Rules for Logical Addition (OR)

$0 + 0 = 0$
$0 + 1 = 1$
$1 + 0 = 1$
$1 + 1 = 1$

If either X or Y is true (1), then Z is true (1) (5.10.1)



X	Y	Z
0	0	0
0	1	1
1	0	1
1	1	1

Truth table

FIGURE 5.10.9 Logical addition and the OR gate.

This rule is embodied in the electronic gates discussed in Chapter 9, in which a logic 1 corresponds, say, to a 5-V signal and a logic 0 to a 0-V signal.

Logical multiplication is denoted by the center dot (·) and is defined by the rules of Table 5.10.10. Figure 5.10.10 depicts the **AND gate**, which corresponds to this operation. The AND gate corresponds to the following logical statement:

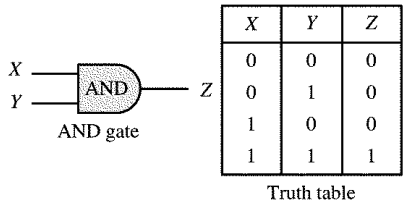
TABLE 5.10.10 Rules for Logical Multiplication (AND)

$0 \cdot 0 = 0$
$0 \cdot 1 = 0$
$1 \cdot 0 = 0$
$1 \cdot 1 = 1$

If both X and Y are true (1), then Z is true (1) (5.10.2)

One can easily envision logic gates (AND and OR) with an arbitrary number of inputs; three- and four-input gates are not uncommon.

The rules that define a logic function are often represented in tabular form by means of a **truth table**. Truth tables for the AND and OR gates are shown in Figures 5.10.9 and 5.10.10. A truth table is nothing more than a tabular summary of all of the possible outputs of a logic gate, given all the possible input



X	Y	Z
0	0	0
0	1	0
1	0	0
1	1	1

Truth table

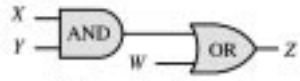
FIGURE 5.10.10 Logical multiplication and the AND gate.

values. If the number of inputs is 3, the number of possible combinations grows from 4 to 8, but the basic idea is unchanged. Truth tables are very useful in defining logic functions. A typical logic design problem might specify requirements such as "the output Z shall be logic 1 only when the condition (X = 1 AND Y = 1) OR (W = 1) occurs, and shall be logic 0 otherwise." The truth table for this particular logic function is shown in Figure 5.10.11 as an illustration. The design consists, then, of determining the combination of logic gates that exactly implements the required logic function. Truth tables can greatly simplify this procedure.

Logic gate realization of the statement "the output Z shall be logic 1 only when the condition (X = 1 AND Y = 1) OR (W = 1) occurs, and shall be logic 0 otherwise."

X	Y	W	Z
0	0	0	0
0	0	1	1
0	1	0	0
0	1	1	1
1	0	0	0
1	0	1	1
1	1	0	1
1	1	1	1

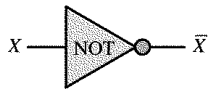
Truth table



Solution using logic gates

FIGURE 5.10.11 Example of logic function implementation with logic gates.

The AND and OR gates form the basis of all logic design in conjunction with the **NOT gate**. The NOT gate is essentially an inverter, and it provides the complement of the logic variable connected to its input. The complement of a logic variable X is denoted by $\overline{X}$. The NOT gate has only one input, as shown in Figure 5.10.12



NOT gate

X	$\overline{X}$
1	0
0	1

Truth table for NOT gate

FIGURE 5.10.12 Complements and the NOT gate.

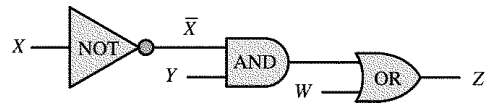
To illustrate the use of the NOT gate, or inverter, we return to the design example of Figure 5.10.11, where we required that the output of a logic circuit be Z = 1 only if X = 0 AND Y = 1 OR if W = 1. We recognize that except for the requirement X = 0, this problem would be identical if we stated it as follows: "The output Z shall be logic 1 only when the condition ($\overline{X}$ = 1 AND Y = 1) OR (W = 1) occurs, and shall be logic 0 otherwise." If we use an inverter to convert X to $\overline{X}$, we see that the required

condition becomes $\overline{X}$ = 1 AND Y = 1) OR (W = 1). The formal solution to this elementary design exercise is illustrated in Figure 5.10.13.

X	$\overline{X}$	Y	W	Z
0	1	0	0	0
0	1	0	1	1
0	1	1	0	1
0	1	1	1	1
1	0	0	0	0
1	0	0	1	1
1	0	1	0	0
1	0	1	1	1

(Required logic function)

Truth table



Solution using logic gates

FIGURE 5.10.13 Solution of a logic problem using logic gates.

In the course of the discussion of logic gates, extensive use will be made of truth tables to evaluate logic expressions. A set of basic rules will facilitate this task. Table 5.10.11 lists some of the rules of Boolean algebra:

TABLE 5.10.11 Rules of Boolean Algebra

1.	$0 + X = X$	
2.	$1 + X = 1$	
3.	$X + X = X$	
4.	$X + \overline{X} = 1$	
5.	$0 \cdot X = 0$	
6.	$1 \cdot X = X$	
7.	$X \cdot X = X$	
8.	$X \cdot \overline{X} = 0$	
9.	$\overline{\overline{X}} = X$	
10.	$X + Y = Y + X$	Commutative law
11.	$X \cdot Y = Y \cdot X$	
12.	$X + (X + Z) = (X + Y) + Z$	Associative law
13.	$X \cdot (Y \cdot Z) = (X \cdot Y) \cdot Z$	
14.	$X \cdot (Y + Z) = X \cdot Y + X \cdot Z$	Distributive law
15.	$X + X \cdot Z = X$	Absorption law
16.	$X \cdot (X + Y) = X$	
17.	$(X + Y) \cdot (X + Z) = X + Y \cdot Z$	
18.	$X + \overline{X} \cdot Y = X + Y$	
19.	$X \cdot Y + Y \cdot Z + \overline{X} \cdot Z = X \cdot Y + \overline{X} \cdot Z$	

To complete the introductory material on Boolean algebra, a few paragraphs need to be devoted to two very important theorems, called **De Morgan's theorems**. These are stated here in the form of logic functions:

$$\left(\overline{X + Y}\right) = \overline{X} \cdot \overline{Y} \tag{5.10.3}$$

$$\left(\overline{X \cdot Y}\right) = \overline{X} + \overline{Y} \tag{5.10.4}$$

These two laws state a very important property of logic functions: any logic function can be implemented using only OR and NOT gates, or using only AND and NOT gates.

De Morgan's laws can easily be visualized in term of logic gates, as shown in Figure 5.10.14. The associated truth tables are proof of these theorems.

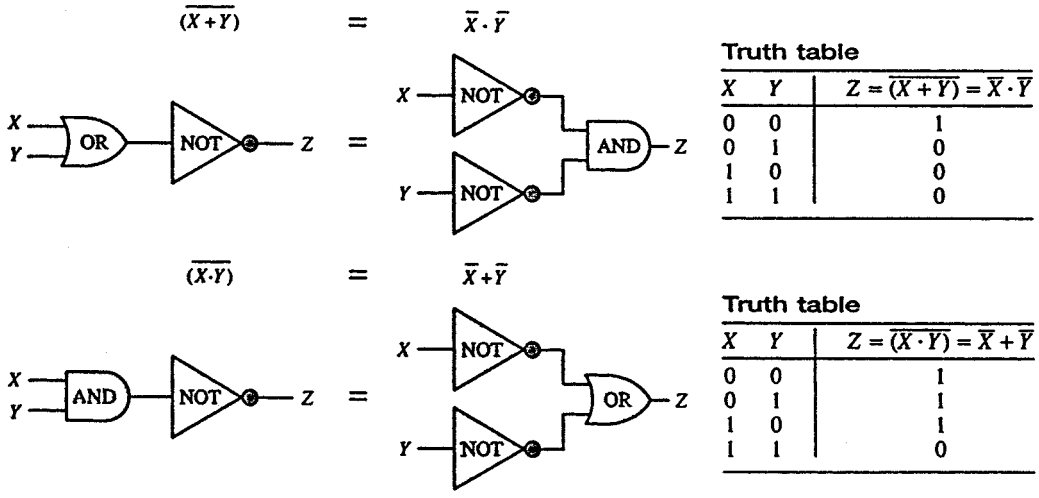


FIGURE 5.10.14 De Morgan's laws.

The importance of De Morgan's laws is in the statement of the **duality** that exists between AND and OR operations: any function can be realized by just one of the two basic operations, plus the complement operation. This gives rise to two families of logic functions: **sums of products** and **products of sums**, as shown in Figure 5.10.15. Any logical expression can be reduced to either one of these two forms. Although the two forms are equivalent, it may well be true that one of the two has a simpler implementation (fewer gates). Example 5.10.1 illustrates this point.

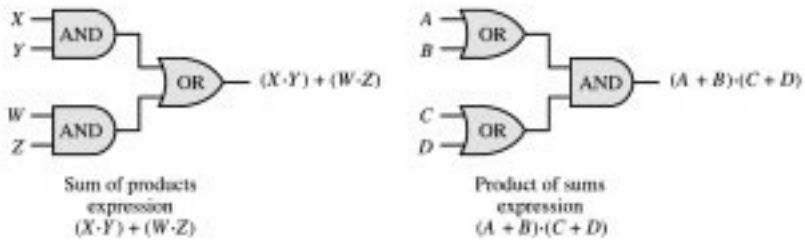


FIGURE 5.10.15 Sun-of-products and product-of-sums logic functions.

Example 5.10.1

Use De Morgan's theorem to realize the function $y = A + (B \cdot C)$ as a product-of-sums expression, and implement it using AND, OR, and NOT gates.

Solution. Knowing that $\overline{\overline{y}} = y$, we can apply the first of De Morgan's laws to the complement of the function y to obtain the expression

$$\overline{y} = \overline{A + (B \cdot C)} = \overline{A} \cdot \left(\overline{B \cdot C}\right) = \overline{A} \cdot \left(\overline{B} + \overline{C}\right)$$

Thus,

$$\overline{y} = y = \overline{A} \cdot \left(\overline{B} + \overline{C}\right)$$

Using logic gates, we can then implement the function as shown in Figure 5.10.16

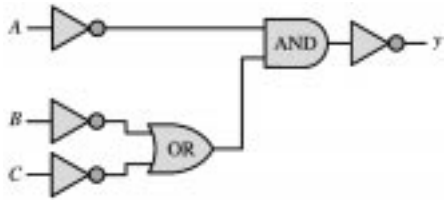
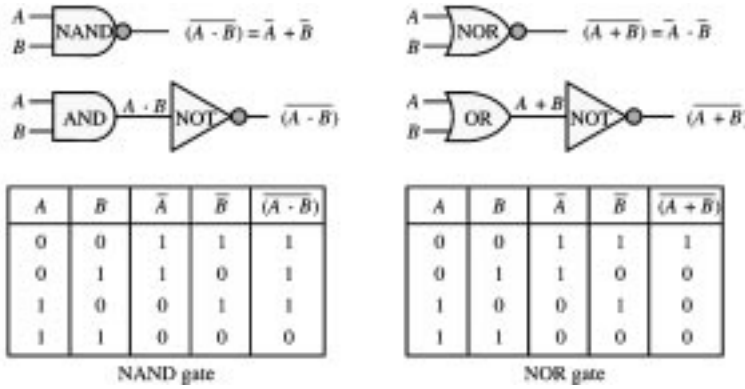


FIGURE 5.10.16 Figure for Example 5.10.1.

NAND and NOR Gates

In addition to the AND and OR gates we have just analyzed, the complementary forms of these gates, called NAND and NOR, are very commonly used in practice. In fact, NAND and NOR gates form the basis of most practical logic circuits. Figure 5.10.17 depicts these two gates and illustrates how they can be easily interpreted in terms of AND, OR, and NOT gates by virtue of De Morgan's laws. You can readily verify that the logic function implemented by the NAND and NOR gates corresponds, respectively, to AND and OR gates followed by an inverter. It is very important to note that, by De Morgan's laws, the NAND gate performs a *logical addition* on the *complements* of the inputs. while the NOR gate performs a *logical multiplication* on the *complements* of the inputs. Functionally, then, any logic function could be implemented with either NOR or NAND gates only.



A	B	$\overline{A}$	$\overline{B}$	$\overline{(A \cdot B)}$
0	0	1	1	1
0	1	1	0	1
1	0	0	1	1
1	1	0	0	0

NAND gate

A	B	$\overline{A}$	$\overline{B}$	$\overline{(A + B)}$
0	0	1	1	1
0	1	1	0	0
1	0	0	1	0
1	1	0	0	0

NOR gate

FIGURE 5.10.17 Equivalence of NAND and NOR gates with AND and OR gates.

In the next section we shall learn how to systematically approach the design of logic functions. First, we provide a few examples to illustrate logic design with NAND and NOR gates.

Example 5.10.2

Realize the following function using only NAND and NOR gates:

$$y = \overline{\left(\overline{A \cdot B}\right) + C}$$

Solution. Since the term in parentheses appears as a complement product, it can be obtained by means of a NAND gate. Further, once the function $\overline{(A \cdot B)}$ has been realized, we can see that y is the

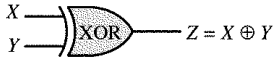
complemented sum of two terms — that is, it can be obtained directly with a NOR gate. The resulting logic circuit is shown in Figure 5.10.18.

Can you find another solution to this problem that employs only two gates?

FIGURE 5.10.18 Figure for Example 5.10.2.

The XOR (Exclusive OR) Gate

It is rather common practice for a manufacturer of integrated circuits to provide common combinations of logic circuits in a single integrated circuit package. An example of this idea is provided by the **exclusive OR** (XOR) **gate**, which provides a logic function similar, but not identical, to the OR gate we have already studied. The XOR gate acts as an OR gate, except when its inputs are all logic 1s; in this case, the output is a logic 0 (thus the term *exclusive)*. Figure 5.10.19 shows the logic circuit symbol adopted for this gate, and the corresponding truth table. The logic function implemented by the XOR gate is the following: "either *X* or *Y*, but not both." This description can be extended to an arbitrary number of inputs.



X	Y	Z
0	0	0
0	1	1
1	0	1
1	1	0

Truth table

FIGURE 5.10.19 XOR gate.

The symbol adopted for the exclusive OR operation is ⊕, and so we shall write

$$Z = X \oplus Y$$

to denote this logic operation. The XOR gate can be obtained by a combination of the basic gates we are already familiar with. For example, if we observe that the XOR function corresponds to $Z = X \oplus Y = (X + Y) \cdot X \cdot \overline{(X + Y)}$, we can realize the XOR gate by means of the circuit shown in Figure 5.10.20.

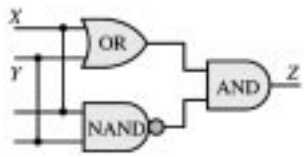


FIGURE 5.10.20 Realization of an XOR gate.

Karnaugh Maps and Logic Design

In examining the design of logic functions by means of logic gates, we have discovered that more than one solution is usually available for the implementation of a given logic expression. It should also be clear by now that some combinations of gates can implement a given function more efficiently than others. How can we be assured of having chosen the most efficient realization? Fortunately, there is a procedure that utilizes a map describing all possible combinations of the variables present in the logic function of interest. This map is called a **Karnaugh map**, after its inventor. Figure 5.10.21 depicts the appearance of Karnaugh maps for two-, three-, and four-variable expressions in two different forms. As can be seen, the row and column assignments for two or more variables are arranged so that all adjacent terms change by only one bit. For example, in the three- or four-variable map, the columns next to

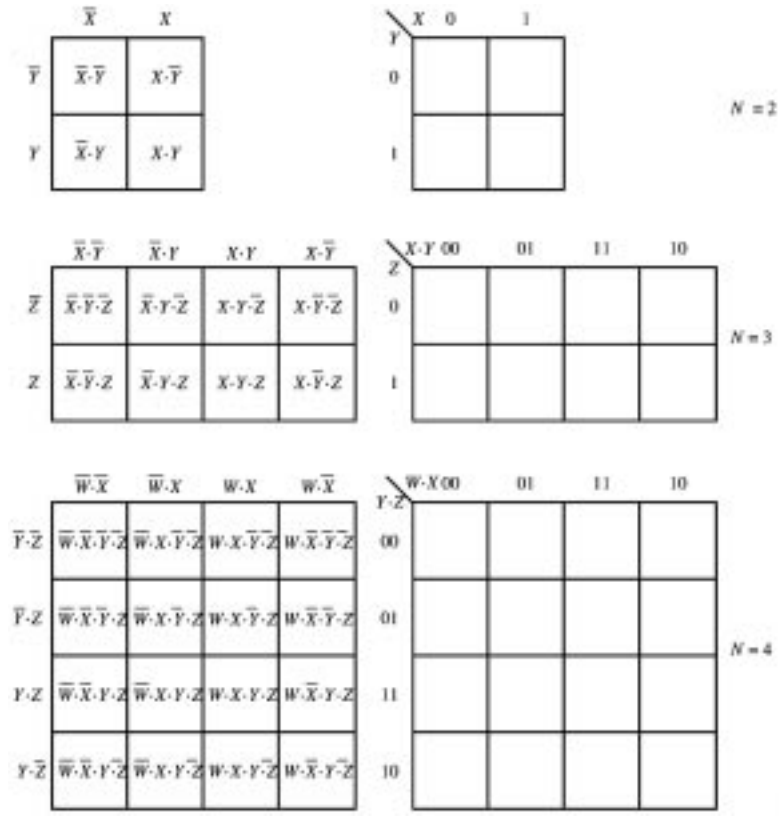


FIGURE 5.10.21 Two-, three-, and four-variable Karnaugh maps.

column 01 are columns 00 and 10. Also note that each map consists of 2^N **cells**, where N is the number of logic variables.

Each cell in a Karnaugh map contains a **minterm**, that is, a product of the N variables that appear in our logic expression (in either uncomplemented or complemented form). For example, for the case of three variables ($N = 3$), there are $2^3 = 8$ such combination, or minterms: $\overline{X} \cdot \overline{Y} \cdot \overline{Z}$, $\overline{X} \cdot \overline{Y} \cdot Z$, $\overline{X} \cdot Y \cdot \overline{Z}$, $\overline{X} \cdot Y \cdot Z$, $X \cdot \overline{Y} \cdot \overline{Z}$, $X \cdot \overline{Y} \cdot Z$, $X \cdot Y \cdot \overline{Z}$, and $X \cdot Y \cdot Z$. The content of each cell — that is, the minterm — is the product of the variables appearing at the corresponding vertical and horizontal coordinates. For example, in the three-variable map, $X \cdot Y \cdot \overline{Z}$ appears at the intersection of $X \cdot Y$ and $\overline{Z}$. The map is filled by placing a value of 1 for any combination of variables for which the desired output is a 1. For example, consider the function of three variables for which we desire to have an output of 1 whenever the variables X, Y, and Z have the following values:

$$\begin{array}{lll} X = 0 & Y = 1 & Z = 0 \\ X = 0 & Y = 1 & Z = 1 \\ X = 1 & Y = 1 & Z = 0 \\ X = 1 & Y = 1 & Z = 1 \end{array}$$

The same truth table is shown in Figure 5.10.22 together with the corresponding Karnaugh map.

The Karnaugh map provides an immediate view of the values of the function in graphical form. Further, the arrangement of the cells in the Karnaugh map is such that any two adjacent cells contain minterms that vary in only one variable. This property, as will be verified shortly, is quite useful in the design of logic functions by means of logic gates, especially if we consider the map to be continuously

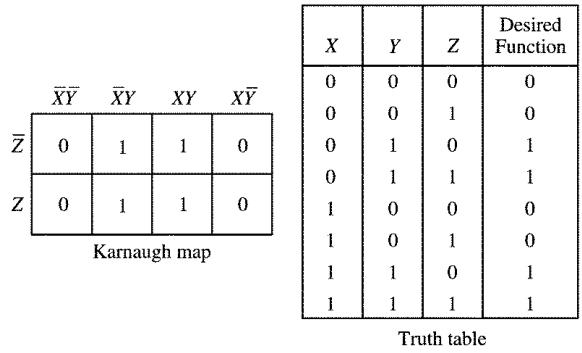


FIGURE 5.10.22 Truth table and Karnaugh map representations of a logic function.

wrapping around itself, as if the top and bottom, and right and left edges were touching each other. For the three-variable map given in Figure 5.10.21, for example, the cell $X \cdot \overline{Y} \cdot \overline{Z}$ is adjacent to $\overline{X} \cdot \overline{Y} \cdot \overline{Z}$ if we "roll" the map so that the right edge touches the left. Note that these two cells differ only in the variable X, a property we earlier claimed adjacent cells have.

Shown in Figure 5.10.23 is a more complex, four-variable logic function which will serve as an example in explaining how Karnaugh maps can be used directly to implement a logic function. First, we define a subcube as a set of 2^m adjacent cells, for $m = 1, 2, 3, \ldots, N$. Thus, a subcube can consist of 1, 2, 4, 8, 16, 32, … cells. All possible subcubes for the four-variable map of Figure 5.10.23 are shown in Figure 5.10.24. Note that there are no four-cell subcubes in this particular case. Note also that there is some overlap between subcubes.

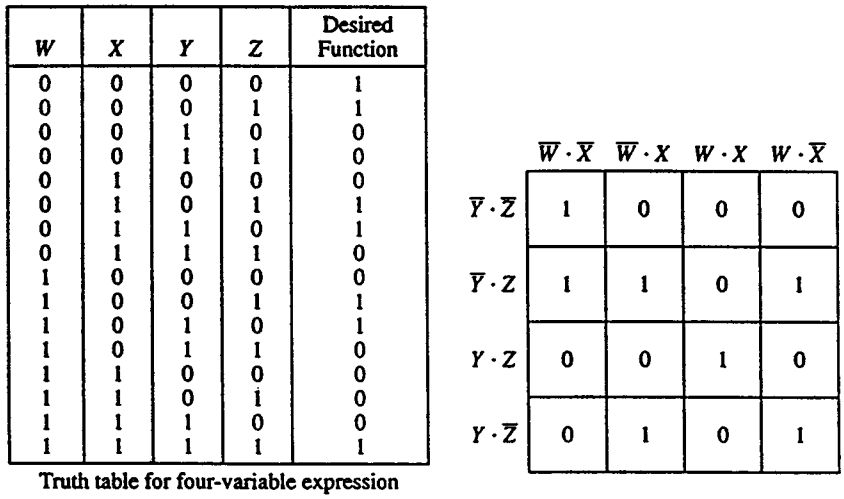


FIGURE 5.10.23 Karnaugh map for a four-variable expression.

Sum-of-Products Realizations

Although not explicitly stated, the logic functions of the preceding section were all in sum-of-products form. As you know, it is also possible to realize logic functions in product-of-sums form. This section discusses the implementation of logic functions in sum-of-products form and gives a set of design rules. The next section will do the same for product-of-sums form logical expressions. The following rules are a useful aid in determining the minimal sum-of-products expression:

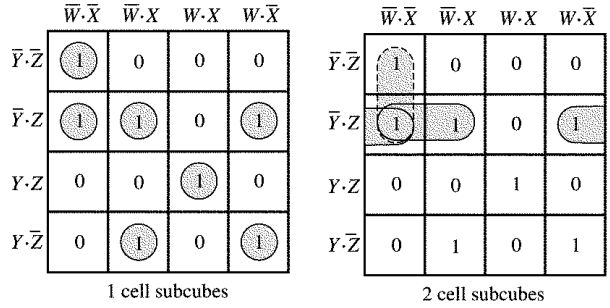


FIGURE 5.10.24 One- and two-cell subcubes for the Karnaugh map of Figure 5.10.23.

1. Begin with isolated cells. These must be used as they are, since no simplification is possible.
2. Find all cells that are adjacent to only one other cell, forming two-cell subcubes.
3. Find cells that form four-cell subcubes, eight-cell subcubes, and so forth.
4. The minimal expression is formed by the collection of the *smallest number of maximal subcubes.*

The following examples illustrate the application of these principles to a variety of problems.

Product-of-Sums Realizations

Thus far, we have exclusively worked with sum-of-products expressions, that is, logic functions of the form $A \cdot B + C \cdot D$. We know, however, that De Morgan's laws state that there is an equivalent form that appears as a product of sums, for example, $(W + Y) \cdot (Y + Z)$. The two forms are completely equivalent, logically, but one of the two forms may lead to a realization involving a smaller number of gates. When using Karnaugh maps, we may obtain the product-of-sums form very simply by following these rules:

1. Solve for the 0s exactly as for the 1s in sum-of-products expressions.
2. Complement the resulting expression.

The same principles stated earlier apply in covering the map with subcubes and determining the minimal expression. The following examples illustrate how one form may result in a more efficient solution than the other.

Example 5.10.3

This example illustrates the design of a logic function using both sum-of-products and product-of-sums implementations, thus showing that it may be possible to realize some savings by using one implementation rather than the other.

1. Realize the function *f* by a Karnaugh map using 0s.
2. Realize the function *f* by a Karnaugh map using 1s.

x	y	z	f
0	0	0	0
0	0	1	1
0	1	0	1
0	1	1	1
1	0	0	1
1	0	1	1
1	1	0	0
1	1	1	0

Solution.

1. Using 0s, we obtain the Karnaugh map of Figure 5.10.25, leading to the product-of-sums expression

$$f = (x + y + z) \cdot (\bar{x} + \bar{y})$$

which requires five gates.

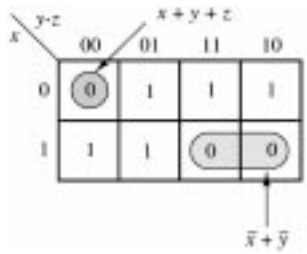


FIGURE 5.10.25 Figure for Example 5.10.3.

2. If 1s are used, as shown in Figure 5.10.26, a sum-of-products expression is obtained, of the form

$$f = \bar{x} \cdot y + x \cdot \bar{y} + \bar{y} \cdot z$$

which requires seven gates.

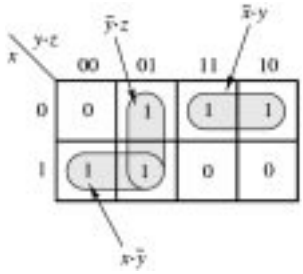


FIGURE 5.10.26 Figure for Example 5.10.3.

Example 5.10.4 Safety Circuit for Operation of a Stamping Press

In this example, the techniques illustrated in the preceding example will be applied to a practical situation. To operate a stamping press, an operator must press two buttons (b_1 and b_2) 1 m apart from each other and away from the press (this ensures that the operator's hands cannot be caught in the press). When the buttons are pressed, the logical variables b_1 and b_2 are equal to 1. Thus, we can define a new variable $A = b_1 \cdot b_2$; when $A = 1$, the operator's hands are safely away from the press. In addition to the safety requirement, however, other conditions must be satisfied before the operator can activate the press. The press is designed to operate on one of two workpieces, part I and part II, but not both. Thus, acceptable logic states for the press to be operated are "part I is in the press, but not part II" and "part II is in the press, but not part I." If we denote the presence of part I in the press by the logical variable $B = 1$ and the presence of part II by the logical variable $C = 1$, we can then impose additional requirements on the operation of the press. For example, a robot used to place either part in the press could activate a pair of switches (corresponding to logical variables B and C) indicating which part, if any, is in the press. Finally, in order for the press to be operable, it must be "ready", meaning that it has to have completed any previous stamping operation. Let the logical variable $D = 1$ represent the ready condition. We have now represented the operation of the press in terms of four logical variables, summarized in the truth

table of Table 5.10.12. Note that only two combinations of the logical variables will result in operation of the press: $ABCD = 1011$ and $ABCD = 1101$. You should verify that these two conditions correspond to the desired operation of the press. Using a Karnaugh map, realize the logic circuitry required to implement the truth table shown.

TABLE 5.10.12 Conditions for Operation of Stamping Press

(*A*) $b_1 \cdot b_2$	(*B*) Part I is in Press	(*C*) Part II is in Press	(*D*) Press is Operable	Press Operation 1 = Pressing; 0 = Not Pressing
0	0	0	0	0
0	0	0	1	0
0	0	1	0	0
0	0	1	1	0
0	1	0	0	0
0	1	0	1	0
0	1	1	0	0
0	1	1	1	0
1	0	0	0	0
1	0	0	1	0
1	0	1	0	0
1	0	1	1	1
1	1	0	0	0
1	1	0	1	1
1	1	1	0	0
1	1	1	1	0

Note: ↑ Both buttons (b_1, b_2) must be pressed for this to be a 1.

Solution. Table 5.10.12 can be converted to a Karnaugh map, as shown in Figure 5.10.27. Since there

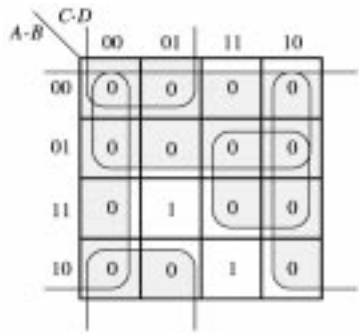


FIGURE 5.10.27 Figure for Example 5.10.4.

are many more 0s than 1s in the table, the use of 0s in covering the map will lead to greater simplification. This will result in a product-of-sums expression. The four subcubes shown in Figure 5.10.27 yield the equation

$$A \cdot D \cdot (C + B) \cdot (\overline{C} + \overline{B})$$

By De Morgan's law, this equation is equivalent to

$$A \cdot D \cdot (C + B) \cdot \overline{(C \cdot B)}$$

which can be realized by the circuit of Figure 5.10.28.

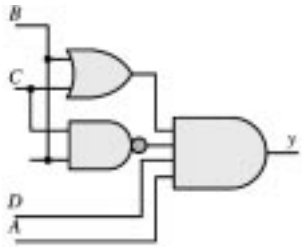


FIGURE 5.10.28 Figure for Example 5.10.4.

For the purpose of comparison, the corresponding sum-of-products circuit is shown in Figure 5.10.29. Note that this circuit employs a greater number of gates and will therefore lead to a more expensive design.

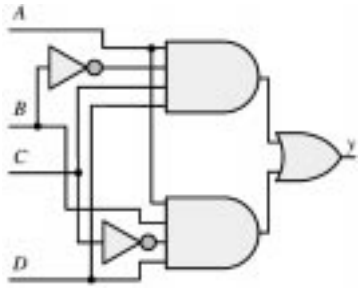


FIGURE 5.10.29 Figure for Example 5.10.4.

Don't Care Conditions

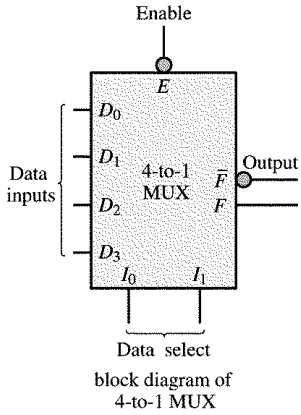
Another simplification technique may be employed whenever the value of the logic function to be implemented can be either a 1 or a 0. This condition may result from the specification of the problem and is not uncommon. Whenever it does not matter whether a position in the map is filled by a 1 or a 0, we use a so-called **don't care** entry, denoted by an x. Then the don't care can be used as either a 1 or a 0, depending on which results in a greater simplification (i.e., helps in forming the smallest number of maximal subcubes).

Combinational Logic Modules

The basic logic gates described in the previous section are used to implement more advanced functions and are often combined to form logic modules, which, thanks to modern technology, are available in compact integrated circuit (IC) packages. In this section and the next, we discuss a few of the more common **combinational logic modules**, illustrating how these can be used to implement advanced logic function.

Multiplexers

Multiplexers, or **data selectors**, are combinational logic circuits that permit the selection of one of many inputs. A typical multiplexer (MUX) has 2^n **data lines**, n **address lines**, and one output. In addition, other control inputs (e.g., enables) may exist. Standard, commercially available MUXs allow for n up to 4; however, two or more MUXs can be combined if a greater range is needed. The MUX allows for one of 2^n inputs to be selected as the data output; the selection of which input is to appear at the output is made by way of the address lines. Figure 5.10.30 depicts the block diagram of a four-input MUX. The input data lines are labeled D_0, D_1, D_2, and D_3; the **data select**, or address, **lines** are labeled I_0 and I_1; and the output is available in both complemented and uncomplemented form, and is thus labeled F, or $\overline{F}$. Finally, an **enable** input, labeled E, is also provided, as a means of enabling or disabling the MUX: if $E = 1$, the MUX is disabled; if $E = 0$, it is enabled. The negative logic (MUX off when $E = 1$



I_1	I_0	F
0	0	D_0
0	1	D_1
1	0	D_2
1	1	D_3

Truth table of 4-to-1 MUX

FIGURE 5.10.30 4:1 MUX.

and on when $E = 0$) is represented by the small "bubble" at the enable input, which represents a complement operation (just as at the output of NAND and NOR gates). The enable input is useful whenever one is interested in a cascade of MUXs; this would be of interest if we needed to select a line from a large number, say $2^8 = 256$. Then two 4-input MUXs could be used to provide the data selection of 1 of 8.

The material described in the previous sections is quite adequate to describe the internal workings of a multiplexer. Figure 5.10.31 shows the internal construction of 4:1 MUX using exclusively NAND gates (inverters are also used, but the reader will recall that a NAND gate can act as an inverter if properly connected).

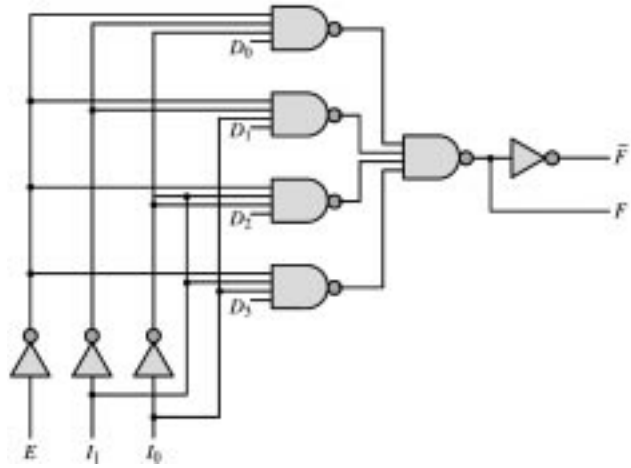


FIGURE 5.10.31 Internal structure of the 4:1 MUX.

In the design of digital systems (for example, microcomputers), a single line is often required to carry two or more different digital signals. However, only one signal at a time can be placed on the line. A MUX will allow us to select, at different instants, the signal we wish to place on this single line. This property is shown here for a 4:1 MUX. Figure 5.10.32 depicts the functional diagram of a 4:1 MUX, showing four data lines, D_0 through D_3, and two select lines, I_0 and I_1.

The data selector function of a MUX is best understood in terms of Table 5.10.13. In this truth table, the x's represent don't care entries. As can be seen from the truth table, the output selects one of the data lines depending on the values of I_1 and I_0, assuming that I_0 is the least significant bit. As an example,

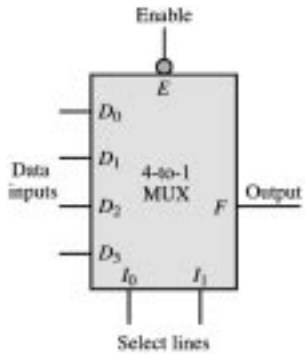


FIGURE 5.10.32 Functional diagram of four-input MUX.

TABLE 5.10.13

I_1	I_0	D_3	D_2	D_1	D_0	F
0	0	x	x	x	0	0
0	0	x	x	x	1	1
0	1	x	x	0	x	0
0	1	x	x	1	x	1
1	0	x	0	x	x	0
1	0	x	1	x	x	1
1	1	0	x	x	x	0
1	1	1	x	x	x	1

$I_1I_0 = 10$ selects D_2, which means that the output, F, will select the value of the data line D_2. Therefore $F = 1$ if $D_2 = 1$ and $F = 0$ if $D_2 = 0$.

Read-Only Memory (ROM)

Another common technique for implementing logic functions uses a **read-only memory**, or ROM. As the name implies, a ROM is a logic circuit that holds in storage ("memory") information — in the form of binary numbers — that cannot be altered but can be "read" by a logic circuit. A ROM is an array of memory cells, each of which can store either a 1 or a 0. The array consists of $2^m \times n$ cells, where n is the number of bits in each word stored in ROM. To access the information stored in ROM, m address lines are required. When an address is selected, in a fashion similar to the operation of the MUX, the binary word corresponding to the address selected appears at the output, which consists of n bits, that is, the same number of bits as the stored words. In some sense, a ROM can be thought of as a MUX that has an output consisting of a word instead of a single bit.

Figure 5.10.33 depicts the conceptual arrangement of a ROM with $n = 4$ and $m = 2$. The ROM table has been filled with arbitrary 4-bit words, just for the purpose of illustration. In Figure 5.10.33, if one were to select an enable input of 0 (i.e., on) and values for the address lines of $I_0 = 0$ and $I_1 = 1$, the output word would be $W_2 = 0110$, so that $b_0 = 0$, $b_1 = 1$, $b_2 = 1$, $b_3 = 0$. Depending on the content of the ROM and the number of address and output lines, one could implement an arbitrary logic function.

Unfortunately, the data stored in read-only memories must be entered during fabrication and cannot be altered later. A much more convenient type of read-only memory is the **erasable programmable read-only memory** (EPROM), the content of which can be easily programmed and stored and may be changed if needed. EPROMs find use in many practical applications because of their flexibility in content and ease of programming. The following example illustrates the use of an EPROM to perform the linearization of a nonlinear function.

Example 5.10.5 EPROM-Based Lookup Table

One of the most common applications of EPROMs is the *arithmetic lookup table*. A lookup table is similar in concept to the familiar multiplication table and is used to store precomputed values of certain

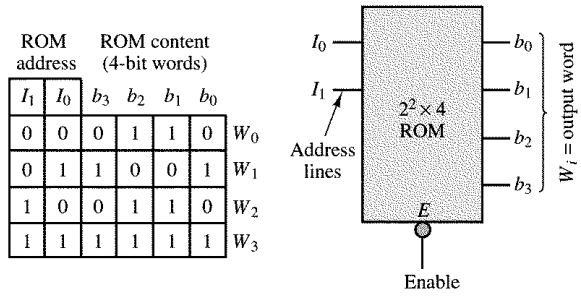


FIGURE 5.10.33 Read-only memory.

functions, eliminating the need for actually computing the function. A practical application of this concept is present in every automobile manufactured in the U.S. since the early 1980s, as part of the exhaust emission control system. In order for the catalytic converter to minimize the emissions of exhaust gases (especially hydrocarbons, oxides of nitrogen, and carbon monoxide), it is necessary to maintain the *air-to-fuel ratio* (A/F) as close as possible to the stoichiometric value, that is, 14.7 parts of air for each part of fuel. Most modern-day engines are equipped with fuel injection systems that are capable of delivering accurate amounts of fuel to each individual cylinder; thus, the task of maintaining an accurate A/F amounts to measuring the mass of air that is aspirated into each cylinder and computing the corresponding mass of fuel. Many automobiles are equipped with a *mass airflow sensor*, capable of measuring the mass of air drawn into each cylinder during each engine cycle. Let the output of the mass airflow sensor be denoted by the variable M_A, and let this variable represent the mass of air (in g) actually entering a cylinder during a particular stroke. It is then desired to compute the mass of fuel, M_F (also expressed in g), required to achieve an A/F of 14.7. This computation is simply

$$M_F = \frac{M_A}{14.7}$$

Although the above computation is a simple division, its actual calculation in a low-cost digital computer (such as would be used on an automobile) is rather complicated. It would be much simpler to tabulate a number of values of M_A, to precompute the variable M_F, and then to store the result of this computation into an EPROM. If the EPROM address were made to correspond to the tabulated values of air mass and the content at each address to the corresponding fuel mass (according to the precomputed values of the expression $M_F = M_A/14.7$), it would not be necessary to perform the division by 14.7. For each measurement of air mass into one cylinder, an EPROM address is specified and the corresponding content is read. The content at the specific address is the mass of fuel required by the particular cylinder.

In practice, the fuel mass needs to be converted into a time interval corresponding to the duration of time during which the fuel injector is open. This final conversion factor can also be accounted for in the table. Suppose, for example, that the fuel injector is capable of injecting K_F g of fuel per second; then the time duration, T_F, during which the injector should be open in order to inject M_F g of fuel into the cylinder is given by

$$T_F = \frac{M_F}{K_F} \text{s}$$

Therefore, the complete expression to be precomputed and stored in the EPROM is

$$T_F = \frac{M_A}{14.7 \times K_F} \text{s}$$

Figure 5.10.34 illustrates this process graphically.

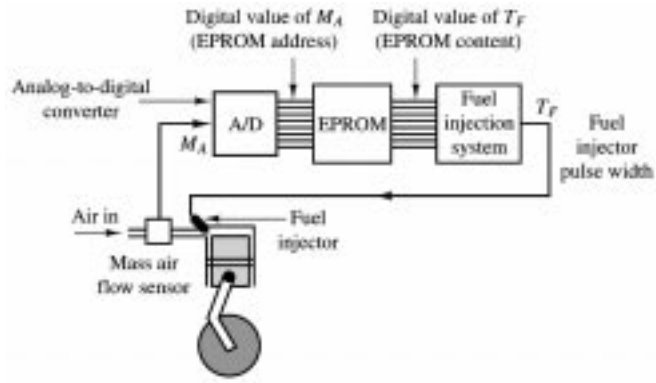


FIGURE 5.10.34 Use of EPROM lookup table in automotive fuel injection system.

To provide a numerical illustration, consider a hypothetical engine capable of aspirating air in the range $0 < M_A < 0.51$ g and equipped with fuel injectors capable of injecting at the rate of 1.36 g/sec. Thus, the relationship between T_F and M_A is

$$T_F = 50 \times M_A \text{ msec} = 0.05 M_A \text{ sec}$$

If the digital value of M_A is expressed in dg (decigrams, or tenths of g), the lookup table of Figure 5.10.35 can be implemented, illustrating the conversion capabilities provided by the EPROM. Note that in order to represent the quantities of interest in an appropriate binary format compatible with the 8-bit EPROM, the units of air mass and of time have been scaled.

M_A(g) × 10^{-2}	Address (digital value of M_A)	Content (digital value of T_F)	T_F(ms) × 10^{-1}
0	00000000	00000000	0
1	00000001	00000101	5
2	00000010	00001010	10
3	00000011	00001111	15
4	00000100	00010100	20
5	00000101	00011001	25
⋮	⋮	⋮	⋮
51	00110011	11111111	255

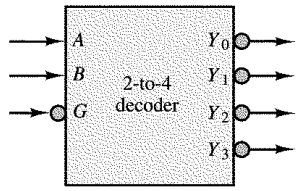
FIGURE 5.10.35 Lookup table for automotive fuel injection application.

Decorders and Read and Write Memory

Decoders, which are commonly used for applications such as address decoding or memory expansion, are combinational logic circuits as well. Our reason for introducing decoders is to show some of the internal organization of semiconductor memory devices.

Figure 5.10.36 shows the truth table for a 2:4 decoder. The decoder has an enable input, $\overline{G}$, and select inputs, B and A. It also has four outputs, Y_0 through Y_3. When the enable input is logic 1, all decoder outputs are forced to logic 1 regardless of the select inputs.

This simple description of decoders permits a brief discussion of the internal organization of an SRAM (**static random-access** or **read and write memory**). SRAM is internally organized to provide memory with high speed (i.e., short access time), a large bit capacity, and low cost. The memory array in this memory device has a column length equal to the number of words, W, and a row length equal to the



Inputs			Outputs			
Enable	Select					
$\overline{G}$	A	B	Y_0	Y_1	Y_2	Y_3
1	**X**	**X**	1	1	1	1
0	0	0	0	1	1	1
0	0	1	1	0	1	1
0	1	0	1	1	0	1
0	1	1	1	1	1	0

FIGURE 5.10.36 2:4 decoder.

number of bits per word, N. To select a word, an n-to-W decoder is needed. Since the address inputs to the decoder select only one of the decoder's outputs, the decoder selects one word in the memory array. Figure 5.10.37 shows the internal organization of a typical SRAM.

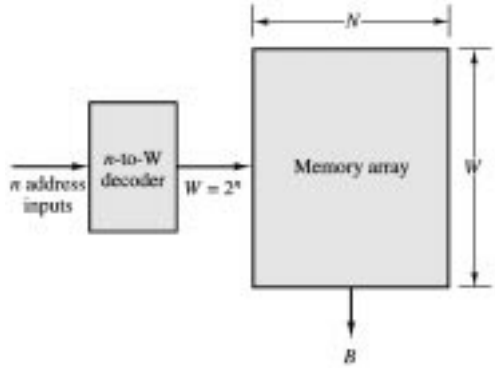


FIGURE 5.10.37 Internal organization of SRAM.

Thus, to choose the desired word from the memory array, the proper address inputs are required. As an example, if the number of words in the memory array is 8, a 3:8 decoder is needed. Data sheets for 2:4 and 3:8 decoders from a CMOS family data book are provided at the end of the chapter.

Sequential Logic Modules

Combinational logic circuits provide outputs that are based on a combination of present inputs only. On the other hand, sequential logic circuits depend on present and past input values. Because of this "memory" property, sequential circuits can store information; this capability opens a whole new area of application for digital logic circuits.

Latches and Flip-Flops

The basic information-storage device in a digital circuit is called a **flip-flop**. There are many different varieties of flip-flops; however, all flip-flops share the following characteristics:

1. A flip-flop is a **bistable device**; that is, it can remain in one of two stable states (0 and 1) until appropriate conditions cause it to change state. Thus, a flip-flop can serve as a memory element.
2. A flip-flop has two outputs, one of which is the complement of the other.

RS Flip-Flop. It is customary to depict flip-flops by their block diagram and a name — such as Q or X — representing the output variable. Figure 5.10.38 represents the so-called **RS flip-flop**, which has two inputs, denoted by S and R, and two outputs, Q and $\overline{Q}$. The value of Q is called the *state* of the flip-

flop. If Q =1, we refer to the device as *being in the 1 state*. Thus, we need define only one of the two outputs of the flip-flop. The two inputs, R and S, are used to change the state of the flip-flop, according to the following rules:

1. When $R = S = 0$, the flip-flop remains in its present state (whether 1 or 0).
2. When $S = 1$ and $R = 0$, the flip-flop is *set* to the 1 state (thus, the letter S, for **set**).
3. When $S = 0$ and $R = 1$, the flip-flop is *reset* to the 0 state (thus, the letter R, for **reset**).
4. It is not permitted for both S and R to be equal to 1. (This would correspond to requiring the flip-flop to set and reset at the same time.)

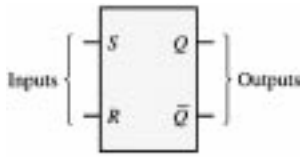


FIGURE 5.10.38 *RS* flip-flop.

The rules just described are easily remembered by noting that 1s on the S and R inputs correspond to the set and reset commands, respectively.

A convenient means of describing the series of transitions that occur as the signals sent to the flip-flop inputs change is the **timing diagram**. A timing diagram is a graph of the inputs and outputs of the *RS* flip-flop (or any other logic device) depicting the transitions that occur over time. In effect, one could also represent these transitions in tabular form; however, the timing diagram provides a convenient visual representation of the evolution of the state of the flip-flop. Figure 5.10.39 depicts a table of transitions for an *RS* flip-flop Q, as well as the corresponding timing diagram.

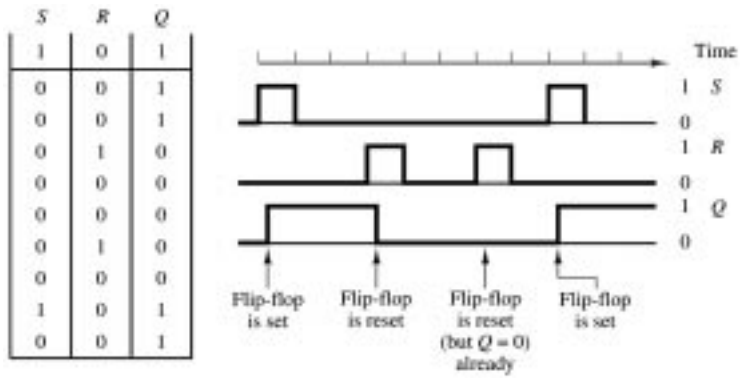


FIGURE 5.10.39 Timing diagram for the *RS* flip-flop.

It is important to note that the *RS* flip-flop is **level-sensitive**. This means that the set and reset operations are completed only after the R and S inputs have reached the appropriate levels. Thus, in Figure 5.10.39 we show the transitions in the Q outputs as occurring with a small delay relative to the transitions in the R and S inputs.

It is instructive to illustrate how an *RS* flip-flop can be constructed using simple logic gates. For example, Figure 5.10.40 depicts a realization of such a circuit consisting of four gates: two inverters and two NAND gates (actually, the same result could be achieved with four NAND gates). Consider the case in which the circuit is in the initial state $Q = 0$ (and therefore $\overline{Q} = 1$). If the input $S = 1$ is applied, the top NOT gate will see inputs $\overline{Q} = 1$ and $\overline{S} = 0$, so that $Q = (\overline{S} \cdot \overline{Q}) = (\overline{0 \cdot 1}) = 1$ — that is, the flip-flop is set. Note that when Q is set to 1, $\overline{Q}$ becomes 0. This, however, does not affect the state of the Q output, since replacing $\overline{Q}$ with 0 in the expression

$$Q = \left(\overline{\overline{S} \cdot \overline{Q}}\right)$$

does not change the result:

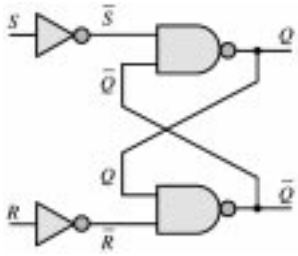


FIGURE 5.10.40 Logic gate implementation of the *RS* flip-flop.

$$Q = \left(\overline{0 \cdot 0}\right) = 1$$

Thus, the cross-coupled feedback from outputs Q and $\overline{Q}$ to the input of the NAND gates is such that the set condition sustains itself. It is straightforward to show (by symmetry) that a 1 input on the R line causes the device to reset (i.e., causes $Q = 0$) and that this condition is also self-sustaining.

An extension of the *RS* flip-flop includes an additional enable input that is *gated* into each of the other two inputs. Figure 5.10.41 depicts an *RS* flip-flop consisting of two NOR gates. In addition, an enable input is connected through two AND gates to the *RS* flip-flop, so that an input to the R and S line will be effective only when the enable input is 1. Thus, any transitions will be controlled by the enable input, which acts as a synchronizing signal. The enable signal may consist of a **clock**, in which case the flip-flop is said to be **clocked** and its operation is said to be **synchronous**.

The same circuit of Figure 5.10.41 can be used to illustrate two additional features of flip-flops: the **preset** and **clear** functions, denoted by the inputs P and C, respectively. When P and C are 0, they do not affect the operation of the flip-flop. Setting $P = 1$ corresponds to setting $S = 1$, and therefore causes the flip-flop to go into the 1 state, thus, the term *preset*: this function allows the user to preset the flip-flop to 1 at any time. When C is 1, the flip-flop is reset, or *cleared* (i.e., Q is made equal to 0). Note that these direct inputs are, in general, asynchronous; therefore, they allow the user to preset or clear the flip-flop at any time. A set of timing waveforms illustrating the function of the enable, preset, and clear inputs is also shown in Figure 5.10.41. Note how transitions occur only when the enable input goes high (unless the preset or clear inputs are used to override the *RS* inputs).

Another extension of the *RS* flip-flop, called the **data latch**, is shown in Figure 5.10.42. In this circuit, the R input is always equal to the inverted S input, so that whenever the enable input is high, the flip-flop is set. This device has the dual advantage of avoiding the potential conflict that might arise if both R and S were high and reducing the number of input connections by eliminating the reset input. This circuit is called a data latch because once the enable input goes low, the flip-flop is latched to the previous value of the input. Thus this device can serve as a basic memory element.

D Flip-Flop. The ***D* flip-flop** is an extension of the data latch that utilizes two *RS* flip-flops, as shown in Figure 5.10.43. In this circuit, a clock is connected to the enable input of each flip-flop. Since Q_1 sees an inverted clock signal, the latch is enabled when the clock waveform goes low. However, since Q_2 is disabled when the clock is low, the output of the D flip-flop will not switch to the 1 state until the clock goes high, enabling the second latch and transferring the state of Q_1 to Q_2. It is important to note that the D flip-flop changes state only on the positive edge of the clock waveform: Q_1 is set on the negative edge of the clock, and Q_2 (and therefore Q) is set on the positive edge of the clock, as shown in the timing diagram of Figure 5.10.43. This type of device is said to be **edge-triggered**. This feature is indicated by the "knife edge" drawn next to the CLK input in the device symbol. The particular device

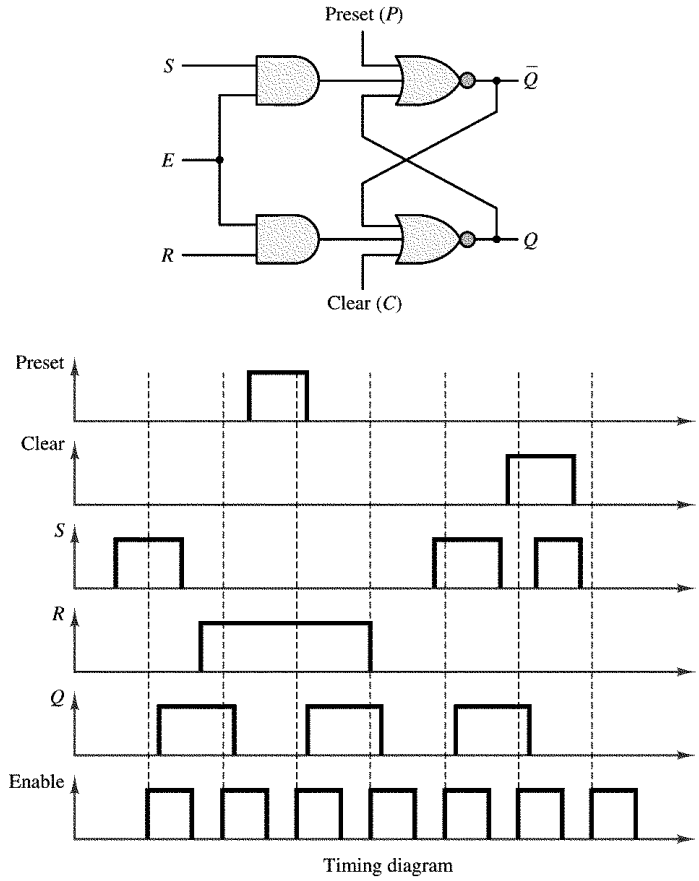


FIGURE 5.10.41 *RS* flip-flop with enable, preset, and clear lines.

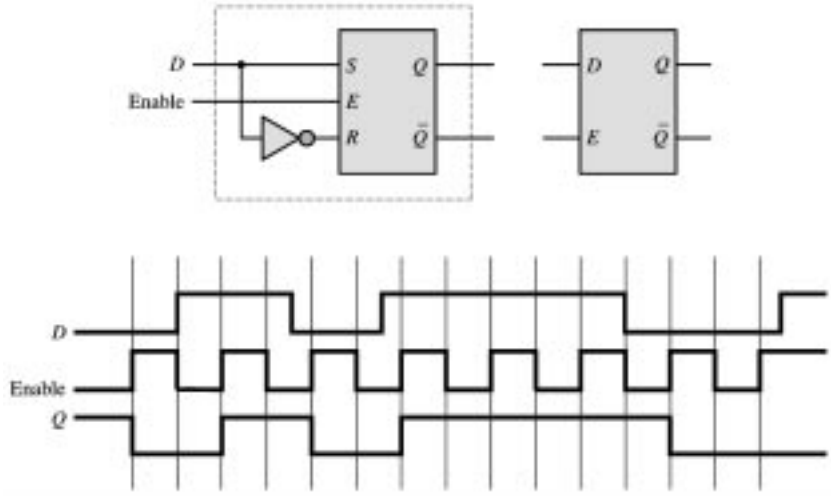


FIGURE 5.10.42 Data latch.

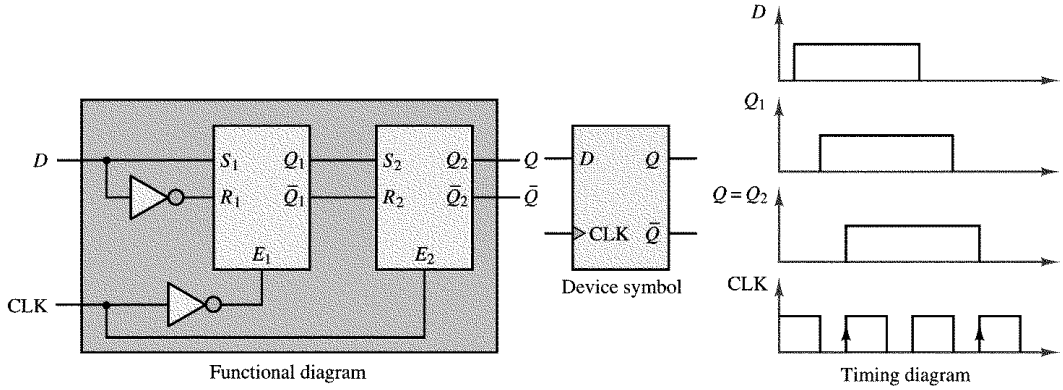


FIGURE 5.10.43 D flip-flop.

described here is said to be positive edge-triggered, or **leading edge-triggered**, since the final output of the flip-flop is set on a positive-going clock transition.

On the basis of the rules stated in this section, the state of the *D* flip-flop can be described by means of the following truth table:

D	CLK	*Q*
0	↑	0
1	↑	1

where the symbol ↑ indicates the occurrence of a positive transition.

JK Flip-Flop. Another very common type of flip-flop is the ***JK* flip-flop**, shown in Figure 5.10.44. The *JK* flip-flop operates according to the following rules:

- When *J* and *K* are both low, no change occurs in the state of the flip-flop.
- When $J = 0$ and $K = \downarrow$, the flip-flop is reset to 0.
- When $J = \downarrow$ and $K = 0$, the flip-flop is set to 1.
- When both *J* and *K* are high, the flip-flop will toggle between states at every transition of the clock input.

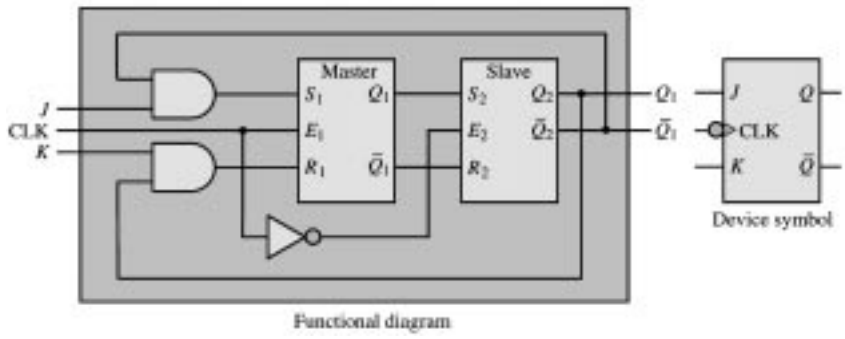


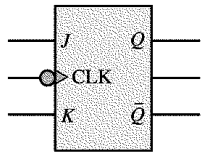
FIGURE 5.10.44 *JK* flip-flop.

The symbol ↓ denotes a negative transition.

Note that, functionally, the operation of the *JK* flip-flop can also be explained in terms of two *RS* flip-flops. When the clock waveform goes high, the "master" flip-flop is enabled; the "slave" receives the

state of the master upon a negative clock transition. The "bubble" at the clock input signifies that the device is negative or **trailing edge-triggered**. This behavior is similar to that of an *RS* flip-flop, except for the $J = 1, K = 1$ condition, which corresponds to a toggle mode rather than to a disallowed combination of inputs.

Figure 5.10.45 depicts the truth table for the *JK* flip-flop. It is important to note that when both inputs are 0 the flip-flop remains in its previous state at the occurrence of a clock transition; when either input is high and the other is low, the *JK* flip-flop behaves like the *RS* flip-flop, whereas if both inputs are high, the output "toggles" between states every time the clock waveform undergoes a negative transition.



JK flip-flop

J_n	K_n	Q_{n+1}
0	0	Q_n
0	1	0 (reset)
1	0	1 (set)
1	1	$\bar{Q}_n$ (toggle)

FIGURE 5.10.45 Truth table for the *JK* flip-flop.

Digital Counters

One of the more immediate applications of flip-flops is in the design of **counters**. A counter is a sequential logic device that can take one of *N* possible states, stepping through these states in a sequential fashion. When the counter has reached its last state, it resets to zero and is ready to start counting again. For example, a three-bit **binary up counter** would have $2^3 = 8$ possible states, and might appear as shown in the functional block of Figure 5.10.46. The input clock waveform causes the counter to step through the eight states, making one transition for each clock pulse. We shall shortly see that a string of *JK* flip-flops can accomplish this task exactly. The device shown in Figure 5.10.46 also displays a reset input, which forces the counter to equal 0: $b_2b_1b_0 = 000$.

Although binary counters are very useful in many applications, one is often interested in a **decade counter**, that is, a counter that counts from 0 to 9 and then resets. A four-bit binary counter can easily be configured in principle to provide this function by means of simple logic that resets the counter when it has reached the count $1001_2 = 9_{10}$. As shown in Figure 5.10.47, if we connect bits b_3 and b_1 to a four-input AND gate, along with $\bar{b}_2$ and $\bar{b}_0$, the output of the AND gate can be used to reset the counter after a count of 10. Additional logic can provide a "carry" bit whenever a reset condition is reached, which could be passed along to another decade counter, enabling counts up to 99. Decade counters can be cascaded so as to represent decimal digits in succession.

Although the decade counter of Figure 5.10.47 is attractive because of its simplicity, this configuration would never be used in practice because of the presence of **propagation delays**. These delays are caused by the finite response time of the individual transistors in each logic device and cannot be guaranteed to be identical for each gate and flip-flop. Thus, if the reset signal — which is presumed to be applied at exactly the same time to each of the four *JK* flip-flops in the four-bit binary counter — does not cause the *JK* flip-flops to reset at exactly the same time on account of different propagation delays, then the binary word appearing at the output of the counter will change from 1001 to some other number, and the output of the four-input NAND gate will no longer be high. In such a condition, the flip-flops that have not already reset will then not be able to reset, and the counting sequence will be irreparably compromised.

What can be done to obviate this problem? The answer is to use a systematic approach to the design of sequential circuits making use of **state transition diagrams**. This topic is discussed in the references.

A simple implementation of the binary counter we have described in terms of its functional behavior is shown in Figure 5.10.48. The figure depicts a three-bit binary **ripple counter**, which is obtained from a cascade of three *JK* flip-flops. The transition table shown in the figure illustrates how the *Q* output of each state becomes the clock input to the next stage, while each flip-flop is held in the toggle mode. The output transitions assume that the clock, CLK, is a simple square wave (all *JK*s are negative edge-triggered).

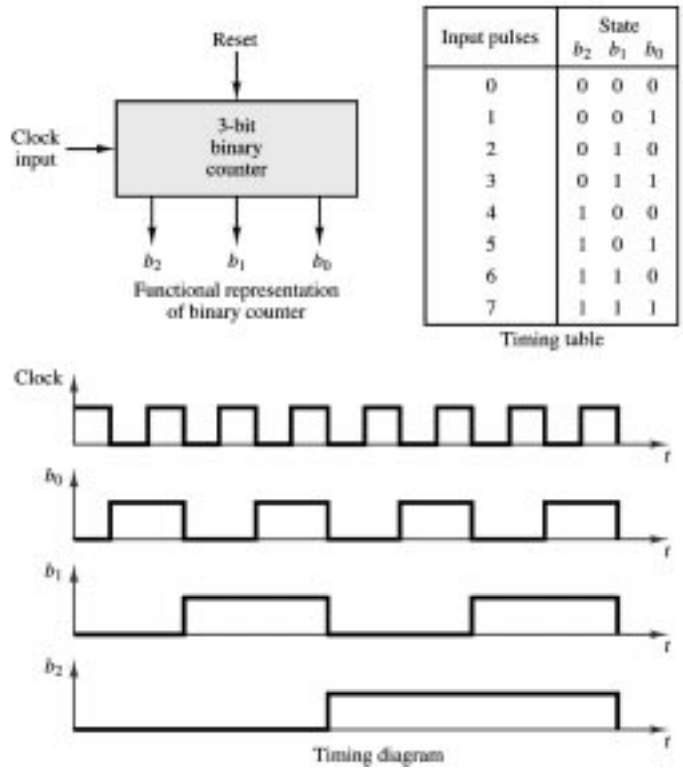


FIGURE 5.10.46 Binary up counter.

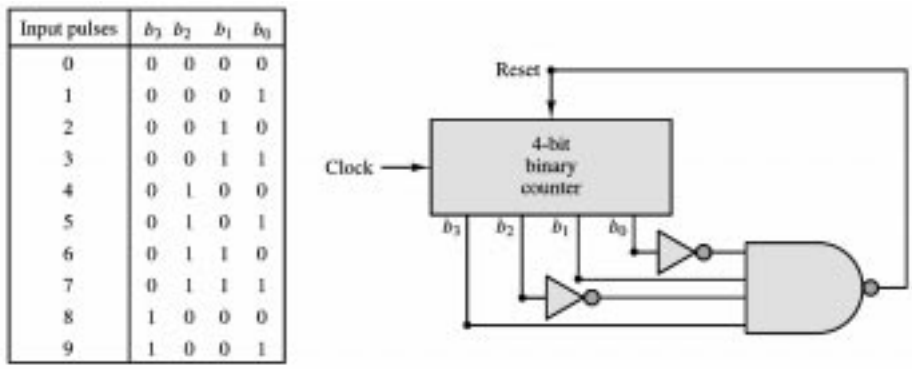


FIGURE 5.10.47 Decade counter.

This 3-bit ripple counter can easily be configured as a divide-by-8 mechanism, simply by adding an AND gate. To divide the input clock rate by 8, one output pulse should be generated for every eight clock pulses. If one were to output a pulse every time a binary 111 combination occurs, a simple AND gate would suffice to generate the required condition. This solution is shown in Figure 5.10.49. Note that the square wave is also included as an input to the AND gate; this ensures that the output is only as wide as the input signal. This application of ripple counters is further illustrated in the following example.

A slightly more complex version of the binary counter is the so-called **synchronous counter**, in which the input clock drives all of the flip-flops simultaneously. Figure 5.10.50 depicts a three-bit synchronous counter. In this figure, we have chosen to represent each flip-flop as a *T* flip-flop. The clocks to all the

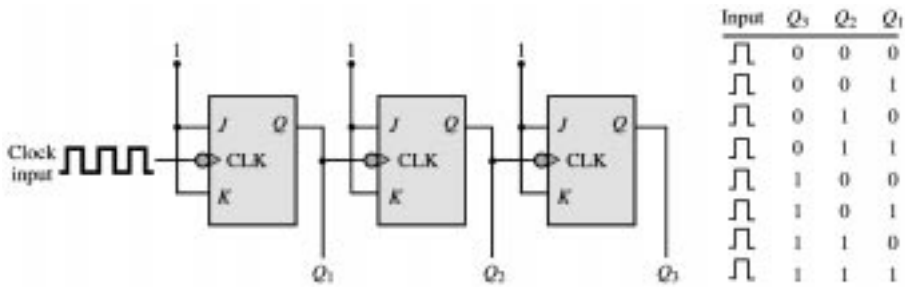


FIGURE 5.10.48 Ripple counter.

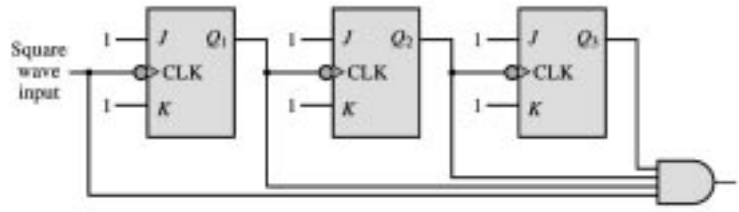


FIGURE 5.10.49 Divide-by-8 circuit.

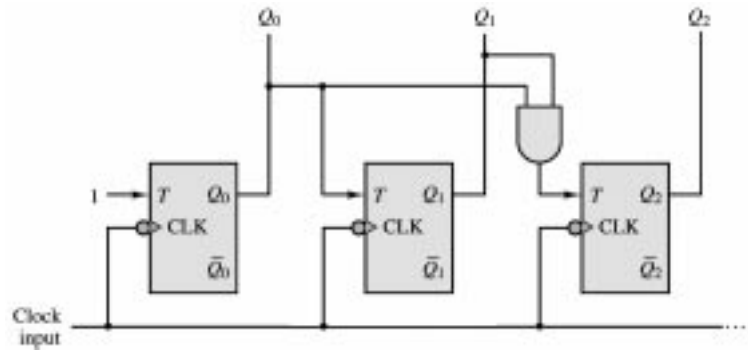


FIGURE 5.10.50 Three-bit synchronous counter.

flip-flops are incremented simultaneously. The reader should verify that Q_0 toggles to 1 first and then Q_1 toggles to 1, and that the AND gate ensures that Q_2 will toggle only after Q_0 and Q_1 have both reached the 1 state ($Q_0 \cdot Q_1 = 1$).

Other common counters are the **ring counter** and the **up-down counter**, which has an additional select input that determines whether the counter counts up or down.

Example 5.10.6 Measurement of Angular Position

One type of angular position encoder is the slotted encoder shown in Figure 5.10.51. This encoder can be used in conjunction with a pair of counters and a high-frequency clock to determine the speed of rotation of the slotted wheel. As shown in Figure 5.10.52, a clock of known frequency is connected to a counter while another counter records the number of slot pulses detected by an optical slot detector as the wheel rotates. Dividing the counter values, one could obtain the speed of the rotating wheel in radians per second. For example, assume a clocking frequency of 1.2 kHz. If both counters are started at zero and at some instant the timer counter reads 2850 and the encoder counter reads 3050, then the speed of the rotating encoder is found to be

$$1200\frac{\text{cycles}}{\text{second}}\cdot\frac{2850\text{ slots}}{3050\text{ cycles}}=1121.3\frac{\text{slots}}{\text{second}}$$

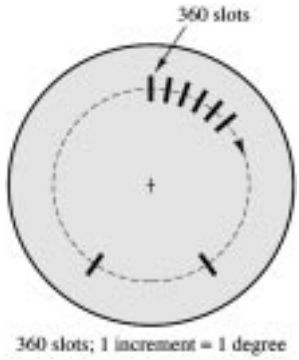


FIGURE 5.10.51

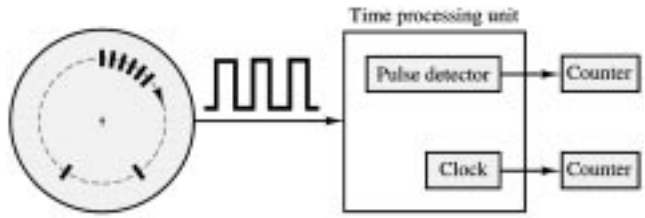


FIGURE 5.10.52 Calculating the speed of rotation of the slotted wheel.

and

$$1121.3\text{ slots/sec}\times 1^\circ\text{ per slot}\times 2\pi/360\text{ rad/degree}=19.6\text{ rad/sec}$$

If this encoder is connected to a rotating shaft, it is possible to measure the angular position and velocity of the shaft. Such shaft encoders are used in measuring the speed of rotation of electric motors, machine tools, engines, and other rotating machinery. A typical application of the slotted encoder is to compute the ignition and injection timing in an automotive engine. In an automotive engine, information related to speed is obtained from the camshaft and the flywheel, which have known reference points. The reference points determine the timing for the ignition firing points and fuel injection pulses and are identified by special slot patterns on the camshaft and crankshaft. Two methods are used to detect the special slots (reference points): *period measurement with additional transition detection* (PMA, and *period measurement with missing transition detection* (PMM). In the PMA method, an additional slot (reference point) determines a known reference position on the crankshaft or camshaft. In the PMM method, the reference position is determined by the absence of a slot. Figure 5.10.53 illustrates a typical PMA pulse sequence, showing the presence of an additional pulse. The additional slot may be used to determine the timing for the ignition pulses relative to a known position of the crankshaft. Figure 5.10.54 depicts a typical PMM pulse sequence. Because the period of the pulses is known, the additional slot of the missing slot can be easily detected and used as a reference position. How would you implement these pulse sequences using ring counters?

Registers

A register consists of a cascade of flip-flops that can store binary data, one bit in each flip-flop. The simplest type of register is the parallel input–parallel output register shown in Figure 5.10.55. In this register, the "load" input pulse, which acts on all clocks simultaneously, causes the parallel inputs $b_0b_1b_2b_3$

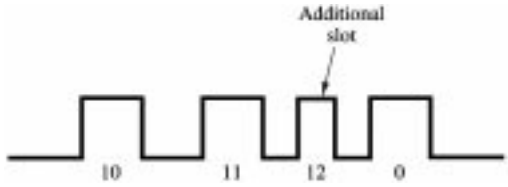


FIGURE 5.10.53 PMA pulse sequence.

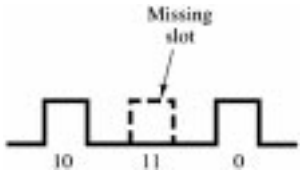


FIGURE 5.10.54 PMA pulse sequence.

to be transferred to the respective flip-flops. The *D* flip-flop employed in this register allows the transfer from b_n to Q_n to occur very directly. Thus, *D* flip-flops are very commonly used in this type of application. The binary word $b_3b_2b_1b_0$ is now "stored", each bit being represented by the state of a flip-flop. Until the "load" input is applied again and a new word appears at the parallel inputs, the register will preserve the stored word.

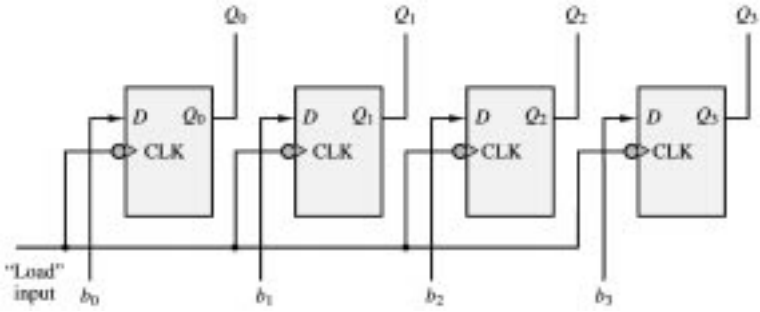


FIGURE 5.10.55 Four-bit parallel register.

The construction of the parallel register presumes that the *N*-bit word to be stored is available in parallel form. However, it is often true that a binary word will arrive in serial form, that is, one bit at a time. A register that can accommodate this type of logic signal is called a **shift register**. Figure 5.10.56 illustrates how the same basic structure of the parallel register applies to the shift register, except that the input is now applied to the first flip-flop and shifted along at each clock pulse. Note that this type of register provides both a serial and a parallel output.

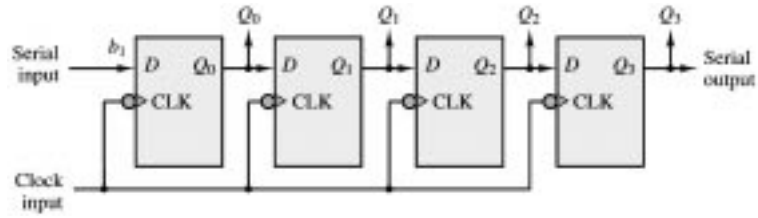


FIGURE 5.10.56 Four-bit shift register.

5.11 Measurements and Instrumentation

Measurement Systems and Transducers

Measurement Systems

In virtually every engineering application there is a need for measuring some physical quantities, such as forces, stresses, temperatures, pressures, flows, or displacements. These measurements are performed by physical devices called **sensors** or **transducers**, which are capable of converting a physical quantity to a more readily manipulated electrical quantity. Most sensors, therefore, convert the change of a physical quantity (e.g., humidity, temperature) to a corresponding (usually proportional) change in an electrical quantity (e.g., voltage or current). Often, the direct output of the sensor requires additional manipulation before the electrical output is available in a useful form. For example, the change in resistance resulting from a change in the surface stresses of a material — the quantity measured by the resistance strain gauges described in Section 5.2 — must first be converted to a change in voltage through a suitable circuit (the Wheatstone bridge) and then amplified from the millivolt to the volt level. The manipulations needed to produce the desired end result are referred to as *signal conditioning*. The wiring of the sensor to the signal conditioning circuitry requires significant attention to *grounding* and *shielding* procedures, to ensure that the resulting signal is as free from noise and interference as possible. Very often, the conditioned sensor signal is then converted to *digital* form and recorded in a computer for additional manipulation, or is displayed in some form. The apparatus used in manipulating a sensor output to produce a result that can be suitably displayed or stored is called a **measurement system**. Figure 5.11.1 depicts a typical computer-based measurement system in block diagram form.

FIGURE 5.11.1 Measurement system.

Sensor Classification

There is no standard and universally accepted classification of sensors. Depending on one's viewpoint, sensors may be grouped according to their physical characteristics (e.g., electronic sensors, resistive sensors), or by the physical variable or quantity measured by the sensor (e.g., temperature, flow rate). Other classifications are also possible. Table 5.11.1 presents a partial classification of sensors grouped according to the variable sensed; we do not claim that the table is complete, but we can safely state that most of the engineering measurements of interest to the reader are likely to fall in the categories listed in Table 5.11.1. Also included in the table are section or example references to sensors described in this chapter.

A sensor is usually accompanied by a set of specifications that indicate its overall effectiveness in measuring the desired physical variable. The following definitions will help the reader understand sensor data sheets:

Accuracy: Conformity of the measurement to the true value, usually in percent of full-scale reading
Error: Difference between measurement and true value, usually in percent of full-scale reading
Precision: Number of significant figures of the measurement
Resolution: Smallest measurable increment
Span: Linear operating range
Range: The range of measurable values
Linearity: Conformity to an ideal linear calibration curve, usually in percent of reading or of full-scale reading (whichever is greater)

TABLE 5.11.1 Sensor Classification

Sensed Variables	Sensors	Ref. in this Chapter
Motion and dimensional variables	Resistive potentiometers	
	Strain gauges	Example 5.2.1
	Differential transformers (LVDTs)	Example 5.12.1
	Variable-reluctance sensors	Example 5.12.2
	Capacitive sensors	Example 5.4.1 and 5.4.2
	Piezoelectric sensors	Example 5.9.2
	Electro-optical sensors	Example 5.10.6
	Moving-coil transducers	Example 5.12.4
	Seismic sensors	
Force, torque, and pressure	Strain gauges	Example 5.2.1
	Piezoelectric sensors	Example 5.9.2
	Capacitive sensors	Example 5.4.1 and 5.4.2
Flow	Pitot tube	
	Hot-wire anemometer	Section 5.11
	Differential pressure sensors	Section 5.11
	Turbine meters	Section 5.11
	Vortex shedding meters	
	Ultrasonic sensors	
	Electromagnetic sensors	
	Imaging systems	
Temperature	Thermocouples	Section 5.11
	Resistance thermometers (RTDs)	Section 5.11
	Semiconductor thermometers	
	Radiation detectors	
Liquid level	Motion transducers	
	Force transducers	
	Differential-pressure measurement devices	
Humidity	Semiconductor sensors	
Chemical composition	Gas analysis equipment	
	Solid-state gas sensors	

Motion and Dimensional Measurements

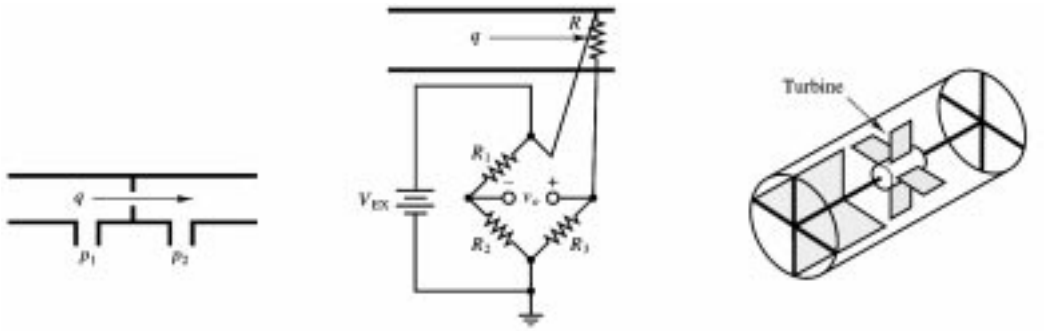
The measurement of motion and dimension is perhaps the most commonly encountered engineering measurement. Measurements of interest include absolute position, relative position (displacement), velocity, acceleration, and jerk (the derivative of acceleration). These can be either translational or rotational measurements; usually, the same principle can be applied to obtain both kinds of measurements. These measurements are often based on changes in elementary properties, such as changes in the resistance of an element (e.g., strain gauges, potentiometers), in an electric field (e.g., capacitive sensors), or in a magnetic field (e.g., inductive, variable-reluctance, or eddy current sensors). Other mechanisms may be based on special materials (e.g., piezoelectric crystals), or on optical signals and imaging systems.

Force, Torque, and Pressure Measurements

Another very common class of measurements is that of pressure and force, and the related measurement of torque. Perhaps the single most common family of force and pressure transducers comprises those based on strain gauges (e.g., load cells, diaphragm pressure transducers). Also very common are piezoelectric transducers. Capacitive transducers again find application in the measurement of pressure.

Flow Measurementrs

In many engineering applications it is desirable to sense the flow rate of a fluid, whether compressible (gas) or incompressible (liquid). The measurement of fluid flow rate is a complex subject; in this section we simply summarize the concepts underlying some of the most common measurement techniques. Shown in Figure 5.11.2 are three different types of flow rate sensors. The sensor in Figure 5.11.2(a) is based on **differential pressure measurement** and on a **calibrated orifice**: the relationship between



Differential-pressure flowmeter: A calibrated orifice and a pair of pressure transducers permit the measurement of flow rate. (a)

Hot-wire anemometer: A heated wire is cooled by the gas flow. The resistance of the wire changes with temperature. (b)

Turbine flowmeter: Fluid flow induces rotation of the turbine; measurement of turbine velocity provides an indication of flow rate. (c)

FIGURE 5.11.2 Devices for the measurement of flow.

pressure across the orifice, $p_1 - p_2$, and flow rate through the orifice, q, is predetermined through the calibration; therefore, measuring the differential pressure is equivalent to measuring flow rate.

The sensor in Figure 5.11.2(b) is called a **hot-wire anemometer**, because it is based on a heated wire that is cooled by the flow of a gas. The resistance of the wire changes with temperature, and a Wheatstone bridge circuit converts this change in resistance to a change in voltage. Also commonly used are **hot-film anemometers**, where a heated film is used in place of the more delicate wire. A very common application of the latter type of sensor is in automotive engines, where control of the air-to-fuel ratio depends on measurement of the engine intake mass airflow rate.

Figure 5.11.2(c) depicts a **turbine flowmeter**, in which the fluid flow causes a turbine to rotate; the velocity of rotation of the turbine (which can be measured by a noncontact sensor — e.g., a magnetic pickup) is related to the flow velocity.

Besides the techniques discussed in this chapter, many other techniques exist for measuring fluid flow, some of significant complexity.

Temperature Measurements

One of the most frequently measured physical quantities is temperature. The need to measure temperature arises in just about every field of engineering. This subsection is devoted to summarizing two common temperature sensors — the **thermocouple** and the **resistance temperature detector** (RTD) — and their related signal conditioning needs.

Thermocouples. A thermocouple is formed by the junction of two dissimilar metals. This junction results on an open-circuit **thermoelectric voltage** due to the **Seebeck effect**, named after Thomas Seebeck, who discovered the phenomenon in 1821. Various types of thermocouples exist; they are usually classified according to the data of Table 5.11.2. The Seebeck coefficient shown in the table is specified at a given temperature because the output voltage of a thermocouple, v, has a nonlinear dependence on temperature. This dependence is typically expressed in terms of a polynomial of the following form:

$$T = a_0 + a_1 v + a_2 v^2 + a_3 v^3 + \ldots + a_n v^n \tag{5.11.1}$$

For example, the coefficients of the J thermocouple in the range –100 to + 1000°C are as follows:

$$a_0 = -0.048868252 \qquad a_1 = 19{,}873.14503 \qquad a_2 = -128{,}614.5353$$

$$a_3 = 11{,}569{,}199.78 \qquad a_4 = -264{,}917{,}531.4 \qquad a_5 = 2{,}018{,}441{,}314$$

TABLE 5.11.2 Thermocouple Data

Type	Elements +/–	Seebeck Coefficient (μV/°C)	Range (°C)	Range (mV)
E	Chromel/constantan	58.70 at 0°C	–270 to 1000	–9.835 to 76.358
J	Iron/constantan	50.37 at 0°C	–210 to 1200	–8.096 to 69.536
K	Chromel/alumel	39.48 at 0°C	–270 to 1372	–6.548 to 54.874
R	Pt(10%)–Rh/Pt	10.19 at 600°C	–50 to 1768	–0.236 to 18.698
T	Copper/constantan	38.74 at 0°C	–270 to 400	–6.258 to 20.869
S	Pt(13%)–Rh/Pt	11.35 at 600°C	–50 to 1768	–0.226 to 21.108

The use of a thermocouple requires special connections, because the junction of the thermocouple wires with other leads (such as voltmeter leads, for example) creates additional thermoelectric junctions that in effect act as additional thermocouples. For example, in the J thermocouple circuit of Figure 5.11.3, junction J_1 is exposed to the temperature to be measured, but junctions J_2 and J_3 also generate a thermoelectric voltage, which is dependent on the temperature at these junctions, that is, the temperature at the voltmeter connections. One would therefore have to know the voltages at these junctions as well, in order to determine the actual thermoelectric voltage at J_1. To obviate this problem, a reference junction at known temperature can be employed; a traditional approach involves the use of a **cold junction**, so called because it consists of an ice bath, one of the easiest means of obtaining a known reference temperature. Figure 5.11.4 depicts a thermocouple measurement using an ice bath. The voltage measured in Figure 5.11.4 is dependent on the temperature difference $T_1 - T_{ref}$, where $T_{ref} = 0°C$. The connections to the voltmeter are made at an *isothermal block*, kept at a constant temperature; note that the same metal is used in both of the connections to the isothermal block. Thus (still assuming a J thermocouple), there is no difference between the thermoelectric voltages at the two copper-iron junctions; these will add to zero at the voltmeter. The voltmeter will therefore read a voltage proportional to $T_1 - T_{ref}$.

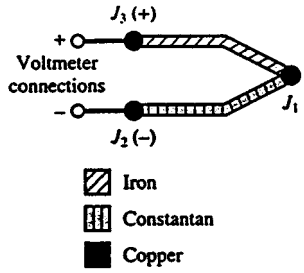


FIGURE 5.11.3 J thermocouple circuit.

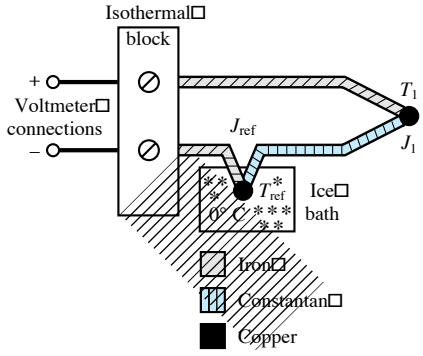


FIGURE 5.11.4 Cold-junction-compensated thermocouple circuit.

An ice bath is not always a practical solution. Other cold junction temperature compensation techniques employ an additional temperature sensor to determine the actual temperature of the junctions J_2 and J_3 of Figure 5.11.3

Resistance Temperature Detectors (RTDs). A resistance temperature detector (RTD) is a variable-resistance device whose resistance is a function of temperature. RTDs can be made with both positive and negative temperature coefficients and offer greater accuracy and stability than thermocouples. **Thermistors** are part of the RTD family. A characteristic of all RTDs is that they are *passive* devices, that is, they do not provide a useful output unless excited by an external source. The change in resistance in an RTD is usually converted to a change in voltage by forcing a current to flow through the device. An indirect result of this method is a **self-heating error**, caused by the $i^2 R$ heating of the device. Self-heating of an RTD is usually denoted by the amount of power that will raise the RTD temperature by 1°C. Reducing the excitation current can clearly help reduce self-heating, but it also reduces the output voltage.

The RTD resistance has a fairly linear dependence on temperature, a common definition of the **temperature coefficient** of an RTD is related to the change in resistance from 0 to 100°C. Let R_0 be the resistance of the device at 0°C and R_{100} the resistance at 100°C. Then the temperature coefficient, α, is defined to be

$$\alpha = \frac{R_{100} - R_0}{100 - 0} \frac{\Omega}{^\circ C} \tag{5.11.2}$$

A more accurate representation of RTD temperature dependence can be obtained by using a nonlinear (cubic) equation and published tables of coefficients. As an example, a platinum RTD could be described either by the temperature coefficient $\alpha = 0.003911$, or by the equation

$$\begin{aligned} R_T &= R_0\left(1 + AT - BT^2 - CT^3\right) \\ &= R_0\left(1 + 3.6962 \times 10^{-3} T - 5.8495 \times 10^{-7} T^2 - 4.2325 \times 10^{-12} T^3\right) \end{aligned} \tag{5.11.3}$$

where the coefficient C is equal to zero for temperatures above 0°C.

Because RTDs have fairly low resistance, they are sensitive to error introduced by the added resistance of the lead wires connected to them; Figure 5.11.5 depicts the effect of the lead resistances, r_L, on the RTD measurement. Note that the measured voltage includes the resistance of the RTD as well as the resistance of the leads. If the leads used are long (greater than 3 m is a good rule of thumb), then the measurement will have to be adjusted for this error. Two possible solutions to the lead problems are the *four-wire* RTD measurement circuit and the *three-wire* Wheatstone bridge circuit, shown in Figure 5.11.6(a) and (b), respectively. In the circuit of Figure 5.11.6(a), the resistance of the lead wires from the excitation, r_{L1} and r_{L4}, may be arbitrarily large, since the measurement is affected by the resistance of only the output lead wires, r_{L2} and r_{L3}, which can be kept small by making these leads short. The circuit of Figure 5.11.6(b) takes advantage of the properties of the Wheatstone bridge to cancel out the unwanted effect of the lead wires while still producing an output dependent on the change in temperature.

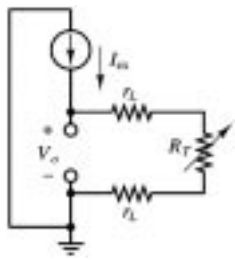


FIGURE 5.11.5 Effect of connection leads on RTD temperature measurement.

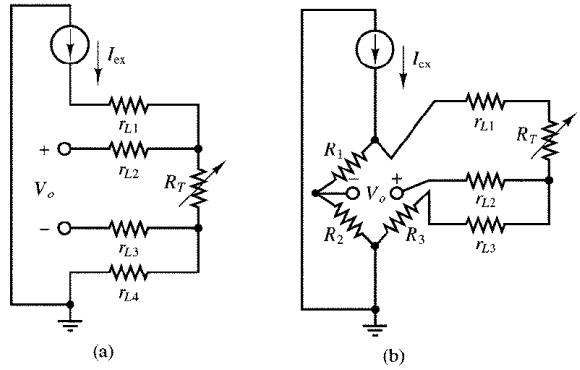


FIGURE 5.11.6 Four-wire RTD circuit (a) and three-wire Wheatstone bridge RTD circuit (b).

Wiring, Grounding, and Noise

The importance of proper circuit connections cannot be overemphasized. Unfortunately, this is a subject that is rarely taught in introductory electrical engineering courses. The present section summarizes some important considerations regarding signal source connections, various types of input configurations, noise sources and coupling mechanisms, and means of minimizing the influence of noise on a measurement.

Signal Sources and Measurement System Configurations

Before proper connection and wiring techniques can be presented, we must examine the difference between **grounded** and **floating signal sources**. Every sensor can be thought of as some kind of signal source; a general representation of the connection of a sensor to a measurement system is shown in Figure 5.11.7(a). The sensor is modeled as an ideal voltage source in series with a source resistance. Although this representation does not necessarily apply to all sensors, it will be adequate for the purposes of the present section. Figures 5.7.11(b) and (c) show two types of signal sources: grounded and floating. A grounded signal source is one in which a ground reference is established — for example, by connecting the *signal low* lead to a case or housing. A floating signal source is one in which neither signal lead is connected to ground; since ground potential is arbitrary, the signal source voltage levels (*signal low* and *signal high*) are at an unknown potential relative to the case ground. Thus, the signal is said to be *floating*. Whether a sensor can be characterized as a grounded or a floating signal source ultimately depends on the connection of the sensor to its case, but the choice of connection may depend on the nature of the source. For example, the thermocouple described earlier is *intrinsically* a floating signal source, since the signal of interest is a difference between voltages. The same thermocouple *could* become a grounded signal source if one or its two leads were directly connected to ground, but this is usually not a desirable arrangement for this particular sensor.

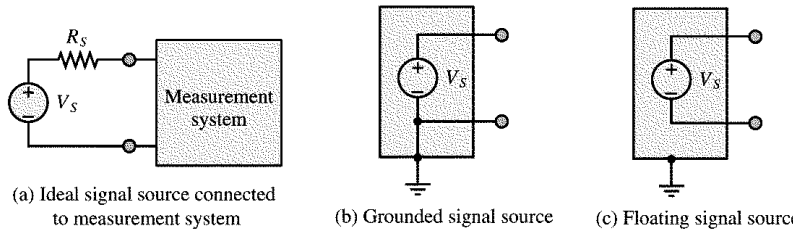


FIGURE 5.11.7 Measurement system and types of signal sources.

In analogy with a signal source, a measurement system can be either **ground-referenced** or **differential**. In a ground-referenced system, the signal low connection is tied to the instrument case ground; in a differential system, neither of the two signal connections is tied to ground. Thus, a differential measurement system is well suited to measuring the difference between two signal levels (such as the output of an ungrounded thermocouple).

One of the potential dangers in dealing with grounded signal sources is the introduction of **ground loops**. A ground loop is an undesired current path caused by the connection of two reference voltages to each other. This is illustrated in Figure 5.11.8, where a grounded signal source is shown connected to a ground-referenced measurement system. Notice that we have purposely denoted the signal source ground and the measurement system ground by two distinct symbols, to emphasize that these are not necessarily at the same potential — as also indicated by the voltage difference ΔV. Now, one might be tempted to tie the two grounds to each other, but this would only result in a current flowing from one ground to the other, through the small (but nonzero) resistance of the wire connecting the two. The net effect of this ground loop would be that the voltage measured by the instrument would include the unknown ground voltage difference ΔV, as shown in Figure 5.11.8. Since this latter voltage is unpredictable, you can see that ground loops can cause substantial errors in measuring systems. In addition, ground loops are the primary cause of conducted noise, as explained later in this section.

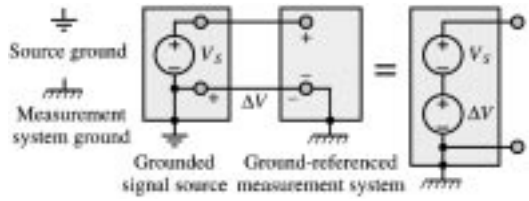


FIGURE 5.11.8 Ground loop in ground-referenced measurement system.

A differential measurement system is often a way to avoid ground loop problems because the signal source and measurement system grounds are not connected to each other, and especially because the signal low input of the measuring instrument is not connected to either instrument case ground. The connection of a grounded signal source and a differential measurement system is depicted in Figure 5.11.9.

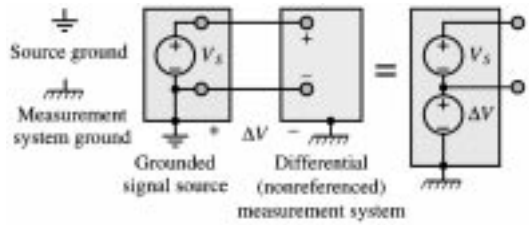


FIGURE 5.11.9 Differential (nonreferenced) measurement system.

If the signal source connected to the differential measurement system is floating, as shown in Figure 5.11.10, it is often a recommended procedure to reference the signal to the instrument ground by means of two identical resistors that can provide a return path to ground for any currents present at the instrument. An example of such input currents could be the input bias currents inevitably present at the input of an operational or instrumentation amplifier.

The simple concepts illustrated in the preceding paragraphs and figures can assist the user and designer of instrumentation systems in making the best possible wiring connections for a given measurement.

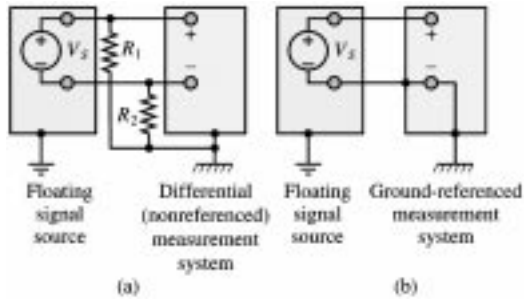


FIGURE 5.11.10 Measuring signals from a floating source: (a) differential input; (b) single-ended input.

Noise Sources and Coupling Mechanisms

Noise — meaning any undesirable signal interfering with a measurement — is an unavoidable element of all measurements. Figure 5.11.11 depicts a block diagram of the three essential stages of a noisy measurement: **a noise source, a noise coupling mechanism**, and a sensor or associated signal-conditioning circuit. Noise sources are always present and are often impossible to eliminate completely; typical sources of noise in practical measurements are the electromagnetic fields caused by fluorescent light fixtures, video monitors, power supplies, switching circuits, and high-voltage (or current) circuits. Many other sources exist, of course, but often the simple sources in our everyday environment are the most difficult to defeat.

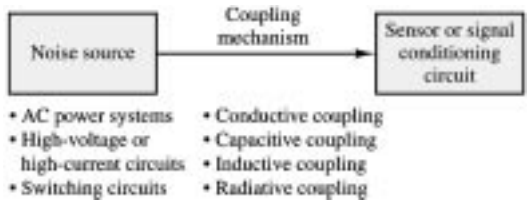


FIGURE 5.11.11 Noise sources and coupling mechanisms.

Figure 5.11.11 also indicates that various coupling mechanisms can exist between a noise source and an instrument. Noise coupling can be conductive; that is, noise currents may actually be conducted from the noise source to the instrument by physical wires. Noise can also be coupled capacitively, inductively and radiatively.

Figure 5.11.12 illustrates how interference can be **conductively coupled** by way of a ground loop. In the figure, a power supply is connected to both a load and a sensor. We shall assume that the load may be switched on and off, and that it carriers substantial currents. The top circuit contains a ground loop: the current i from the supply divides between the load and sensor; since the wire resistance is nonzero, a large current flowing through the load may cause the ground potential at point a to differ from the potential at point b. In this case, the measured sensor output is no longer, v_o, but it is now equal to $v_o + v_{ba}$, where v_{ba} is the potential difference from point b to point a. Now, if the load is switched on and off and its current is therefore subject to large, abrupt changes, these changes will be manifested in the voltage v_{ba} and will appear as noise on the sensor output.

This problem can be cured simply and effectively by providing separate *ground returns* for the load and sensor, thus eliminating the ground loop.

The mechanism of **capacitive coupling** is rooted in electric fields that may be caused by sources of interference. The detailed electromagnetic analysis can be quite complex, but to understand the principle, refer to Figure 5.11.13(a), where a noise source is shown to generate an electric field. If a noise source conductor is sufficiently close to a conductor that is part of the measurement system, the two conductors

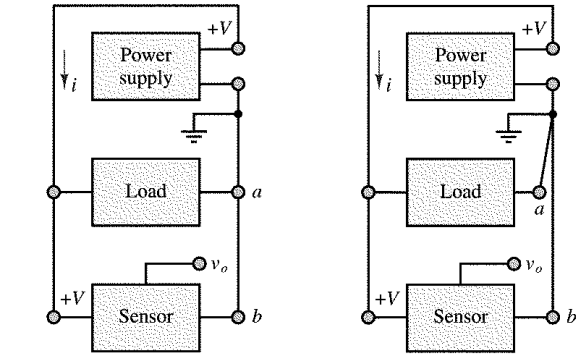


FIGURE 5.11.12 Conductive coupling: ground loop and separate ground returns.

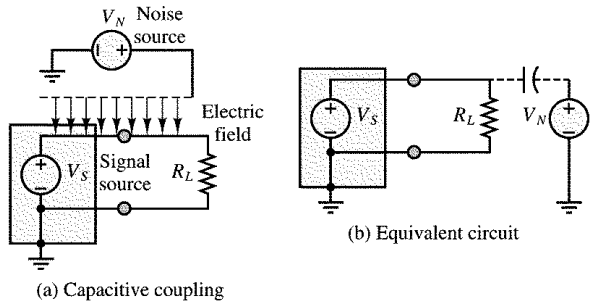


FIGURE 5.11.13 Capacitive coupling and equivalent-circuit representation.

(separated by air, a dielectric) will form a capacitor, through which any time-varying currents can flow. Figure 5.11.13(b) depicts an equivalent circuit in which the noise voltage V_N couples to the measurement circuit through an imaginary capacitor, representing the actual capacitance of the noise path.

The dual of capacitive coupling is **inductive coupling**. This form of noise coupling is due to the magnetic field generated by current flowing through a conductor. If the current is large, the magnetic fields can be significant, and the **mutual inductance** between the noise source and the measurement circuit causes the noise to couple to the measurement circuit. Thus, inductive coupling, as shown in Figure 5.11.14 results when undesired (unplanned) magnetic coupling ties the noise source to the measurement circuit.

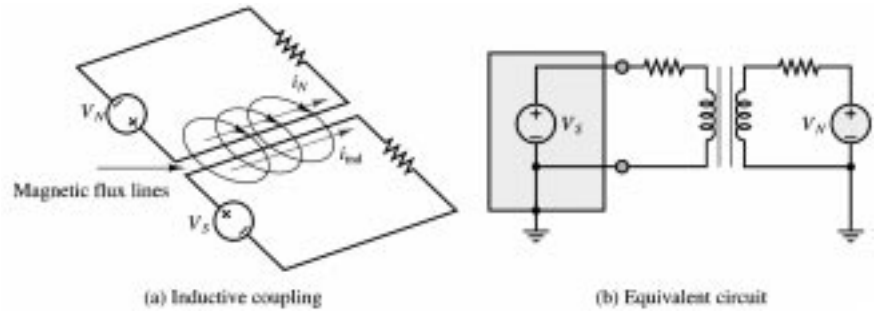


FIGURE 5.11.14 Inductive coupling and equalent-circuit representation.

Noise Reduction

Various techniques exist for minimizing the effect of undesired interference, in addition to proper wiring and grounding procedures. The two most common methods are **shielding** and the use of **twisted-pair wire**. A shielded cable is shown in Figure 5.11.15. The shield is made of a copper braid or of foil and is usually grounded at the source end, *but not at the instrument end*, because this would result in a ground loop. The shield can protect the signal from a significant amount of electromagnetic interference, especially at lower frequencies. Shielded cables with various numbers of conductors are available commercially. However, shielding cannot prevent inductive coupling. The simplest method for minimizing inductive coupling is the use of twisted-pair wire; the reason for using twisted pair is that untwisted wire can offer large loops that can couple a substantial amount of electromagnetic radiation. Twisting drastically reduces the loop area, and with it the interference. Twisted pair is available commercially.

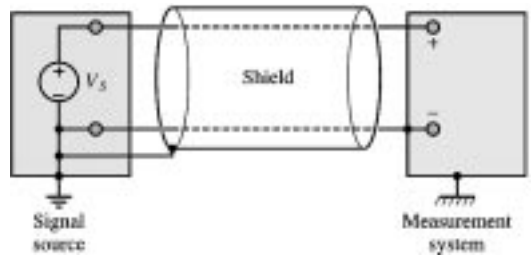


FIGURE 5.11.15 Shielding.

Signal Conditioning

A properly wired, grounded, and shielded sensor connection is a necessary first stage of any well-designed measurement system. The next stage consists of any **signal conditioning** that may be required to manipulate the sensor output into a form appropriate for the intended use. Very often, the sensor output is meant to be fed into a digital computer, as illustrated in Figure 5.11.1. In this case, it is important to condition the signal so that it is compatible with the process of data acquisition. Two of the most important signal-conditioning functions are *amplification* and *filtering*. Both are discussed in the present section.

Instrumentation Amplifiers

An **instrumentation amplifier** (IA) is a differential amplifier with very high input impedance, low bias current, and programmable gain that finds widespread application when low-level signals with large common-mode components are to be amplified in noisy environments. This situation occurs frequently when a low-level transducer signal needs to be preamplified, prior to further signal conditioning (e.g., filtering).

The functional structure of an IC instrumentation amplifier is depicted in Figure 5.11.16. Specifications for a common IC instrumentation amplifier (and a more accurate circuit description) are shown in Figure 5.11.17. Among the features worth mentioning here are the programmable gains, which the user can set by suitably connecting one or more of the resistors labeled R_1 to the appropriate connection. Note that the user may also choose to connect additional resistors to control the amplifier gain, without adversely affecting the amplifier's performance, since R_1 requires no matching. In addition to the pin connection that permits programmable gains, two additional pins are provided, called **sense** and **reference**. These additional connections are provided to the user for the purpose of referencing the output voltage to a signal other than ground, by means of the reference terminal, or of further amplifying the output current (e.g., with a transistor stage), by connecting the sense terminal to the output of the current amplifier.

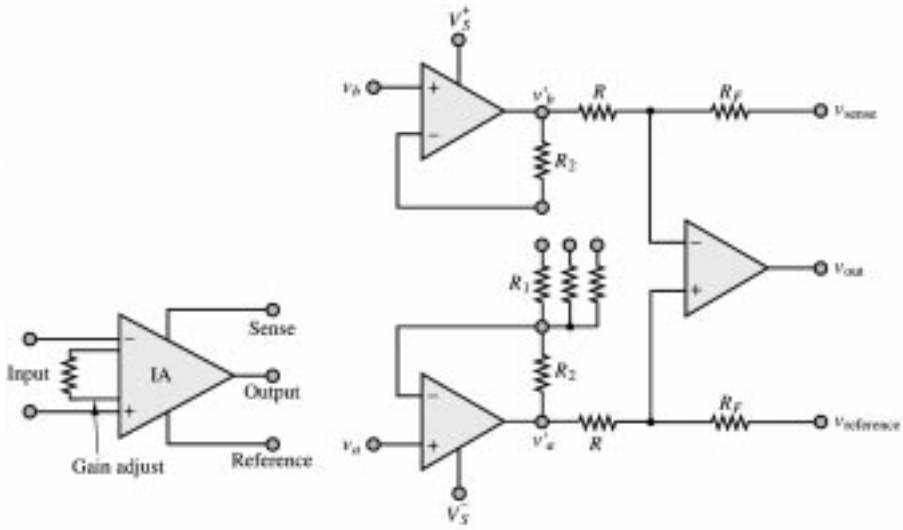


FIGURE 5.11.16 IC instrumentation amplifier.

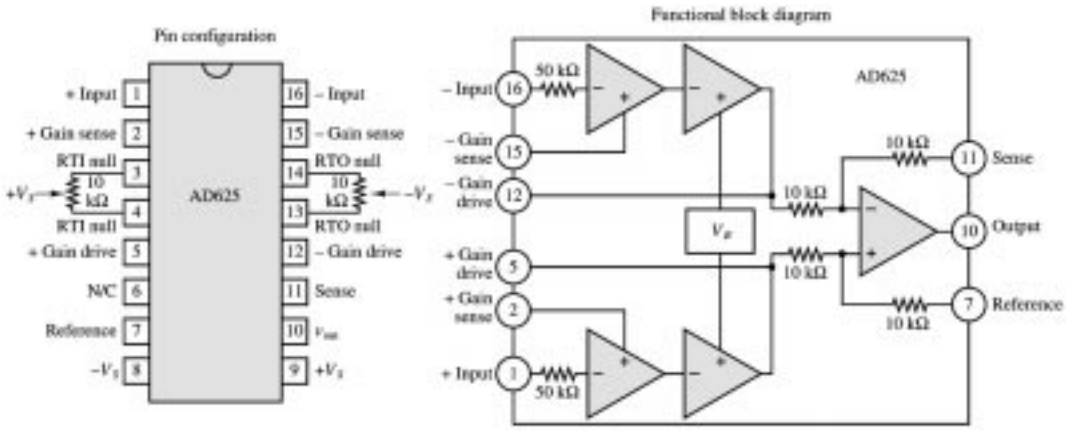


FIGURE 5.11.17 AD625 instrumentation amplifier data sheet.

Active Filters

The need to filter sensor signals that may be corrupted by noise or other interfering or undesired inputs has already been approached in two earlier chapters. In Section 5.6, simple passive filters made of resistors, capacitors, and inductors were analyzed. It was shown that three types of filter frequency response characteristics can be achieved with these simple circuits: low-pass, high-pass, and band-pass. In Section 5.9, the concept of active filters was introduced, to suggest that it may be desirable to exploit the properties of operational amplifiers to simplify filter design, to more easily match source and load impedances, and to eliminate the need for inductors. The aim of this section is to discuss more advanced active filter designs, which find widespread application in instrumentation circuits.

Figure 5.11.18 depicts the general characteristics of a low-pass active filter, indicating that within the pass-band of the filter, a certain deviation from the nominal filter gain, A, is accepted, as indicated by the minimum and maximum pass-band gains, $A + \varepsilon$ and $A - \varepsilon$. The width of the pass-band is indicated by the cutoff frequency, ω_c. On the other hand, the stop-band, starting at the frequency ω_s, does not

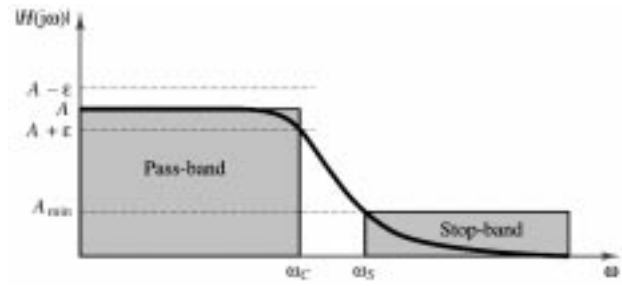


FIGURE 5.11.18 Prototype low-pass filter response.

allow a gain greater than A_{min}. Different types of filter designs achieve different types of frequency responses, which are typically characterized either by having a particularly flat pass-band frequency response (**Butterworth filters**) or by a very rapid transition between pass-band and stop-band (**Chebyshev filters**, and **Cauer**, or **elliptical, filters**), or by some other characteristic, such as a very linear phase response (**Bessel filters**). Achieving each of these properties usually involves trade-offs; for example, a very flat pass-band response will usually result in a relatively slow transition from pass-band to stop-band.

In addition to selecting a filter from a certain family, it is also possible to select the *order* of the filter; this is equal to the order of the differential equation that describes the input-output relationship of a given filter. In general, the higher the order, the faster the transition from pass-band to stop-band (at the cost of greater phase shifts and amplitude distortion, however). Although the frequency response of Figure 5.11.18 pertains to a low-pass filter, similar definitions also apply to the other types of filters.

Butterworth filters are characterized by a *maximally flat* pass-band frequency response characteristic; their response is defined by a magnitude-squared function of frequency:

$$\left|H(j\omega)\right|^2 = \frac{H_0^2}{1+\varepsilon^2\omega^{2n}} \tag{5.11.14}$$

where $\varepsilon = 1$ for maximally flat response and n is the order of the filter. Figure 5.11.19 depicts the frequency response (normalized to $\omega_c = 1$) of first-, second-, third-, and fourth-order Butterworth low-pass filters. The **Butterworth polynomials**, given in Table 5.11.3 in factored form, permit the design of the filter by specifying the denominator as a polynomial in s. For $s = j\omega$, one obtains the frequency response of the filter.

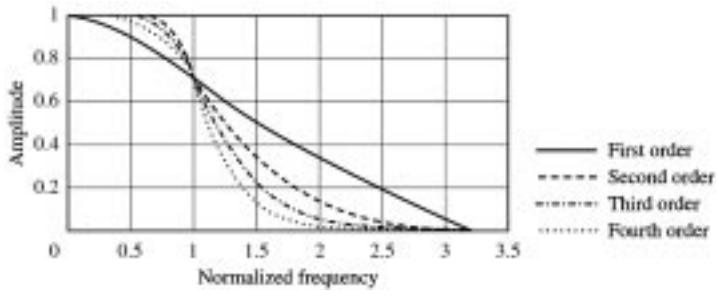


FIGURE 5.11.19 Butterworth low-pass filter frequency response.

Figure 5.11.20 depicts the normalized frequency response of first- to fourth-order low-pass Chebyshev filters ($n = 1$ to 4), for $\varepsilon = 1.06$. Note that a certain amount of ripple is allowed in the pass-band; the

TABLE 5.11.3 Butterworth Polynomials in Quadratic Form

Order, *n*	Quadratic Factors
1	$(s + 1)$
2	$(s^2 + \sqrt{2}s + 1)$
3	$(s + 1)\ (s^2 + s + 1)$
4	$(s^2 + 0.7654s + 1)\ (s^2 + 1.8478s + 1)$
5	$(s + 1)\ (s^2 + 0.6180s + 1)\ (s^2 + 1.6180s + 1)$

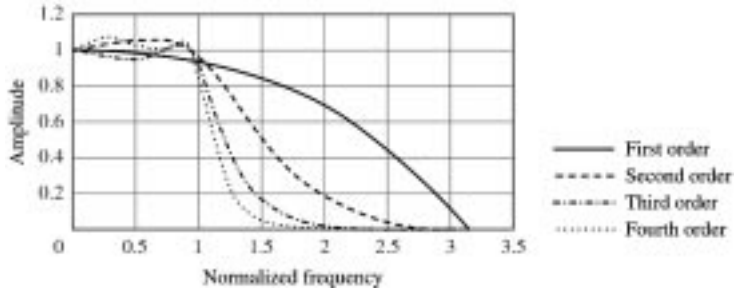


FIGURE 5.11.20 Chebyshev low-pass filter frequency response.

amplitude of the ripple is defined by the parameter ε and is constant throughout the pass-band. Thus, these filters are also called **equiripple filters**. Cauer, or elliptical, filters are similar to Chebyshev filters, except for being characterized by equiripple both in the pass-band and in the stop-band. Design tables exist to select the appropriate order of Butterworth, Chebyshev, or Cauer filter for a specific application.

Three common configurations of second-order active filters, which can be used to implement **second-order** (or **quadratic**) **filter sections** using a single of op-amp, are shown in Figure 5.11.21. These filters are called **constant-*K***, or **Sallen and Key, filters** (after the name of the inventors). The analysis of these active filters, although somewhat more involved than that of the active filters presented in the preceding chapter, is based on the basic properties of the ideal operational amplifier discussed earlier. Consider, for example, the low-pass filter of Figure 5.11.21. The first unusual aspect of the filter is the presence of both negative and **positive feedback**; that is, feedback connections are provided to both the inverting and the noninverting terminals of the op-amp. The analysis method consists of finding expressions for the input terminal voltages of the op-amp, V^+ and V^-, and using these expressions to derive the input-output relationship for the filter. The frequency response of the low-pass filter is given by

$$H(j\omega) = \frac{K\left(1/R_1R_2C_1C_2\right)}{(j\omega)^2 + \left[\frac{1}{R_1C_1} + \frac{1}{R_2C_1} + \frac{1}{R_2C_2}(1-K)\right] + \frac{1}{R_1R_2C_1C_2}} \tag{5.11.5}$$

This frequency response can be expressed in more general form as follows:

$$H(j\omega) = \frac{H_0\omega_C^2}{(j\omega)^2 + (\omega_C/Q)j\omega + \omega_C^2}$$

where

$$H_0 = K \tag{5.11.6}$$

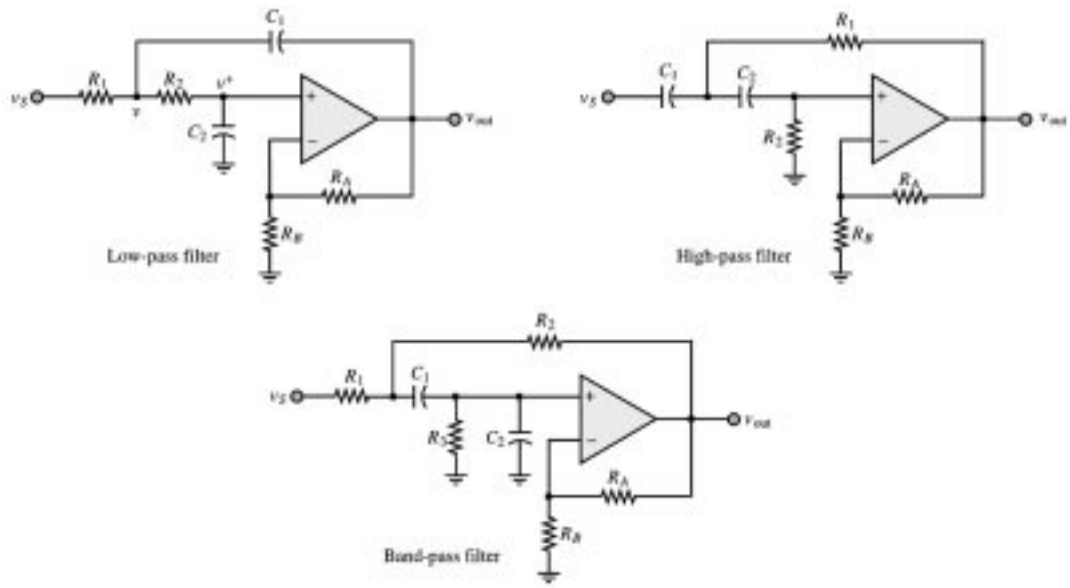


FIGURE 5.11.21 Sallen and Key active filters.

is the DC gain of the filter, and where

$$\omega_C = \sqrt{\frac{1}{R_1 R_2 C_1 C_2}}$$

is the cutoff frequency, and where

$$\frac{1}{Q} = \sqrt{\frac{R_2 C_1 C_2}{R_1 C_1}} + \sqrt{\frac{R_1 C_2}{R_2 C_1}} + (1 - K)\sqrt{\frac{R_1 C_1}{R_2 C_2}} \tag{5.11.17}$$

is the inverse of the **quality factor**, Q. The Q of a filter is related to the overshoot in the transient response of the filter; and to the peaking (i.e., sharpness of the resonant peak) of the frequency response, a high-Q circuit will display more peaking, or overshoot, than a low-Q circuit.

Analog-to-Digital and Digital-to-Analog Conversion

To take advantage of the capabilities of a microcomputer, it is necessary to suitably interface signals to and from external devices with the microcomputer. Such signals may be analog or digital. Depending on the nature of the signal, either an analog or a digital interface circuit will be required. The advantages in memory storage, programming flexibility, and computational power afforded by today's digital computers are such that the instrumentation designer often chooses to convert an analog signal to an equivalent digital representation, to exploit the capabilities of a microprocessor in processing the signal. In many cases, the data converted from analog to digital form remain in digital form for ease of storage, or for further processing. In some instances it is necessary to convert the data back to analog form. The latter condition arises frequently in the context of control system design, where an analog measurement is converted to digital form and processed by a digital computer to generate a control action (e.g., raising or lowering the temperature of a process, or exerting a force or a torque); in such cases, the output of the digital computer is converted back to analog form, so that a continuous signal becomes available to

the actuators. Figure 5.11.22 illustrates the general appearance of a digital measuring instrument and of a digital controller acting on a plant or process.

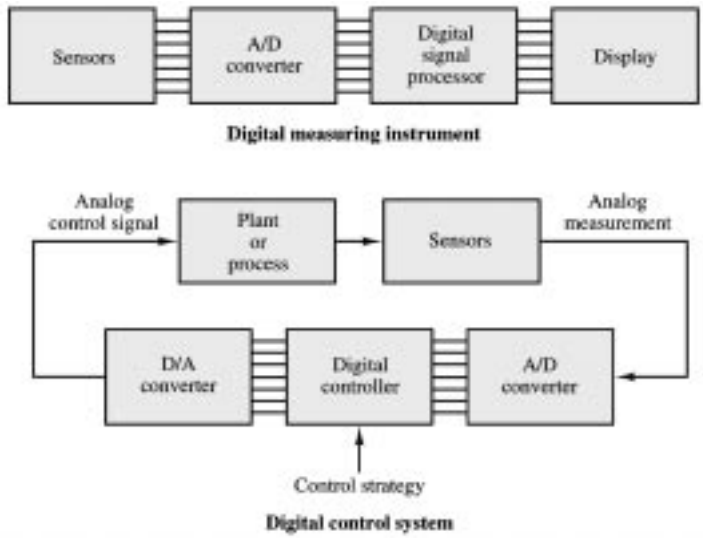


FIGURE 5.11.22 Block diagrams of a digital measuring instrument and a digital control system.

The objective of this section is to describe how the digital-to-analog (D/A) and analog-to-digital (A/D) conversion blocks of Figure 5.11.22 function. After illustrating discrete circuits that can implement simple A/D and D/A converters, we shall emphasize the use of ICs specially made for these tasks. Nowadays, it is uncommon (and impractical) to design such circuits using discrete components: the performance and ease of use of IC packages make them the preferred choice in virtually all applications.

Digital-to-Analog Converters

We discuss digital-to-analog conversion first because it is a necessary part of analog-to-digital conversion in many A/D conversion schemes. A **digital-to-analog converter** (DAC) will convert a binary word to an analog output voltage (or current). The binary word is represented in terms of 1s and 0s, where typically (but not necessarily) 1s correspond to a 5-V level and 0s to a 0-V signal. As an example, consider a four-bit binary word:

$$B = \left(b_3 b_2 b_1 b_0\right)_2 = \left(b_3 \cdot 2^3 + b_2 \cdot 2^2 + b_1 \cdot 2^1 + b_0 \cdot 2^0\right)_{10} \tag{5.11.8}$$

The analog voltage corresponding to the digital word B would be

$$v_a = \left(8b_3 + 4b_2 + 2b_1 + b_0\right)\delta v \tag{5.11.9}$$

where δv is the smallest *step size* by which v_a can increment. This least step size will occur whenever the least significant bit (LSB), b_0, changes from 0 to 1, and is the smallest increment the digital number can make. We shall also shortly see that the analog voltage obtained by the D/A conversion process has a "staircase" appearance because of the discrete nature of the binary signal.

The step size is determined on the basis of each given application and is usually determined on the basis of the number of bits in the digital word to be converted to an analog voltage. We can see that, by extending the previous example for an n-bit word, the maximum value v_a can attain is

$$v_{a\max} = \left(2^{n-1} + 2^{n-2} + \ldots + 2^1 + 2^0\right)\delta v$$
$$= \left(2^n - 1\right)\delta v \tag{5.11.10}$$

It is relatively simple to construct a DAC by taking advantage of the summing amplifier illustrated in Section 5.9. Consider the circuit shown in Figure 5.11.23, where each bit in the word to be converted is represented by means of a 5-V source and a switch. When the switch is closed, the bit takes a value of 1 (5 V); when the switch is not closed, the bit has value 0. Thus, the output of the DAC is proportional to the word $b_{n-1}\ldots b_{n-2}\ldots b_1 b_0$.

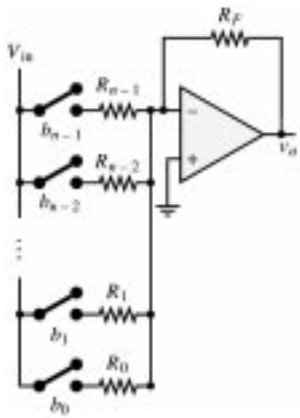


FIGURE 5.11.23 *n*-bit digital-to-analog converter (DAC).

If we select

$$R_i = \frac{R_0}{2^i} \tag{5.11.11}$$

we can obtain weighted gains for each bit so that

$$v_a = -\frac{R_F}{R_0}\left(2^{n-1}b_{n-1} + \ldots + 2^1 b_1 + 2^0 b_0\right)\cdot 5\ \text{V} \tag{5.11.12}$$

and so that the analog output voltage is proportional to the decimal representation of the binary word.

The practical design of a DAC is generally not carried out in terms of discrete components, because of problems such as the accuracy required of the resistor value. Many of the problems associated with this approach can be solved by designing the complete DAC circuit in integrated circuit (IC) form. The specifications stated by the IC manufacturer include the resolution, that is, the minimum nonzero voltage; the full-scale accuracy; the output range; the output settling time; the power supply requirements; and the power dissipation. The following example illustrates the use of a common integrated circuit DAC.

Example 5.11.1

A typical DAC one would use in conjunction with the 8086 microprocessor is the AD558. This is an IC that can be used in a "stand-alone" configuration (without a microprocessor) or with a microprocessor interfaced to it.

1. If one were to set up the AD558 for an output swing of 0 to 10 V, what would be the smallest voltage output increment attainable?

2. On what is the maximum operating frequency (the largest frequency on which the DAC can perform conversion) of the AD558 dependent? Determine the maximum frequency attainable if the converter is to be run at full-scale input.

Solution.

1. Since this DAC is an eight-bit device, the total number of digital increments one could expect is 256. Thus, the finest voltage steps one could expect at the output would be

$$\frac{10}{255} = 39.2 \text{ mV}$$

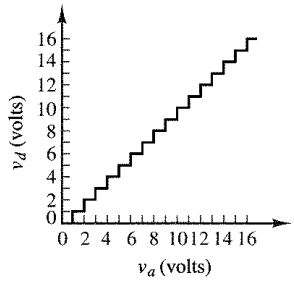
   This means that for every increment of one bit, the output would jump (in a stepwise fashion) by 39.2 mV.
2. The maximum frequency at which a DAC can run is dependent on the settling time. This is defined as the time it takes for the output to settle to within one half of the least significant bit of its final value. Thus, only one transition can be made per settling time. The settling time for the AD558 depends on the voltage range and is defined for a positive-going full-scale step to ±1/2 LSB. The settling time is 1 μsec, and the corresponding maximum conversion frequency is 1 MHz.

Analog-to-Digital Converters

You should by now have developed an appreciation for the reasons why it is convenient to process data in digital form. The device that makes conversion of analog signals to digital form is the **analog-to-digital converter** (ADC), and, just like the DAC, it is also available as a single IC package. In addition to discussing analog-to-digital conversion, we shall also introduce the *sample-and-hold amplifier.*

Quantization. The process of converting an analog voltage (or current) to digital form requires that the analog signal be quantized and encoded in binary form. The process of **quantization** consists of subdividing the range of the signal into a finite number of intervals; usually, one employs $2^n - 1$ intervals, where n is the number of bits available for the corresponding binary word. Following this quantization, a binary word is assigned to each interval (i.e., to each range of voltages or currents); the binary word is then the digital representation of any voltage (current) that falls within that interval. You will note that the smaller the interval, the more accurate the digital representation is. However, some error is necessarily always present in the conversion process; this error is usually referred to as **quantization error**. Let v_a represent the analog voltage and v_d its quantized counterpart, as shown in Figure 5.11.24 for an analog voltage in the range 0 to 0 16 V. In the figure, the analog voltage v_a takes on a value of $v_d = 0$ whenever it is in the range 0 to 1 V; for $1 \le v_a < 2$, the corresponding value is $v_d = 1$; for $2 \le v_a < 3$, $v_d = 2$; and so on, until, for $15 \le v_a < 16$, we have $v_d = 15$. You see that if we now represent the quantized voltage v_d by its binary counterpart, as shown in the table of Figure 5.11.24, each 1-V analog interval corresponds to a unique binary word. In this example, a four-bit word is sufficient to represent the analog voltage, although the representation is not very accurate. As the number of bits increases, the quantized voltage is closer and closer to the original analog signal; however, the number of bits required to represent the quantized value increases.

Tracking ADC. Although not the most efficient in all applications, the **tracking ADC** is an easy starting point to illustrate the operation of an ADC, in that it is based on the DAC presented in the previous section. The tracking ADC, shown in Figure 5.11.25, compares the analog input signal with the output of a DAC; the comparator output determines whether the DAC output is larger or smaller than the analog input to be converted to binary form. If the DAC output is smaller, then the comparator output will cause an up-down counter to count up, until it reaches a level close to the analog signal; if the DAC output is larger than the analog signal, then the counter is forced to count down. Note that the rate at which the up-down counter is incremented is determined by the external clock, and that the binary counter output



Quantized voltage / Binary representation

v_d	b_3	b_2	b_1	b_0
0	0	0	0	0
1	0	0	0	1
2	0	0	1	0
3	0	0	1	1
4	0	1	0	0
⋮			⋮	
14	1	1	1	0
15	1	1	1	1

FIGURE 5.11.24 A digital voltage representation of an analog voltage.

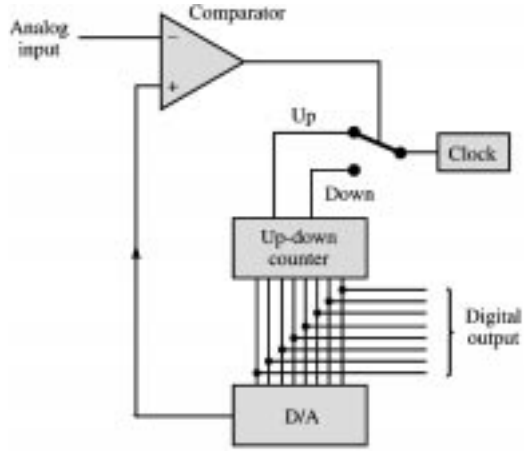


FIGURE 5.11.25 Tracking ADC.

corresponds to the binary representation of the analog signal. A feature of the tracking ADC is that it follows ("tracks") the analog signal by changing one bit at a time.

Integrating ADC. The **integrating ADC** operates by charging and discharging a capacitor, according to the following principle: if one can ensure that the capacitor charges (discharges) linearly, then the time it will take for the capacitor to discharge is linearly related to the amplitude of the voltage that has charged the capacitor. In practice, to limit the time it takes to perform a conversion, the capacitor is not required to charge fully. Rather, a clock is used to allow the input (analog) voltage to charge the capacitor for a short period of time, determined by a fixed number of clock pulses. Then the capacitor is allowed to discharge through a known circuit, and the corresponding clock count is incremented until the capacitor is fully discharged. The latter condition is verified by a comparator, as shown in Figure 5.11.26. The clock count accumulated during the discharge time is proportional to the analog voltage.

In the figure, the switch causes the counter to reset when it is connected to the reference voltage, V_{ref}. The reference voltage is used to provide a known, linear discharge characteristic through the capacitor (see the material on the op-amp integrator in Section 5.9. When the comparator detects that the output of the integrator is equal to zero, it switches state and disables the NAND gate, thus stopping the count. The binary counter output is now the digital counterpart of the voltage v_a.

Other common types of ADC are the so-called successive-approximation ADC and the flash ADC.

Flash ADC. The **flash ADC** is fully parallel and is used for high-speed conversion. A resistive divider network of $2''$ resistors divides the known voltage range into that many equal increments. A network of

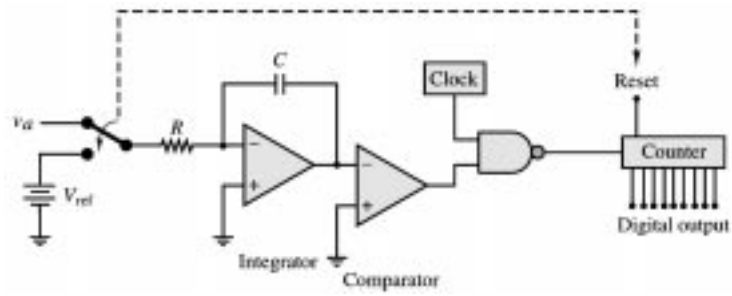


FIGURE 5.11.26 Integrating ADC.

$2'' - 1$ comparators then compares the unknown voltage with that array of test voltages. All comparators with inputs exceeding the unknown are "on"; all others are "off". This comparator code can be converted to conventional binary by a digital priority encoder circuit. For example, assume that the three-bit flash ADC of Figure 5.11.27 is set up with V_{ref} = 8 V. An input of 6.2 V is provided. If we number the comparators from the top of Figure 5.11.27, the state of each of the seven comparators is as given in Table 5.11.4.

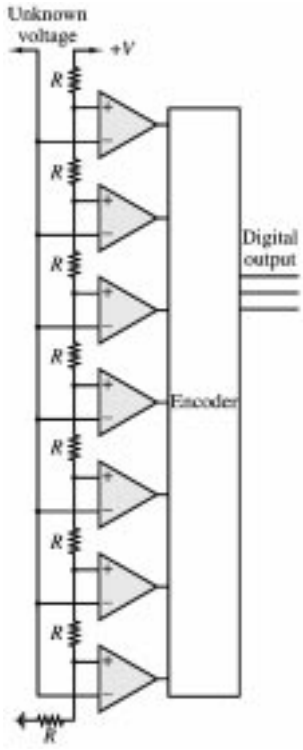


FIGURE 5.11.27 A three-bit flash ADC.

To resolve the uncertainty generated by the finite ADC conversion time of any practical converter, it is necessary to use a sample-and-hold amplifier. The objective of such an amplifier is to "freeze" the value of the analog waveform for a time sufficient for the ADC to complete its task.

A typical sample-and-hold amplifier is shown in Figure 5.11.28. It operates as follows. A MOSFET analog switch is used to "sample" the analog waveform. Recall that when a voltage pulse is provided to the sample input of the MOSFET switch (the gate), the MOSFET enters the ohmic region and in effect becomes nearly a short circuit for the duration of the sampling pulse. While the MOSFET conducts,

TABLE 5.11.4 State of Comparators in a 3-Bit Flash ADC

Comparator	Input on + Line	Input on – Line	Output
1	7 V	6.2 V	H
2	6 V	6.2 V	L
3	5 V	6.2 V	L
4	4 V	6.2 V	L
5	3 V	6.2 V	L
6	2 V	6.2 V	L
7	1 V	6.2 V	L

the analog voltage, v_a, charges the "hold" capacitor, C, at a fast rate through the small "on" resistance of the MOSFET. The duration of the sampling pulse is sufficient to charge C to the voltage v_a. Because the MOSFET is virtually a short circuit for the duration of the sampling pulse, the charging (RC) time constant is very small, and the capacitor charges very quickly. When the sampling pulse is over, the MOSFET returns to its nonconducting state, and the capacitor holds the sampled voltage without discharging, thanks to the extremely high input impedance of the voltage-follower (buffer) stage. Thus, v_{SH} is the sampled-and-held value of v_a at any given sampling time.

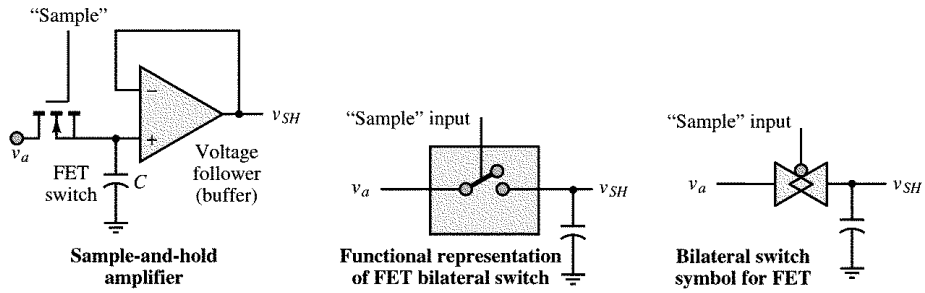


FIGURE 5.11.28 Description of the sample-and-hold process.

The appearance of the output of a typical sample-and-hold circuit is shown in Figure 5.11.29, together with the analog signal to be sampled. The time interval between samples, or **sampling interval**, $t_n - t_{n-1}$, allows the ADC to perform the conversion and make the digital version of the sampled signal available, say, to a computer or to another data acquisition and storage system. The sampling interval needs to be at least as long as the A/D conversion time, of course, but it is reasonable to ask how frequently one needs to sample a signal to preserve its fundamental properties (e.g., peaks and valleys, "ringing", fast transition). One might instinctively be tempted to respond that it is best to sample as frequently as possible, within the limitations of the ADC, so as to capture all the features of the analog signal. In fact, this is not necessarily the best strategy. How should we select the appropriate sampling frequency for a given application? Fortunately, an entire body of knowledge exists with regard to sampling theory, which enables the practicing engineer to select the best sampling rate for any given application. Given the scope of this chapter, we have chosen not to delve into the details of sampling theory, but, rather, to provide the student with a statement of the fundamental result: the **Nyquist sampling criterion**. The Nyquist criterion states that to prevent aliasing** when sampling a signal, *the sample rate should be selected to be at least twice the highest-frequency component present in the signal*. Thus, if we were sampling an audio signal (say, music), we would have to sample at a frequency of at least 40 kHz (twice the highest audible frequency, 20 kHz). In practice, it is advisable to select sampling frequencies substantially greater than the Nyquist rate; a good rule of thumb is five to ten times greater. The following

** Aliasing is a form of signal distortion that occurs when an analog signal is sampled at an insufficient rate.

example illustrates how the designer might take the Nyquist criterion into account in designing a practical A/D conversion circuit.

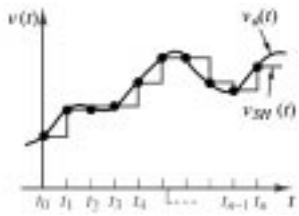


FIGURE 5.11.29 Sampled data.

Example 5.11.2

A typical ADC one would use in conjunction with the 8086 microprocessor is the AD574. This is a successive-approximation converter.

1. What is the accuracy (in volts) the AD574 can provide if $V_{CC} =$ 15.0 V and $0 \le \text{Vin} \le 15.0$ V?
2. On the basis of the data sheet, what is the highest-frequency signal you could convert using the AD574? (Assume that $V_{CC} =$ 15.0 V.)
3. If the maximum conversion time available were 40 msec, what would be the highest-frequency signal you could expect to sample on the basis of the Nyquist criterion?

Solution.

1. According to the data sheet, the least significant bit (LSB) of this converter limits its accuracy, meaning that the output is accurate within ± 1 bit. For the 0- to 15-V swing, this gives a voltage accuracy of

$$\frac{V_{\max} - V_{\min}}{2^n - 1} \quad \text{or} \quad \frac{15}{2^{12} - 1} \times (\pm 1 \text{ bit}) = \pm 3.66 \text{ mV}$$

2. On the basis of the data sheet, the maximum conversion time is 35 μsec. Therefore, the highest frequency of data conversion using the AD574 is

$$f_{\max} = \frac{1}{35\ \mu\text{s}} = 28.57 \text{ kHz}$$

   Thus, the highest signal frequency that could be represented, according to the Nyquist principle, is

$$\frac{1}{2} f_{\max} = \frac{28.57 \times 10^3}{2} = 14.285 \text{ kHz}$$

   This is the maximum theoretical signal frequency that can be represented without distortion, according to the Nyquist principle.
3. Following the same procedure discussed in part 2,

$$\frac{1}{2} f_{\max} = \left(\frac{1}{40 \times 10^{-6}}\right)\left(\frac{1}{2}\right) = 12.5 \text{ kHz}$$

Data Acquisition Systems

A typical data acquisition system, shown in Figure 5.11.30, often employs an *analog multiplexer*, to process several different input signals. A bank of bilateral analog MOSFET switches provides a simple

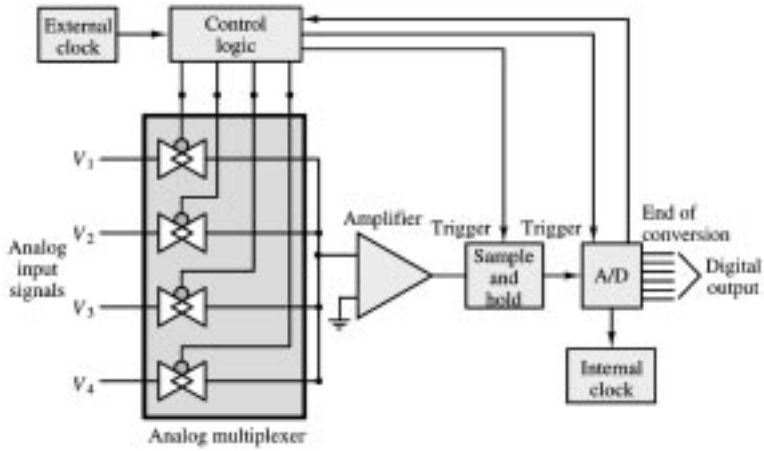


FIGURE 5.11.30 Data acquisition system.

and effective means of selecting which of the input signals should be sampled and converted to digital form. Control logic, employing standard gates and counters, is used to select the desired *channel* (input-signal) and to trigger the sampling circuit and the ADC. When the A/D conversion is completed, the ADC sends an appropriate *end of conversion* signal to the control logic, thereby enabling the next channel to be sampled.

In the block diagram of Figure 5.11.30, four analog inputs are shown; if these were to be sampled at regular intervals, the sequence of events would appear as depicted in Figure 5.11.31. We notice, from a qualitative analysis of the figure, that the effective sampling rate for each channel is one fourth the actual external clock rate; thus, it is important to ensure that the sampling rate for each individual channel satisfies the Nyquist criterion. Further, although each sample is held for four consecutive cycles of the external clock, we must notice that the ADC can use only one cycle of the external clock to complete the conversion, since its services will be required by the next channel during the next clock cycle. Thus, the internal clock that times the ADC must be sufficiently fast to allow for a complete conversion of any sample within the design range.

Timer ICs: the NE555

This section introduces a multipurpose integrated circuit that can perform basic timing functions. The NE555 is a timer circuit capable of producing accurate time delays (pulses) or oscillation. In the time-delay, or monostable, mode, the time delay or pulse duration is controlled by an external RC network. In the astable, or clock generator, mode, the frequency is controlled by two external resistors and one capacitor. Figure 5.11.32 depicts typical circuits for monostable and astable operation of the NE555. Note that the threshold level and the trigger level can also be externally controlled. For the monostable circuit, the pulse width can be computed from the following equation:

$$T = 1.1R_1C \tag{5.11.13}$$

For the astable circuit, the positive pulse width can be computed from the following equation:

$$T_+ = 0.69\left(R_1 + R_2\right)C \tag{5.11.14}$$

and the negative pulse width can be computed from

$$T_- = 0.69R_2C \tag{5.11.15}$$

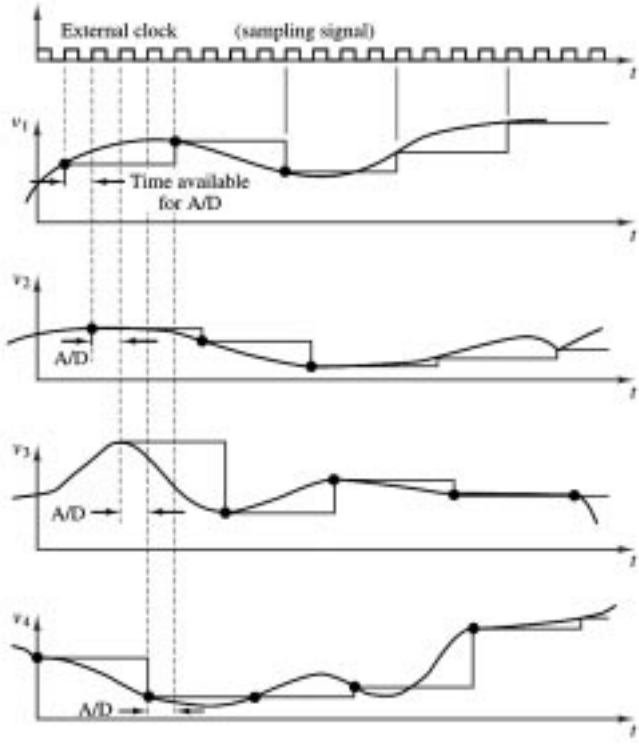


FIGURE 5.11.31 Multiplexed sampled data.

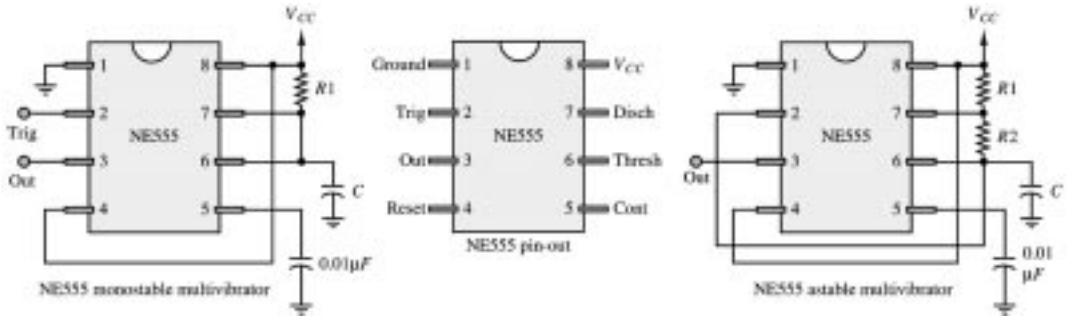


FIGURE 5.11.32 NE555 timer.

Data Transmission in Digital Instruments

One of the necessary aspects of data acquisition and control systems is the ability to transmit and receive data. Often, a microcomputer-based data acquisition system is interfaced to other digital devices, such as digital instruments or other microcomputers. In these cases it is necessary to transfer data directly in digital form. In fact, it is usually preferable to transmit data that are already in digital form, rather than analog voltages or currents. Among the chief reasons for the choice of digital over analog is that digital data are less sensitive to noise and interference than analog signals: in receiving a binary signal transmitted over a data line, the only decision to be made is whether the value of a bit is 0 or 1. Compared with the difficulty in obtaining a precise measurement of an analog voltage or current, either of which could be corrupted by noise or interference in a variety of ways, the probability of making an error in discerning between binary 0s and 1s is very small. Further, as will be shown shortly, digital data are often coded in such a way that many transmission errors may be detected and corrected for. Finally,

storage and processing of digital data are much more readily accomplished than would be the case with analog signals. This section explores a few of the methods that are commonly employed in transmitting digital data; both parallel and serial interfaces are considered.

Digital signals in a microcomputer are carried by a bus, consisting of a set of parallel wires each carrying one bit of information. In addition to the signal-carrying wires, there are also control lines that determine under what conditions transmission may occur. A typical computer data bus consists of eight parallel wires and therefore enables the transmission of one byte; digital data are encoded in binary according to one of a few standard codes, such as the BCD code described in Section 5.10, or the ASCII code, which is summarized in Table 5.11.5. This bus configuration is usually associated with **parallel transmission**, whereby all of the bits are transmitted simultaneously, along with some control bits.

TABLE 5.11.5 ASCII Code

Graphic or Control	ASCII (hex)	Graphic or Control	ASCII (hex)	Graphic or Control	ASCII (hex)
NUL	00	+	2B	V	56
SOH	01	,	2C	W	57
STX	02	-	2D	X	58
ETX	03	.	2E	Y	59
EOT	04	/	2F	Z	5A
ENQ	05	0	30	[	5B
ACK	06	1	31	\	5C
BEL	07	2	32	]	5D
BS	08	3	33	↑	5E
HT	09	4	34	←	5F
LF	0A	5	35	`	60
VT	0B	6	36	a	61
FF	0C	7	37	b	62
CR	0D	8	38	c	63
SO	0E	9	39	d	64
SI	0F	:	3A	e	65
DLE	10	;	3B	f	66
DCl	11	<	3C	g	67
DC2	12	=	3D	h	68
DC3	13	>	3E	i	69
DC4	14	?	3F	j	6A
NAK	15	@	40	k	6B
SYN	16	A	41	l	6C
ETB	17	B	42	m	6D
CAN	18	C	43	n	6E
EM	19	D	44	o	6F
SUB	1A	E	45	p	70
ESC	1B	F	46	q	71
FS	1C	G	47	r	72
GS	1D	H	48	s	73
RS	1E	I	49	t	74
US	1F	J	4A	u	75
SP	20	K	4B	v	76
!	21	L	4C	w	77
”	22	M	4D	x	78
#	23	N	4E	y	79
$	24	O	4F	z	7A
%	25	P	50	{	7B
&	26	Q	51	l	7C
’	27	R	52	}	7D
(	28	S	53	~	7E
)	29	T	54	DEL	7F
*	2A	U	55		

Figure 5.11.33 depicts the general appearance of a parallel connection. Parallel data transmission can take place in one of two modes: **synchronous** or **asynchronous**. In synchronous transmission, a timing clock pulse is transmitted along with the data over a control line. The arrival of the clock pulse indicates that valid data have also arrived. While parallel synchronous transmission can be very fast, it requires the added complexity of a synchronizing clock and is typically employed only for internal computer data transmission. Further, this type of communication can take place only over short distances (approximately 4 m). Asynchronous data transmission, on the other hand, does not take place at a fixed clock rate, but requires a **handshake protocol** between sending and receiving ends. The handshake protocol consists of the transmission of *data ready* and *acknowledge* signals over two separate control wires. Whenever the sending device is ready to transmit data, it sends a pulse over the *data ready* line. When this signal reaches the receiver, and if the receiver is ready to receive the data, an *acknowledge* pulse is sent back, indicating that the transmission may occur; at this point, the parallel data are transmitted.

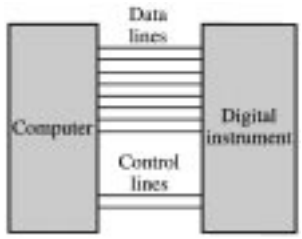


FIGURE 5.11.33 Parallel data transmission.

Perhaps the most common parallel interface is based on the **IEEE 488 standard**, leading to the so-called IEEE 488 bus, also referred to as GPIB (for **general-purpose instrument bus**).

The IEEE 488 Bus

The IEEE 488 bus, shown in Figure 5.11.34, is an eight-bit parallel asynchronous interface that has found common application in digital instrumentation applications. The physical bus consists of 16 lines, of which 8 are used to carry the data, 3 for the handshaking protocol, and the rest to control the data flow. The bus permits connection of up to 15 instruments and data rates of up to 1 Mbyte/sec. There is a limitation, however, in the maximum total length of the bus cable, which is 20 m. The signals transmitted are TTL-compatible and employ negative logic, whereby a logic 0 corresponds to a TTL high state (>2 V) and a logic 1 to a TTL low state (<0.8 V). Often, the eight-bit word transmitted over an IEEE 488 bus is coded in ASCII format (see Table 5.11.5).

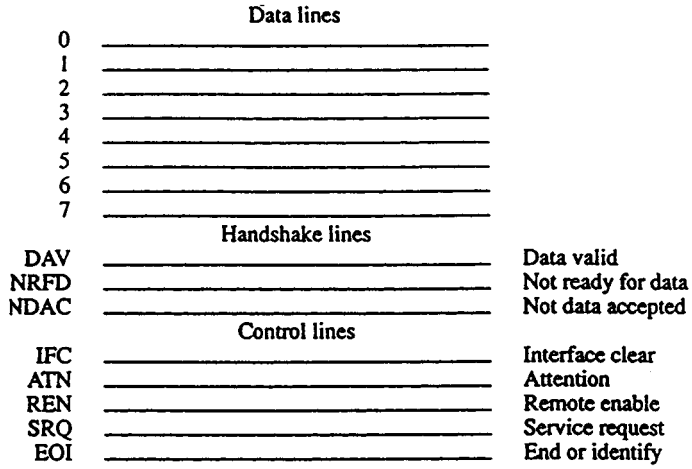


FIGURE 5.11.34 IEEE 488 bus.

In an IEEE 488 bus system, devices may play different roles and are typically classified as *controllers*, which manage the data flow; *talkers* (e.g., a digital voltmeter), which can only send data; *listeners* (e.g., a printer), which can only receive data; and *talkers/listeners* (e.g., a digital oscilloscope), which can receive as well as transmit data. The simplest system configuration might consist of just a talker and a listener. If more than two devices are present on the bus, a controller is necessary to determine when and how data transmission is to occur on the bus. For example, one of the key rules implemented by the controller is that only one talker can transmit at any one time; it is possible, however, for several listeners to be active on the bus simultaneously. If the data rates of the different listeners are different, the talker will have to transmit at the slowest rate, so that all of the listeners are assured of receiving the data correctly.

The set of rules by which the controller determines the order in which talking and listening are to take place is determined by a **protocol**. One aspect of the protocol is the handshake procedure, which enables the transmission of data. Since different devices (with different data rate capabilities) may be listening to the same talker, the handshake protocol must take into account these different capabilities. Let us discuss a typical handshake sequence that leads to transmission of data on an IEEE 488 bus. The three handshake lines used in the IEEE 488 have important characteristics that give the interface system wide flexibility, allowing interconnection of multiple devices that may operate at different speeds. The slowest active device controls the rate of data transfer, and more than one device can accept data simultaneously. The timing diagram of Figure 5.11.35 is used to illustrate the sequence in which the handshake and data transfer are performed:

1. All active listeners use the not ready for data (NRFD) line to indicate their state of readiness to accept a new piece of information. Nonreadiness to accept data is indicated if the NRFD line is held at 0 V. If even one active listener is not ready, the NRFD line of the entire bus is kept at 0 V and the active talker will not transmit the next byte. When all active listeners are ready and they have released the NRFD line, it now goes high.
2. The designated talker drives all eight data input/output lines, causing valid data to be placed on them.
3. Two microseconds after putting valid data on the data lines, the active talker pulls the data valid (DAV) line to 0 V and thereby signals the active listeners to read the information on the data bus. The 2-μsec interval is required to allow the data put on the data lines to reach (settle to) valid logic levels.
4. After the DAV is asserted, the listeners respond by pulling the NRFD line back down to zero. This prevents any additional data transfers from being initiated. The listeners also begin accepting the data byte at their own rates.
5. When each listener has accepted the data, it releases the not data accepted (NDAC) line. Only when the last active listener has released its hold on the NDAC line will that line go to its high-voltage-level state.
6. (a) When the active talker sees that NDAC has come up to its high state, it stops driving the data line. (b) At the same time, the talker releases the DAV line, ending the data transfer. The talker may now put the next byte on the data bus.
7. The listeners pull down the NDAC line back to 0 V and put the byte “away”.

Each of the instruments present on the data bus is distinguished by its own address, which is known to the controller; thus, the controller determines who the active talkers and listeners are on the bus by *addressing* them. To implement this and other functions, the controller uses the five control lines. Of these, ATN (attention) is used as a switch to indicate whether the controller is addressing or instructing the devices on the bus, or whether data transmission is taking place: when ATN is logic 1, the data lines contain either control information or addresses; with ATN = 1, only the controller is enabled to talk. When ATN = 0, only the devices that have been addressed can use the data lines. The IFC (interface clear) line is used to initialize the bus, or to clear it and reset it to a known condition in case of incorrect transmission. The REN (remote enable) line enables a remote instrument to be controlled by the bus;

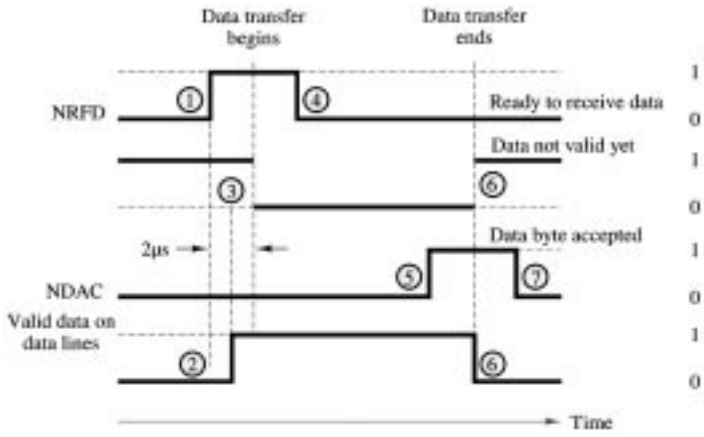


FIGURE 5.11.35 IEEE 488 data transmission protocol.

thus, any function that might normally be performed manually on the instrument (e.g., selecting a range or mode of operation) is now controlled by the bus via the data lines. The SRQ (service request) line is used by instruments on the bus whenever the instrument is ready to send or receive data; however, it is the controller who decides when to service the request. Finally, the EOI (end or identify) line can be used in two modes; when it is used by a talker, it signifies the end of a message; when it is used by the controller, it serves as a *polling* line, that is, a line used to interrogate the instrument about its data output.

Although it was mentioned earlier that the IEEE 488 bus can be used only over distances of up to 20 m, it is possible to extend its range of operation by connecting remote IEEE 488 bus systems over telephone communication lines. This can be accomplished by means of *bus extenders*, or by converting the parallel data to serial form (typically, in RS-232 format) and by transmitting the serial data over the phone lines by means of a modem. Serial communications and the RS-232 standard are discussed in the next section.

The RS-232 Standard

The primary reason why parallel transmission of data is not used exclusively is the limited distance range over which it is possible to transmit data on a parallel bus. Although there are techniques which permit extending the range for parallel transmission, these are complex and costly. Therefore, **serial transmission** is frequently used whenever data is to be transmitted over a significant distance. Since serial data travel along one single path and are transmitted one bit at a time, the cabling costs for long distances are relatively low; further, the transmitting and receiving units are also limited to processing just one signal and are also much simpler and less expensive. Two modes of operation exist for serial transmission: **simplex**, which corresponds to transmission in one direction only; and **duplex**, which permits transmission in either direction. Simplex transmission requires only one receiver and one transmitter, at each end of the link; on the other hand, duplex transmission can occur in one of two manners: **half-duplex** and **full-duplex**. In the former, although transmission can take place in both directions, it cannot occur simultaneously in both directions; in the latter case, both ends can simultaneously transmit and receive. Full-duplex transmission is usually implemented by means of four wires.

The data rate of a serial transmission line is measured in bits per second, since the data are transmitted one bit at a time. The unit of 1 bit/sec is called a **baud**; thus, reference is often made to the baud rate of a serial transmission. The baud rate can be translated into a parallel transmission rate in words per second if the structure of the word is known; for example, if a word consists of 10 bits (start and stop bits plus an 8-bit data word) and the transmission takes place at 1200 baud, 120 words are being transmitted every second. Typical data rates for serial transmission are standardized; the most common rates (familiar to the users of personal computer modem connections) are 300, 600, 1200, and 2400 baud. Baud rates can be as low as 50 baud or as high as 19,200 baud.

Like parallel transmission, serial transmission can also occur either synchronously or asynchronously. In the serial case, it is also true that asynchronous transmission is less costly but not as fast. A handshake protocol is also required for asynchronous serial transmission, as explained in the following. The most popular data-coding scheme for serial transmission is, once again, the ASCII code, consisting of a 7-bit word plus a **parity bit**, for a total of 8 bits per character. The role of the parity bit is to permit error detection in the event of erroneous reception (or transmission) of a bit. To see this, let us discuss the sequence of handshake events for asynchronous serial transmission and the use of parity bits to correct for errors. In serial asynchronous systems, handshaking is performed by using start and stop bits at the beginning and end of each character that is transmitted. The beginning of the transmission of a serial asynchronous word is announced by the "start" bit, which is always a 0-state bit. For the next five to eight successive bit times (depending on the code and the number of bits that specify the word length in the code), the line is switched to the 1 and 0 states required to represent the character being sent. Following the last bit of the data and the parity bit (which will be explained next), there is 1 bit or more in the 1 state, indicating "idle". The time period associated with this transmission is called the "stop" bit interval.

If noise pulses affect the transmission line, it is possible that a bit in the transmission could be misread. Thus, following the 5 to 8 transmitted data bits, there is a party bit that is used for error detection. Here is how the parity bit works. If the transmitter keeps track of the number of 1s in the word being sent, it can send a parity bit, a 1 or a 0, to ensure that the total number of 1s sent is always even (even parity) or odd (odd parity). Similarly, the receiver can keep track of the 1s received to see whether there was an error with the transmission. If an error is detected, retransmission of the word can be repeated.

Serial data transmission occurs most frequently according to the **RS-232 standard**. The RS-232 standard is based on the transmission of voltage pulses at a preselected baud rate; the voltage pulses are in the range –3 to –15 V for a logic 0 and in the range of +3 to +15 V for a logic 1. It is important to note that this amounts to a negative logic convention and that the signals are *not* TTL-compatible. The distance over which such transmission can take place is up to approximately 17 m (50 ft). The RS-232 standard was designed to make the transmission of digital data compatible with existing telephone lines; since phone lines were originally designed to carry analog voice signals, it became necessary to establish some standard procedures to make digital transmission possible over them. The resulting standard describes the mechanical and electrical characteristics of the interface between *data terminal equipment* (DTE) and *data communication equipment* (DCE). DTE consists of computers, terminals, digital instruments, and related peripherals; DCE includes all of those devices that are used to encode digital data in a format that permits their transmission over telephone lines. Thus, the standard specified how data should be presented by the DTE to the DCE so that digital data can be transmitted over voice lines.

A typical example of DCE is the **modem**. A modern converts digital data to audio signals that are suitable for transmission over a telephone line and is also capable of performing the reverse function, by converting the audio signals back to digital form. The term *modem* stands for *mod*ulate-*dem*odulate, because a modem modulates to a sinusoidal carrier using digital pulses (for transmission) and demodulates the modulated sinusoidal signal to recover the digital pulses (at reception). Three methods are commonly used for converting digital pulses to an audio signal: **amplitude-shift keying, frequency-shift keying**, and **phase-shift keying**, depending on whether the amplitude, phase, or frequency of the sinusoid is modulated by the digital pulses. Figure 5.11.36 depicts the essential block of a data transmission system based on the RS-232 standard, as well as examples of digital data encouded for transmission over a voice line.

In addition to the function just described, however, the RS-232 standard also provides a very useful set of specifications for the direct transmission of digital data between computers and instruments. In other words, communication between digital terminal instruments may occur directly in digital form (i.e., without digital communication devices encoding the digital data with analog voice lines). Thus, this standard is also frequently used for direct digital communication.

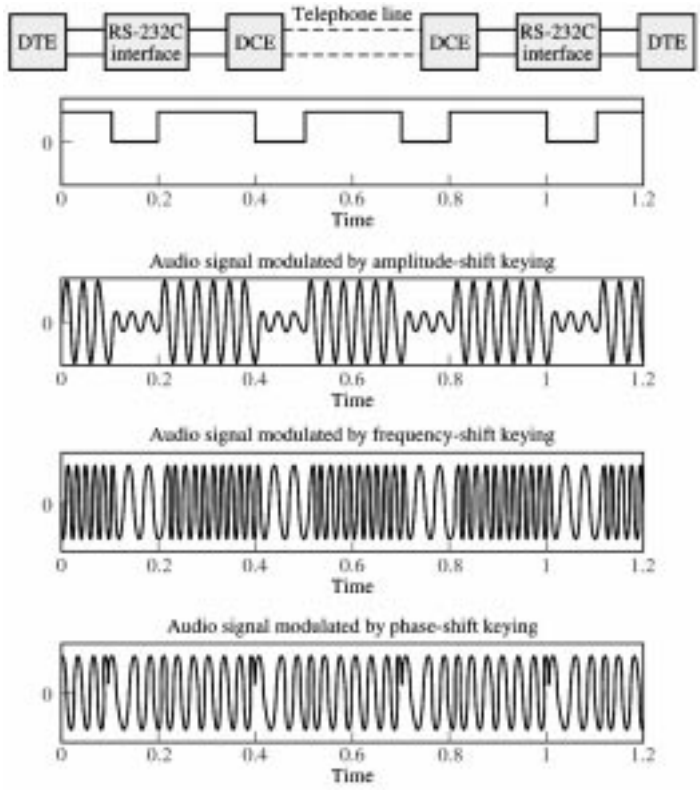


FIGURE 5.11.36 Digital data encoded for analog transmission.

The RS-232 standard can be summarized as follows:

- Data signals are encoded according to a negative logic convention using voltage levels of –3 to –15 V for logic 1 and +3 to +15 V for logic 0.
- Control signals use a positive logic convention (opposite to that of data signals).
- The maximum shunt capacitance of the load cannot exceed 2500 pF; this, in effect, limits the maximum length of the cables used in the connection.
- The load resistance must be between 300 Ω and 3 kΩ.
- Three wires are used for data transmission. One wire each is used for receiving and transmitting data; the third wire is a signal return line (signal ground). In addition, there are 22 wires that can be used for a variety of control purposes between the DTE and DCE.
- The male part of the connector is assigned to the DTE and the female part to the DCE. Figure 5.11.37 labels each of the wires in the 25-pin connector. Since each side of the connector has a *receive* and a *transmit* line, it has been decided by convention that the DCE transmits on the transmit line and receives on the receive line, while the DTE receives on the transmit line and transmits on the receive line.
- The baud rate is limited by the length of the cable; for a 17-m length, any rate from 50 baud to 19.2 kbaud is allowed. If a longer cable connection is desired, the maximum baud rate will decrease according to the length of the cable and **line drivers** can be used to amplify the signals, which are transmitted over twisted-air wires. Line drivers are simply signal amplifiers that are used directly on the digital signal, prior to encoding. For example, the signal generated by a DTE device (say, a computer) may be transmitted over a distance of up to 3300 m (at a rate of 600 baud) prior to being encoded by the DCE.

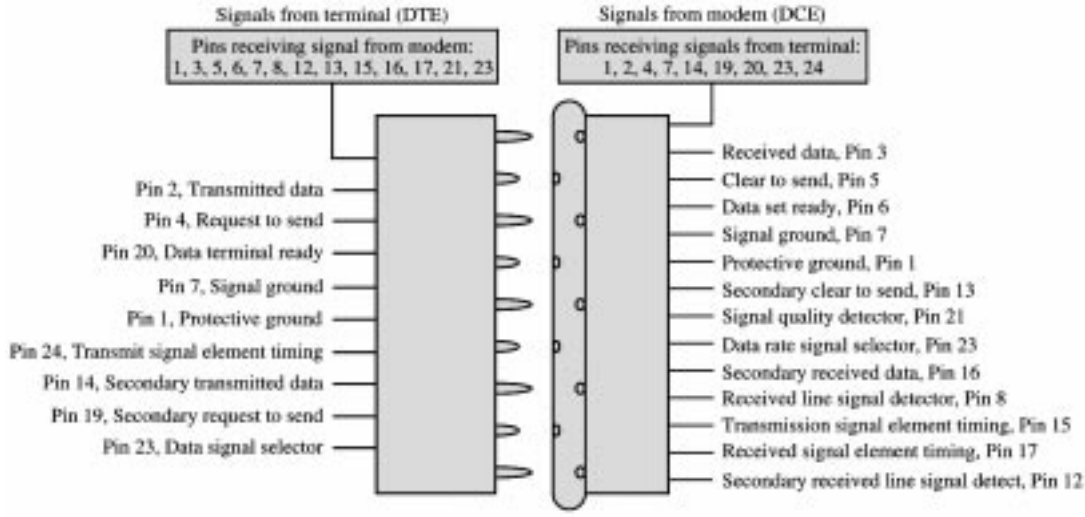


FIGURE 5.11.37 RS-232 connections.

- The serial data can be encoded according to any code, although the ASCII code is by far the most popular.

Other Communication Network Standards

In addition to the popular RS-232 and IEEE bus standards we should mention other communication standards that have become or are rapidly becoming commonplace. Once is the *Ethernet*, which operates at 10 Mb/sec and is based on IEEE Standard 802.3. This is commonly used in office networks. Higher speed networks include FDDI (fiber distributed data interface), which specifies an optical fiber ring with a data rate of 100 Mb/sec, and ATM (asynchronous transfer mode), a packet oriented transfer mode moving data in fixed-size packets called cells. ATM does not operate at a fixed speed. Typical speed is 155 Mb/sec, but there are implementations running as fast as 2 Gb/sec.

5.12 Electromechanical Systems

The objective of this section is to introduce the fundamental notions of electromechanical energy conversion, leading to an understanding of the operation of various electromechanical transducers

The Magnetic Field and Faraday's Law

The quantities used to measure the strength of a magnetic field are the **magnetic flux**, ϕ, in units of **webers** (Wb); and the **magnetic flux density**, B, in units of webers per square(Wb/m^2), or **teslas** (T). The latter quantity, as well as the associated **magnetic field intensity,** H (in units of amperes per meter, or A/m), are vectors. Thus, the density of the magnetic flux and its intensity are in general described in vector form, in terms of the components present in each spatial direction (e.g., on the *x*, *y*, and *z* axes). In discussing magnetic flux density and field intensity in this chapter and the next, we shall almost always assume that the field is a *scalar field*, that is, that it lies in a single spatial direction. This will simplify many explanations.

It is customary to represent the magnetic field by means of the familiar *lines of force* (a concept also due to Faraday); we visualize the strength of a magnetic field by observing the density of these lines in space. You probably know from a previous course in physics that such lines are closed in a magnitude field, that is, that they form continuous loops exiting at a magnetic north pole (by definition) and entering at a magnetic south pole. The relative strengths of the magnetic fields generated by two magnets could be depicted as shown in Figure 5.12.1.

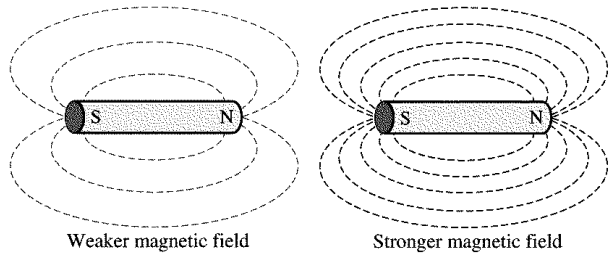


FIGURE 5.12.1 Lines of force in a magnetic field.

Magnetic fields are generated by electric charge in motion, and their effect is measured by the force they exert on a moving charge. As you may recall from previous physics courses, the vector force **f** exerted on a charge of *q* moving at velocity **u** in the presence of a magnetic field with flux density **B** is given by the equation

$$\mathbf{f} = q\mathbf{u} \times \mathbf{B} \tag{5.12.1}$$

where the symbol × denotes the (vector) cross product. If the charge is moving at a velocity **u** in a direction that makes an angle θ with the magnetic field, then the magnitude of the force is given by

$$f = quB\sin\theta \tag{5.12.2}$$

and the direction of this force is at right angles with the plane formed by the vectors **B** and **u**. This relationship is depicted in Figure 5.12.2.

The magnetic flux, ϕ, is then defined as the integral of the flux density over some surface area. For the simplified (but often useful) case of magnetic flux lines perpendicular to a cross-sectional area *A*, we can see that the flux is given by the following integral:

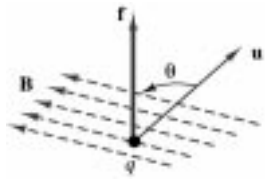


FIGURE 5.12.2 Charge moving in a constant magnetic field.

$$\phi = \int_A B\, dA \tag{5.12.3}$$

in webers (Wb), where the subscript A indicates that the integral is evaluated over the surface A. Furthermore, if the flux were to be uniform over the cross-sectional area A (a simplification that will be useful), the preceding integral could be approximated by the following expression:

$$\phi = B \cdot A \tag{5.12.4}$$

Figure 5.12.3 illustrates this idea by showing hypothetical magnetic flux lines traversing a surface, delimited in the figure by a thin conducting wire.

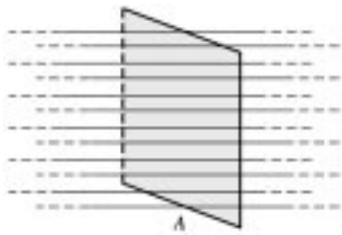


FIGURE 5.12.3 Magnetic flux lines crossing a surface.

Faraday's law states that if the imaginary surface A were bounded by a conductor — for example, the thin wire of Figure 5.12.3 — then a *changing* magnetic field would induce a voltage, and therefore a current, in the conductor. More precisely, Faraday's law states that a time-varying flux causes an induced **electromotive force**, or emf, e, as follows:

$$e = -\frac{d\phi}{dt} \tag{5.12.5}$$

In practical applications, the size of the voltages induced by the changing magnetic field can be significantly increased if the conducting wire is coiled many times around, so as to multiply the area crossed by the magnetic flux lines many times over. For an N-turn coil with cross-sectional area A, for example, we have the emf

$$e = N\frac{d\phi}{dt} \tag{5.12.6}$$

Figure 5.12.4 shows an N-turn coil *linking* a certain amount of magnetic flux; you can see that if N is very large and the coil is tightly wound (as is usually the case in the construction of practical devices), it is not unreasonable to presume that each turn of the coil links the same flux. It is convenient, in practice, to define the **flux linkage**, λ, as

$$\lambda = N\phi \tag{5.12.7}$$

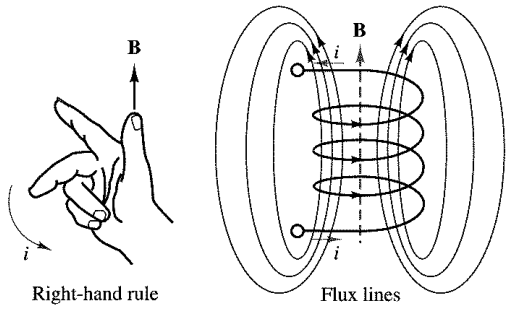


FIGURE 5.12.4 Concept of flux linkage.

so that

$$e = \frac{d\lambda}{dt} \tag{5.12.8}$$

In the analysis of linear circuits, we implicitly assumed that the relationship between flux linkage and current was a linear one:

$$\lambda = Li \tag{5.12.9}$$

so that the effect of a time-varying current was to induce a transformer voltage across an inductor coil, according to the expression

$$v = L\frac{di}{dt} \tag{5.12.10}$$

This is, in fact, the defining equation for the ideal **self-inductance**, *L*. In addition to self-inductance, however, it is also important to consider the **magnetic coupling** that can occur between neighboring circuits. Self-inductance measures the voltage induced in a circuit by the magnetic field generated by a current flowing in the same circuit. It is also possible that a second circuit in the vicinity of the first may experience an induced voltage as a consequence of the magnetic field generated in the first circuit. This principle underlies the operation of all transformers.

Self- and Mutual Inductance

Figure 5.12.5 depicts a pair of coils, one of which, L_1, is excited by a current, i_1, and therefore develops a magnetic field and a resulting induced voltage, v_1, The second coil, L_2, is not energized by a current, but links some of the flux generated by the current i_1 and L_1 because of its close proximity to the first coil. The magnetic coupling between the coils established by virtue of their proximity is described by a quantity called **mutual inductance** and defined by the symbol *M*. The mutual inductance is defined by the equation

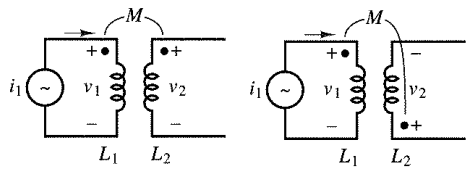


FIGURE 5.12.5 Mutual inductance.

$$v_2 = M\frac{di_1}{dt} \tag{5.12.11}$$

The dots shown in the two figures indicate the polarity of the coupling between the coils. If the dots are at the same end of the coils, the voltage induced in coil 2 by a current in coil 1 has the same polarity as the voltage induced by the same current in coil 1; otherwise, the voltages are in opposition, as shown in the lower part of Figure 5.12.5. Thus, the presence of such dots indicates that magnetic coupling is present between two coils, It should also be pointed out that if a current (and therefore a magnetic field) were present in the second coil, an additional voltage would be induced across coil 1. The voltage induced across a coil is, in general, equal to the sum of the voltages induced by self-inductance and mutual inductance.

Example 5.12.1 Linear Variable Differential Transformer (LVDT)

The linear variable differential transformer (LVDT) is a displacement transducer based on the mutual inductance concept just discussed. Figure 5.12.6 shows a simplified representation of an LVDT, which consists of a primary coil, subject to AC excitation (v_{ex}) and of a pair of identical secondary coils, which are connected so as to result in the output voltage

$$v_{out} = v_1 - v_2$$

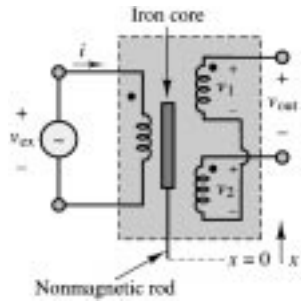


FIGURE 5.12.6 Linear variable differential transformer.

The ferromagnetic core between the primary and secondary coils can be displaced in proportion to some external motion, x, and determines the magnetic coupling between primary and secondary coils. Intuitively, as the core is displaced upward, greater coupling will occur between the primary coil and the top secondary coil, thus inducing a greater voltage in the top secondary coil. Hence, $v_{out} > 0$ for positive displacements. The converse is true for negative displacements. More formally, if the primary coil has resistance R_p and self-inductance L_p, we can write

$$iR_p + L_p\frac{di}{dt} = v_{ex}$$

and the voltages induced in the secondary coils are given by

$$v_1 = M_1\frac{di}{dt}$$

$$v_2 = M_2\frac{di}{dt}$$

so that

$$v_{\text{out}} = \left(M_1 - M_2\right)\frac{di}{dt}$$

where M_1 and M_2 are the mutual inductances between the primary and the respective secondary coils, It should be apparent that each of the mutual inductances is dependent on the position of the iron core. For example, with the core at the *null position*, $M_1 = M_2$ and $v_{\text{out}} = 0$. The LVDT is typically designed so that $M_1 - M_2$ is linearly related to the displacement of the core, x.

Because the excitation is by necessity an AC signal, the output voltage is actually given by the difference of two sinusoidal voltages at the same frequency and is therefore itself a sinusoid, whose amplitude and phase depend on the displacement, x. Thus, v_{out} is an *amplitude-modulated* (AM) signal. To recover a signal proportional to the actual displacement, it is therefore necessary to use a demodulator circuit.

In practical electromagnetic circuits, the self-inductance of a circuit is not necessarily constant; in particular, the inductance parameter, L, is not constant, in general, but depends on the strength of the magnetic field intensity, so that it will not be possible to use such a simple relationship as $v = L\, di/dt$, with L constant. If we revisit the definition of the transformer voltage,

$$e = N\frac{d\phi}{dt} \tag{5.12.12}$$

We see that in an inductor coil, the inductance is given by

$$L = \frac{N\phi}{i} = \frac{\lambda}{i} \tag{5.12.13}$$

This expression implies that the relationship between current and flux in a magnetic structure is linear (the inductance being the slope of the line). In fact, the properties of ferromagnetic materials are such that the flux-current relationship is nonlinear, so that the simple linear inductance parameter used in electric circuit analysis is not adequate to represent the behavior of the magnetic circuits of the present chapter. In any practical situation, the relationship between the flux linkage, λ, and the current is nonlinear, and might be described by a curve similar to that shown in Figure 5.12.7. Wherever the i-λ curve is not a straight line, it is more convenient to analyze the magnetic system in terms of energy calculations, since the corresponding circuit equation would be nonlinear.

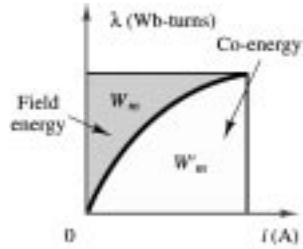


FIGURE 5.12.7

In a magnetic system, the energy stored in the magnetic field is equal to the integral of the instantaneous power, which is the product of voltage and current, just as in a conventional electrical circuit.

$$W_m = \int ei\, dt \tag{5.12.14}$$

However, in this case, the voltage corresponds to the induced emf, according to Faraday's law:

$$e = \frac{d\lambda}{dt} = N\frac{d\phi}{dt} \tag{5.12.15}$$

and is therefore related to the rate of change of the magnetic flux. The energy stored in the magnetic field could therefore be expressed in terms of the current by the integral

$$W_m = \int ei\,dt = \int \frac{d\lambda}{dt} i\,dt = \int i\,d\lambda \tag{5.12.16}$$

It should be straightforward to recognize that this energy is equal to the area above the λ-i curve of Figure 5.12.7. From the same figure, it is also possible to define a fictitious (but sometimes useful) quantity called **co-energy**, equal to the area under and identified by the symbol W'_m. From the figure, it is also possible to see that the co-energy can be expressed in terms of the stored energy by means of the following relationship:

$$W'_m = i\lambda - W_m \tag{5.12.17}$$

Ampère's Law

As explained in the previous section, Faraday's law is one of two fundamental laws relating electricity to magnetism. The second relationship, which forms a counterpart to Faraday's law, is **Ampère's law**. Qualitatively, Ampère's law states that the magnetic field intensity, **H**, in the vicinity of a conductor is related to the current carried by the conductor; thus Ampère's law establishes a dual relationship with Faraday's law.

In the previous section, we described the magnetic field in terms of its flux density, **B**, and flux ϕ, To explain Ampère's law and the behavior of magnetic materials, we need to define a relationship between the magnetic field intensity, **H**, and the flux density, **B**. These quantities are related by:

$$\begin{aligned} \mathbf{B} &= \mu\mathbf{H} \\ &= \mu_r\mu_0\mathbf{H} \quad \text{Wb/m}^2 \text{ or T} \end{aligned} \tag{5.12.18}$$

where the parameter μ is a scalar constant for a particular physical medium (at least, for the applications we consider here) and is called the **permeability** of the medium. The permeability of a material can be factored as the product of the permeability of free space, $\mu_0 = 4\pi \times 107$ H/m, times the relative permeability, μ_r, which varies greatly according to the medium. For example, for air and for most electrical conductors and insulators, μ_r is equal to 1. For ferromagnetic materials, the value of μ_r can take values in the hundreds or thousands. The size of μ_r represents a measure of the magnetic properties of the material. A consequence of Ampère's law is that the larger the value of μ, the smaller the current required to produce a large flux density in an electromagnetic structure. Consequently, many electromechanical devices make use of ferromagnetic materials, called iron cores, to enhance their magnetic properties. Table 5.12.1 gives approximate values of μ_r for some common materials.

Conversely, the reason for introducing the magnetic field intensity is that it is dependent of the properties of the materials employed in the construction of magnetic circuits. Thus, a given magnetic field intensity, **H**, will give rise to different flux densities in different materials. It will therefore be useful to define *sources* of magnetic energy in terms of the magnetic field intensity, so that different magnetic structures and materials can then be evaluated or compared for a given source. In analogy with electromotive force, this "source" will be termed **magnetomotive force** (mmf). As stated earlier, both the

TABLE 5.12.1 Relative Permeabilities for Common Materials

Material	μ_r
Air	1
Permalloy	100,000
Cast steel	1,000
Sheet steel	4,000
Iron	5,195

magnetic flux density and field intensity are vector quantities; however, for ease of analysis, scalar fields will be chosen by appropriately selecting the orientation of the fields, wherever possible.

The field generated by a single conducting wire is not very strong; however, if we arrange the wire into a tightly wound coil with many turns, we can greatly increase the strength of the magnetic field. For such a coil with N turns, one can verify visually that the lines of force associated with the magnetic field link all of the turns of the conducting coil, so that we have effectively increased the current linked by the flux lines N-fold. The product $N \cdot i$ is a useful quantity in electromagnetic circuits and is called the magnetomotive force, $\mathcal{F}$ (often abbreviated mmf), in analogy with the electromotive force defined earlier:

$$\mathcal{F} = Ni \quad \text{ampere-turns } (A \cdot t) \tag{5.12.19}$$

Typical arrangements are the iron-core inductor and the toroid of Figure 5.12.8. The flux densities for these inductors are given by the expressions

$$B = \frac{\mu Ni}{l} \quad \text{Flux density for tightly wound circular coil} \tag{5.12.20}$$

$$B = \frac{\mu Ni}{2\pi r_2} \quad \text{Flux density for toroidal coil} \tag{5.12.21}$$

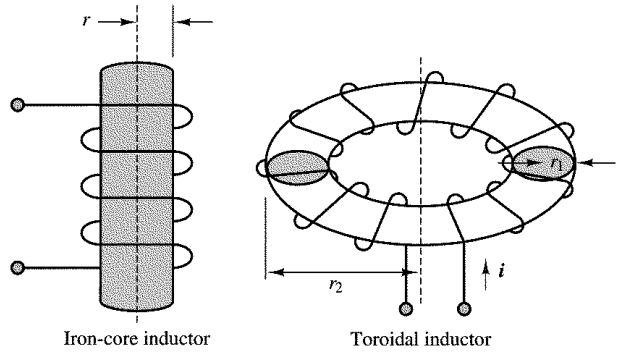


FIGURE 5.12.8 Practical inductors.

Intuitively, the presence of a high-permeability material near a source of magnetic flux causes the flux to preferentially concentrate in the high-μ material, rather than in air, much as a conducting path concentrates the current produced by an electric field in an electric circuit. Figure 5.12.9 depicts an example of a simple electromagnetic structure, which, as we shall see shortly, forms the basis of the practical transformer.

Table 5.12.2 summarizes the variables introduced thus far in the discussion of electricity and magnetism.

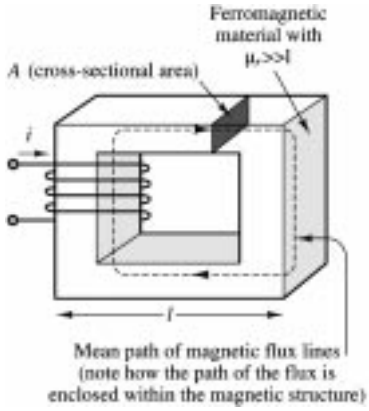


FIGURE 5.12.9 A simple electromagnetic structure.

TABLE 5.12.2 Magnetic Variables and Units

Variable	Symbol	Units
Current	I	A
Magnetic flux density	B	$Wb/m^2 = T$
Magnetic flux	ϕ	Wb
Magnetic field intensity	H	A/m
Electromotive force	e	V
Magnetomotive force	$\mathcal{F}$	A · t
Flux linkage	λ	Wb · t

Magnetic Circuits

It is possible to analyze the operation of electromagnetic devices such as the one depicted in Figure 5.12.9 by means of magnetic equivalent circuits, similar in many respects to the equivalent electrical circuits of the earlier chapters. Before we can present this technique, however, we need to make a few simplifying approximations. The first of these approximations assumes that there exists a **mean path** for the magnetic flux, and that the corresponding mean flux density is approximately constant over the cross-sectional area of the magnetic structure. Thus, a coil wound around a coil with cross-sectional area A will have flux density

$$B = \frac{\phi}{A} \tag{5.12.22}$$

where A is assumed to be perpendicular to the direction of the flux lines. Figure 5.12.9 illustrates such a mean path and the cross-sectional area, A. Knowing the flux density, we obtain the field intensity:

$$H = \frac{B}{\mu} = \frac{\phi}{A\mu} \tag{5.12.23}$$

But then, knowing the field intensity, we can relate the mmf of the coil, $\mathcal{F}$, to the product of the magnetic field intensity, H, and the length of the magnetic (mean) path, l, for one leg of the structure:

$$\mathcal{F} = N \cdot i = H \cdot l \tag{5.12.24}$$

In summary, the mmf is equal to the magnetic flux times the length of the magnetic path, divided by the permeability of the material times the cross-sectional area:

$$\mathcal{F} = \phi \frac{l}{\mu A} \tag{5.12.25}$$

A review of this formula reveals that the magnetomotive force, $\mathcal{F}$, may be viewed as being analogous to the voltage source in a series electrical circuit, and that the flux, ϕ, is then equivalent to the electrical current in a series circuit and the term $l/\mu A$ to the *magnetic resistance* of one leg of the magnetic circuit. You will note that the term $l/\mu A$ is very similar to the term describing the resistance of a cylindrical conductor of length l and cross-sectional A, where the permeability, μ, is analogous to the conductivity, σ. The term $l/\mu A$ occurs frequently enough to be assigned the name of reluctance, and the symbol $\mathcal{R}$.

In summary, when an N-turn coil carrying a current i is wound around a magnetic core such as the one indicated in Figure 5.12.9, the mmf, $\mathcal{F}$, generated by the coil produces a flux, ϕ, that is *mostly* concentrated with the core and is assumed to be uniform across the cross section. Within this simplified picture, then, the analysis of a magnetic circuit is analogous to that of resistive electrical circuits. This analogy is illustrated in Table 5.12.3 and in the examples in this section.

TABLE 5.12.3 Analogy between Electric and Magnetic Circuits

Electrical Quantity	Magnetic Quantity
Electrical field intensity, E, V/m	Magnetic field intensity, H, A · t/m
Voltage, v, V	Magnetomotive force, $\mathcal{F}$, A · t
Current, i, A	Magnetic flux, ϕ, Wb
Current density, J, A/m^2	Magnetic flux density, B, Wb/m^2
Resistance, R, Ω	Reluctance, $\mathcal{R} = l/\mu A$, A · t/Wb
Conductivity, σ, l/Ω · m	Permeability, μ, Wb/A · m

The usefulness of the magnetic circuit analogy can be emphasized by analyzing a magnetic core similar to that of Figure 5.12.9, but with a slightly modified geometry. Figure 5.12.10 depicts the magnetic structure and its equivalent circuit analogy. In the figure, we see that the mmf, $\mathcal{F}$ = Ni, excites the magnetic circuit, which is composed of four legs: two of mean path length l_1 and cross-sectional area $A_1 = d_1 w$, and the other two of mean length l_2 and cross section $A_2 = d_2 w$. Thus, the reluctance encountered by the flux in its path around the magnetic core is given by the quantity $\mathcal{R}_{\text{series}}$, with

$$\mathcal{R}_{\text{series}} = 2\mathcal{R}_1 + 2\mathcal{R}_2$$

and

$$\mathcal{R}_1 = \frac{l_1}{\mu A_1}$$
$$\mathcal{R}_2 = \frac{l_2}{\mu A_2} \tag{5.12.26}$$

It is important at this stage to review the assumptions and simplifications made in analyzing the magnetic structure of Figure 5.12.10:

1. All of the magnetic flux is linked by all of the turns of the coil.
2. The flux is confined exclusively within the magnetic core.
3. The density of the flux is uniform across the cross-sectional area of the core.

You can probably see intuitively that the first of these assumptions might not hold true near the ends of the coil, but that it might be more reasonable if the coil is tightly wound. The second assumption is equivalent to stating that the relative permeability of the core is infinitely higher than that of air

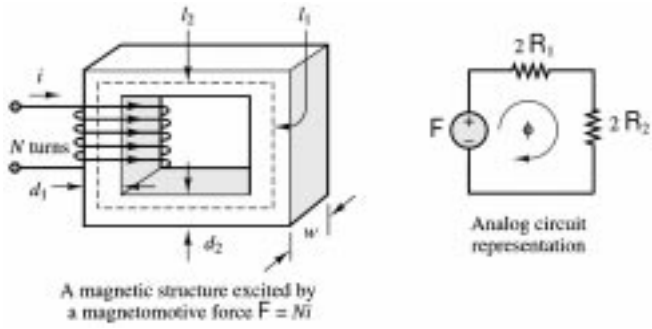


FIGURE 5.12.10 Analogy between magnetic and electric circuits.

(presuming that this is the medium surrounding the core): if this were the case, the flux would indeed be confined within the core. It is worthwhile to note that we made a similar assumption when we treat wires in electric circuits as perfect conductors: the conductivity of copper is substantially greater than that of free space, by a factor of approximately 10^{15}. In the case of magnetic materials, however, even for the best alloys, we have a relative permeability only on the order of 10^3 to 10^4. Thus, an approximation that is quite appropriate for electric circuits is not nearly as good in the case of magnetic circuits. Some of the flux in a structure such as those of Figures 5.12.9 and 5.12.10 would thus not be confined within the core (this is usually referred to as **leakage flux**). Finally, the assumption that the flux is uniform across the core cannot hold for a finite-permeability medium, but it is very helpful in giving an approximate *mean* behavior of the magnetic circuit.

The magnetic circuit analogy is therefore far from being exact. However, short of employing the tools of electromagnetic field theory and of vector calculus, or advanced numerical simulation software, it is the most convenient tool at the engineer's disposal for the analysis of magnetic structures.

Example 5.12.2 Magnetic Reluctance Position Sensor

A simple magnetic structure, very similar to those examined in the previous examples, finds very common application in the so-called variable-reluctance position sensor, which, in turn, finds widespread application in a variety of configurations for the measurement of linear and angular position and velocity. Figure 5.12.11 depicts one particular configuration that is used in many applications. In this structure, a permanent magnet with a coil of wire wound around it forms the sensor; a steel disk (typically connected to a rotating shaft) has a number of tabs that pass between the pole pieces of the sensor. The area of the tab is assumed equal to the area of the cross section of the pole pieces and is equal to a^2. The reason for the name *variable-reluctance sensor* is that the reluctance of the magnetic structure is variable, depending on whether or not a ferromagnetic tab lies between the pole pieces of the magnet.

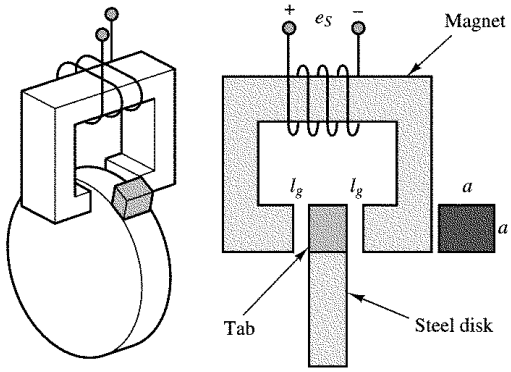


FIGURE 5.12.11 Variable-reluctance position sensor.

The principle of operation of the sensor is that an electromotive force, e_s, is induced across the coil by the change in magnetic flux caused by the passage of the tab between the pole pieces when the disk is in motion. As the tab enters the volume between the pole pieces, the flux will increase because of the lower reluctance of the configuration, until it reaches a maximum when the tab is centered between the poles of the magnet. Figure 5.12.12 depicts the approximate shape of the resulting voltage, which, according to Faraday's law, is given by

$$e_S = -\frac{d\phi}{dt}$$

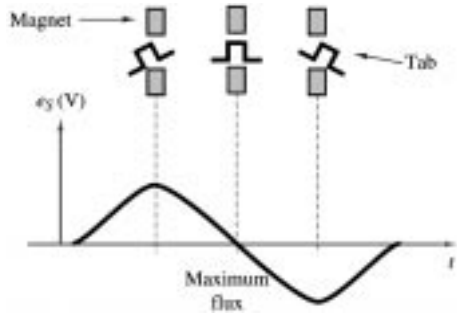


FIGURE 5.12.12 Variable-reluctance position sensor waveform.

The rate of change of flux is dictated by the geometry of the tab and of the pole pieces, and by the speed of rotation of the disk. It is important to note that, since the flux is changing only if the disk is rotating, this sensor cannot detect the static position of the disk.

One common application of this concept is in the measurement of the speed of rotation of rotating machines, including electric motors and internal combustion engines. In these applications, use is made of a *60-tooth wheel*, which permits the conversion of the speed rotation directly to units of revolutions per minute. The output of a variable-reluctance position sensor magnetically coupled to a rotating disk equipped with 60 tabs (teeth) is processed through a comparator or Schmitt trigger circuit. The voltage waveform generated by the sensor is nearly sinusoidal when the teeth are closely spaced, and it is characterized by one sinusoidal cycle for each tooth on the disk. If a negative zero-crossing detector is employed, the trigger circuit will generate a pulse corresponding to the passage of each tooth, as shown in Figure 5.12.13. If the time between any two pulses is measured by means of a high-frequency clock, the speed of the engine can be directly determined in units of rev/min by means of a digital counter.

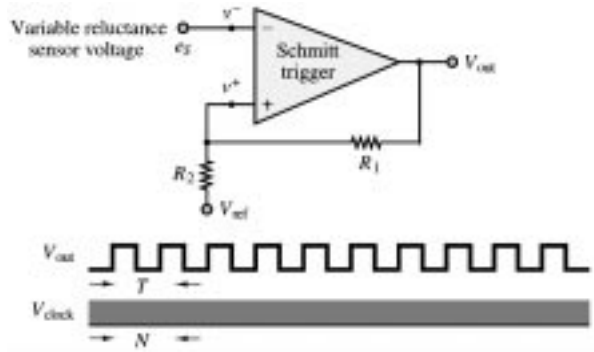


FIGURE 5.12.13 Signal processing for a 60-tooth wheel RPM sensor.

Magnetic Materials and *B-H* Curves

In the analysis of magnetic circuits presented in the previous sections, the relative permeability, μ_r, was treated as a constant. In fact, the relationship between the magnetic flux density, **B**, and the associated field intensity, **H**,

$$\mathbf{B} = \mu\mathbf{H} \tag{5.12.27}$$

is characterized by the fact that the relative permeability of magnetic materials is not a constant, but is a function of the magnetic field intensity. In effect, all magnetic materials exhibit a phenomenon called **saturation**, whereby the flux density increases in proportion to the field intensity until it cannot do so any longer. Figure 5.12.14 illustrates the general behavior of all magnetic materials. You will note that since the *B-H* curve shown in the figure is nonlinear, the value of μ (which is the slope of the curve) depends on the intensity of the magnetic field.

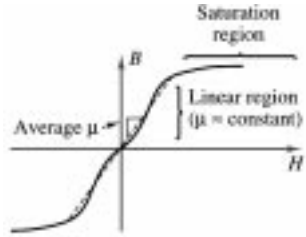


FIGURE 5.12.14 Permeability and magnetic saturation effects.

To understand the reasons for the saturation of a magnetic material, we need to briefly review the mechanism of magnetization. The basic idea behind magnetic materials is that the spin of electrons constitutes motion of charge and therefore leads to magnetic effects, as explained in the introductory section of this chapter. In most materials, the electron spins cancel out, on the whole, and no net effect remains. In ferromagnetic materials, on the other hand, atoms can align so that the electron spins cause a net magnetic effect. In such materials, there exist small regions with strong magnetic properties (called **magnetic domains**), the effects of which are neutralized in unmagnetized material by other, similar regions that are oriented differently, in a random pattern. When the material is magnetized, the magnetic domains tend to align with each other, to a degree that is determined by the intensity of the applied magnetic field.

In effect, the large number of miniature magnets within the material are *polarized* by the external magnetic field. As the field increases, more and more domains become aligned. When all of the domains have become aligned, any further increase in magnetic field intensity does not yield an increase in flux density beyond the increase that would be caused in a nonmagnetic material. Thus, the relative permeability, μ_r, approaches 1 in the saturation region. It should be apparent that an exact value of μ_r cannot be determined; the value of μ_r used in the earlier examples is to be interpreted as an average permeability, for intermediate values of flux density. As a point of reference, commercial magnetic steels saturate at flux densities around a few teslas. Figure 5.12.17, shown later in this section, will provide some actual *B-H* curves for common ferromagnetic materials.

The phenomenon of saturation carries some interesting implications with regard to the operation of magnetic circuits: the results of the previous section would seem to imply that an increase in the mmf (that is, an increase in the current driving the coil) would lead to a proportional increase in the magnetic flux. This is true in the *linear region* of Figure 5.12.14; however, as the material reaches saturation, further increases in the driving current (or, equivalently, in the mmf) do not yield further increases in the magnetic flux.

There are two more features that cause magnetic materials to further deviate from the ideal model of the linear *B-H* relationship: **eddy currents** and **hysteresis**. The first phenomenon consists of currents that are caused by any time-varying flux in the core material. As you know, a time-varying flux will

induce a voltage, and therefore a current. When this happens inside the magnetic core, the induced voltage will cause "eddy" currents (the terminology should be self-explanatory) in the core, which depend on the resistivity of the core. Figure 5.12.15 illustrates the phenomenon of eddy currents. The effect of these currents is to dissipate energy in the form of heat. Eddy currents are reduced by selecting high-resistivity core materials, or by *laminating* the core, introducing tiny, discontinuous air gaps between core layers (see Figure 5.12.15). Lamination of the core reduces eddy currents greatly without affecting the magnetic properties of the core.

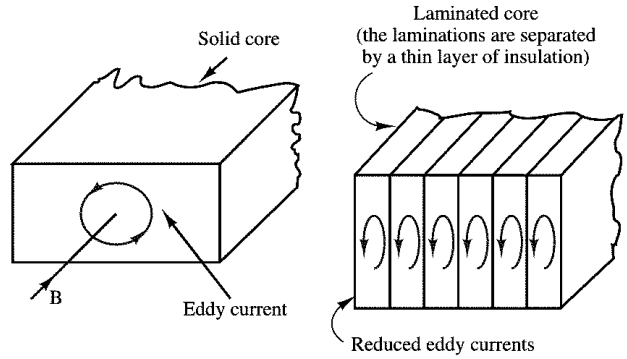


FIGURE 5.12.15 Eddy currents in magnetic structures.

It is beyond the scope of this section to quantify the losses caused by induced eddy currents, but it will be important to be aware of this source of energy loss.

Hysteresis is another loss mechanism in magnetic materials; it displays a rather complex behavior, related to the magnetization properties of a material. The curve of Figure 5.12.16 reveals that the *B*-*H* curve for a magnetic material during magnetization (as H is increased) is displaced with respect to the curve that is measured when the material is demagnetized. To understand the hysteresis process, consider a core that has been energized for some time, with a field intensity of $H_1A \cdot$ t/m. As the current required to sustain the mmf corresponding to H_1 is decreased, we follow the hysteresis curve from the point α to the point β. When the mmf is exactly zero, the material displays the **remanent** (or **residual**) **magnetization** B_r. To bring the flux density to zero, we must further decrease the mmf (i.e., produce a negative current), until the field intensity reaches the value $-H_0$ (point γ on the curve). As the mmf is made more negative, the curve eventually reaches the point α'. If the excitation current go the coil is now increased, the magnetization curve will follow the path $\alpha' = \beta' = \gamma' = \alpha$, eventually returning to the original point in the *B*-*H* plane, but via a different path.

The result of this process, by which an *excess magnetomotive force* is required to magnetize or demagnetize the material, is a net energy loss. It is difficult to evaluate this loss exactly; however, it can be shown that it is related to the area between the curves of Figure 5.2.16. There are experimental techniques that enable the approximate measurement of these losses.

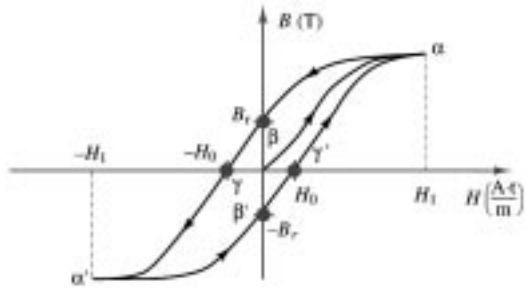
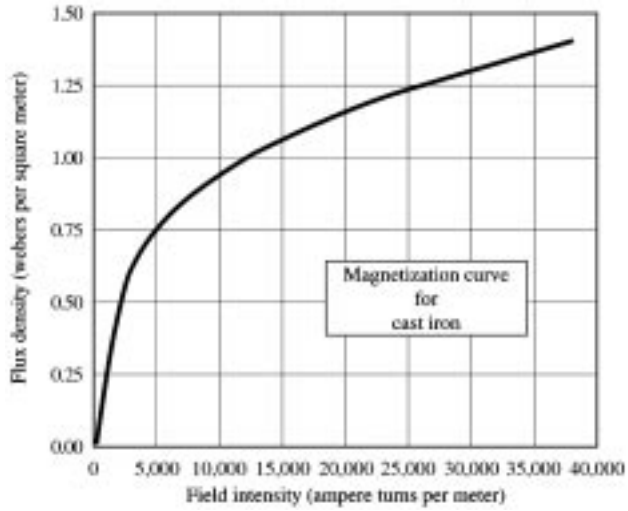
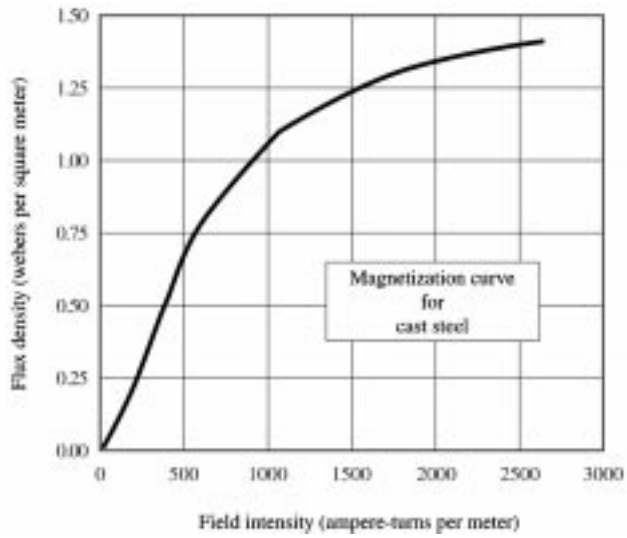


FIGURE 5.12.16 Hysteresis in magnetization curves.



(a)



(b)

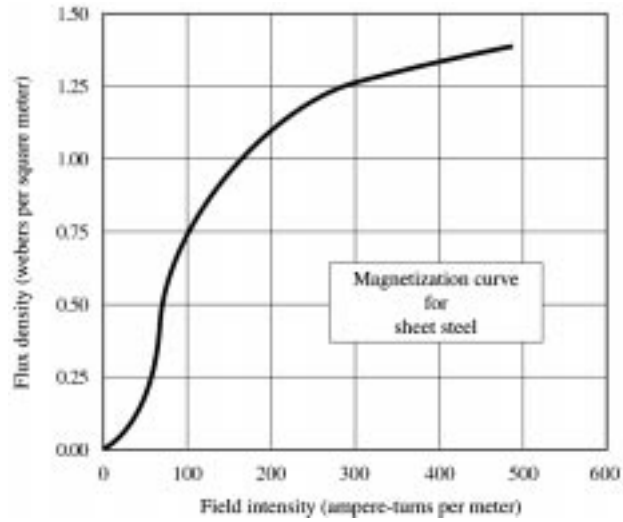
FIGURE 5.12.17 (a) Magnetization curve for cast iron; (b) magnetization curve for cast steel;

Figures 5.2.17(a) to (c) depict magnetization curves for three very common ferromagnetic materials: cast iron, cast steel, and sheet steel.

Electromechanical Energy Conversion

From the material developed thus far, it should be apparent that electromagnetomechanical devices are capable of converting mechanical forces and displacements to electromagnetic energy, and that the converse is also possible. The objective of this section is to formalize the basic principles of energy conversion in electromagnetomechanical systems, and to illustrate its usefulness and potential for application by presenting several examples of **energy transducers**. A transducer is a device that can convert electrical to mechanical energy (in this case, it is often called an **actuator**), or vice versa (in which case it is called a sensor).

Several physical mechanisms permit conversion of electrical to mechanical energy and back, the principal phenomena being the **piezoelectric effect**, consisting of the generation of a change in electric



(c)

FIGURE 5.12.17 (c) magnetization curve for sheet steel.

field in the presence of strain in certain crystals (e.g., quartz), and **electrostriction** and **magnetostriction**, in which changes in the dimension of certain materials lead to a change in their electrical (or magnetic) properties. Although these effects lead to some interesting applications, this chapter is concerned only with transducers in which electrical energy is converted to mechanical energy through the coupling of a magnetic field. It is important to note that all rotating machines (motors and generators) fit the basic definition of electromechanical transducers we have just given.

Forces in Magnetic Structures

It should be apparent by now that it is possible to convert mechanical forces to electrical signals, and vice versa, by means of the coupling provided by energy stored in the magnetic field. In this subsection, we discuss the computation of mechanical forces and of the corresponding electromagnetic quantities of interest; these calculations are of great practical importance in the design and application of electromechanical actuators. For example, a problem of interest is the computation of the current required to generate a given force in an electromechanical structure. This is the kind of application that is likely to be encountered by the engineer in the selection of an electromechanical device for a given task.

As already seen in this chapter, an electromechanical system includes an electrical system and a mechanical system, in addition to means through which the two can interact. The principal focus of this chapter has been the coupling that occurs through an electromagnetic field common to both the electrical and the mechanical system; to understand electromechanical energy conversion, it will be important to understand the various energy storage and loss mechanisms in the electromagnetic field. Figure 5.12.18 illustrates the coupling between the electrical and mechanical systems. In the mechanical system, energy loss can occur because of the heat developed as a consequence of *friction*, while in the electrical system, analogous losses are incurred because of *resistance*. Loss mechanisms are also present in the magnetic coupling medium, since *eddy current losses* and *hysteresis losses* are unavoidable in ferromagnetic materials. Either system can supply energy, and either system can store energy. Thus, the figure depicts the flow of energy from the electrical to the mechanical system, accounting for these various losses. The same flow could be reversed if mechanical energy were converted to electrical form.

Moving-Iron Transducers

One important class of electromagnetomechanical transducers is that of **moving-iron transducers**. The aim of this section is to derive an expression for the magnetic forces generated by such transducers and to illustrate the application of these calculations to simple, yet common devices such as electromagnets,

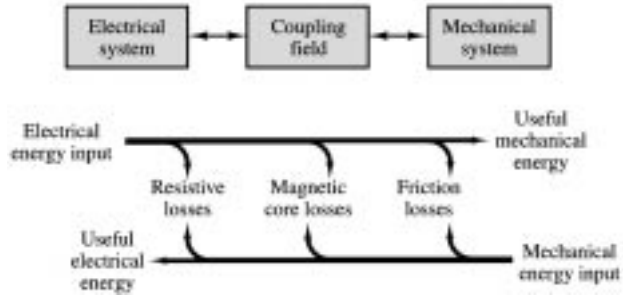


FIGURE 5.12.18 Electromechanical system.

solenoids, and relays. The simplest example of a moving-iron transducer is the electromagnet of Figure 5.12.19, in which the U-shaped element is fixed and the bar is movable. In the following paragraphs, we shall derive a relationship between the current applied to the coil, the displacement of the movable bar, and the magnetic force acting in the air gap.

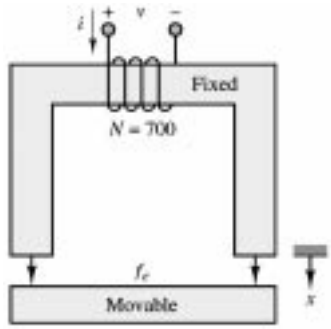


FIGURE 5.12.19 Simple electromagnet.

The principle that will be applied throughout the section is that in order for a mass to be displaced, some work needs to be done; this work corresponds to a change in the energy stored in the electromagnetic field, which causes the mass to be displaced. With reference to Figure 5.12.19, let f_e represent the magnetic force acting on the bar and x the displacement of the bar, in the direction shown. Then the net work into the electromagnetic field, W_m, is equal to the sum of the work done by the electrical circuit plus the work done by the mechanical system. On the basis of a linear approximation, it can be shown that the stored energy in a magnetic structure is given by the expression

$$W_m = \frac{\phi \mathcal{F}}{2} \tag{5.12.28}$$

and since the flux and the mmf are related by the expression

$$\phi = \frac{Ni}{\mathcal{R}} = \frac{\mathcal{F}}{\mathcal{R}} \tag{5.12.29}$$

the stored energy can be related to the reluctance of the structure according to

$$W_m = \frac{\phi^2 \mathcal{R}(x)}{2} \tag{5.12.30}$$

where the reluctance has been explicitly shown to be a function of displacement, as is the case in a moving-iron transducer. Finally, then, we shall use the following approximate expression to compute the magnetic force acting on the moving iron:

$$f = -\frac{dW_m}{dx} = -\frac{\phi^2}{2}\frac{d\mathcal{R}(x)}{dx} \tag{5.12.31}$$

Example 5.12.13 An Electromagnet

An electromagnet is used to support a solid piece of steel as shown in Figure 5.12.19. A force of 8900 N is required to support the weight. The cross-sectional area of the magnet core (the fixed part) is 0.01 m^2. Determine the minimum current that can keep the weight from falling for x = 1.5 mm. Assume negligible reluctance for the steel parts, and negligible fringing in the air gap.

Solution. We have already shown that in magnetic structures with air gaps, the reluctance is mostly due to the air gaps. This explains the assumption that the reluctance of the structure is negligible. For the structure of Figure 5.12.19, the reluctance is therefore given by

$$\mathcal{R}(x) = \frac{l}{\mu_0 A}$$

where A = 0.01 m^2 and $l = 2x$, and therefore

$$\mathcal{R}(x) = \frac{2x}{4\pi \times 10^{-7} \times 0.01} = \frac{x}{1.2566 \times 10^{-8}}$$

The magnitude of the force in the air gap is given by the expression

$$|f| = \frac{\phi^2}{2}\frac{d\mathcal{R}(x)}{dx} = \frac{N^2 i^2}{2\mathcal{R}^2}\frac{d\mathcal{R}(x)}{dx}$$

$$= \frac{i^2}{2}\frac{N^2}{\mathcal{R}^2}\frac{d\mathcal{R}}{dx} = \frac{i^2}{2}(700)^2\frac{6.2832 \times 10^{-9}}{x^2} = 8900 \text{ N}$$

from which the current can be computed:

$$i^2 = 2 \times \frac{8900\left(1.5 \times 10^{-3}\right)^2}{(700)^2\left(6.2832 \times 10^{-9}\right)} 6.504 \text{ A}$$

or

$$i = 2.55 \text{ A}$$

You should recognize the practical importance of these calculations in determining approximate current requirements and force-generation capabilities of electromechanical transducers.

Moving-Coil Transducers

Another important class of electromagnetomechanical transducers is that of **moving-coil transducers**. This class of transducers includes a number of common devices, such as microphones, loudspeakers, and all electric motors and generators. The aim of this section is to explain the relationship between a

fixed magnetic field, the emf across the moving coil, and the forces and motions of the moving element of the transducer.

Motor Action. A moving-coil transducer can act as a motor when an externally supplied current flowing through the electrically conducting part of the transducer is converted into a force that can cause the moving part of the transducer to be displaced. Such a current would flow, for example, if the support of Figure 5.12.20 were made of conducting material, so that the conductor and the right-hand side of the support "rail" were to form a loop (in effect, a 1-turn coil).

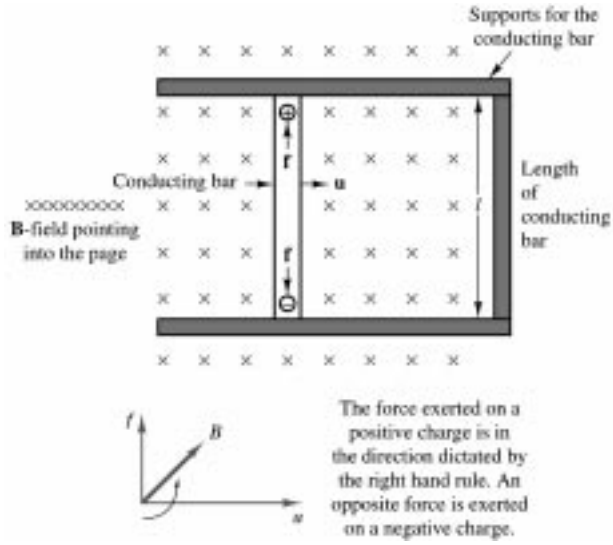


FIGURE 5.12.20 A simple electromechanical motion transducer.

The phenomenon we have just described is sometimes referred to as the "*Bli* law".

Generator Action. The other mode of operation of a moving-coil transducer occurs when an external force causes the coil (i.e., the moving bar, in Figure 5.12.20) to be displaced. This external force is converted to an emf across the coil, as will be explained in the following paragraphs.

It is important to observe that since positive and negative charges are forced in opposite directions in the transducer of Figure 5.12.20, a potential difference will appear across the conducting bar; this potential difference is the electromotive force, or emf. The emf must be equal to the force exerted by the magnetic field. In short, the electric force per unit charge (or electric field) e/l must equal the magnetic force per unit charge $f/q = Bu$. Thus, the relationship

$$e = Blu \tag{5.12.32}$$

which holds whenever **B, l,** and **u** are mutually perpendicular, as in Figure 5.12.20.

It was briefly mentioned that the *Blu* and *Bli* laws indicate that, thanks to the coupling action of the magnetic field, a conversion of mechanical to electrical energy — or the converse — is possible. The simple structure of Figures 5.12.20 can, again, serve as an illustration of this energy-conversion process, although we have not yet indicated how these idealized structures can be converted into a practical device. In this section we shall begin to introduce some physical considerations. Before we proceed any further, we should try to compute the power — electrical and mechanical — that is generated (or is required) by our ideal transducer. The electrical power is given by

$$P_E = ei = Blui \quad \text{(W)} \tag{5.12.33}$$

while the mechanical power required, say, to move the conductor from left to right is given by the product of force and velocity:

$$P_M - f_{\text{ext}} u = Bliu \quad \text{(W)} \tag{5.12.34}$$

Example 5.12.4 The Loudspeaker

A loudspeaker, shown in Figure 5.12.21, uses a permanent magnet and a moving coil to produce the vibrational motion that generates the pressure waves we perceive as sound. Vibration of the loudspeaker is caused by changes in the input current to a coil; the coil is, in turn, coupled to a magnetic structure that can produce time-varying forces on the speaker diaphragm. A simplified model for the mechanics of the speaker is also shown in Figure 5.12.21. The force exerted on the coil is also exerted on the mass of the speaker diaphragm, as shown in Figure 5.12.22, which depicts a free-body diagram of the forces acting on the loudspeaker diaphragm.

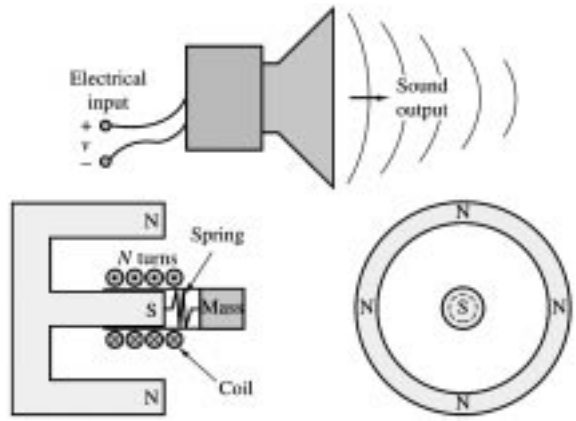


FIGURE 5.12.21 Loudspeaker.

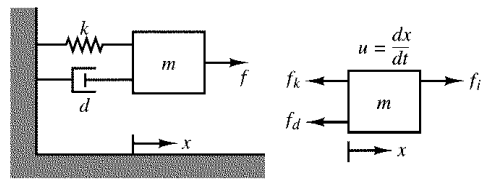


FIGURE 5.12.22 Forces acting on loudspeaker diaphragm.

The force exerted on the mass, f_i, is the magnetic force due to current flow in the coil. The electrical circuit that describes the coil is shown in Figure 5.12.23, where L represents the inductance of the coil, R represents the resistance of the windings, and e is the emf induced by the coil moving through the magnetic field.

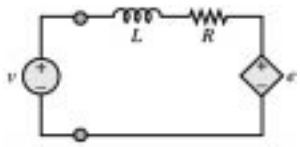


FIGURE 5.12.23 Model of transducer electrical field.

Determine the frequency response, $(\mathbf{U}/\mathbf{V})(j\omega)$, of the speaker using phasor analysis if the model parameters are

$L \approx 0$ H $\quad R = 8\ \Omega \quad m = 0.001$ kg $\quad d = 22.75$

$k \approx 500{,}000$ N/m $\quad N = 47 \quad B = 1T \quad$ Radius of coil = 5 cm

Solution: To determine the frequency response of the loudspeaker, we need to write the fundamental equations that describe the two subsystems that make up the loudspeaker. The electrical subsystem is described by the usual KVL relationship, applied to the circuit of Figure 5.12.23.

$$v = L\frac{di}{dt} + Ri + e$$

where e is the emf generated by the motion of the coil in the magnetic field. Next, according to Newton's law, we can write a force balance equation to describe the dynamics of the mechanical subsystem:

$$m\frac{du}{dt} = f_i - f_d - f_k = f_i - du - kx$$

Now, the coupling between the electrical and mechanical systems is expressed in each of the two preceding equations by the terms e and f_i:

$$e = Blu$$

$$f_i = Bli$$

Since we desire the frequency response, we use phasor techniques to represent the electrical and mechanical subsystem equations:

$$\mathbf{V}(j\omega) = j\omega L\mathbf{I}(j\omega) + R\mathbf{I}(j\omega) + Bl\mathbf{U}(j\omega) \qquad \text{Electrical equation}$$

$$(j\omega m + d)\mathbf{U}(j\omega) + \frac{K}{j\omega}\mathbf{U}(j\omega) = Bl\mathbf{I}(j\omega) \qquad \text{Mechanical equation}$$

Having assumed that the inductance of the coil is negligible, we are able to simplify the electrical equation and to solve for $\mathbf{I}(j\omega)$:

$$\mathbf{I}(j\omega) = \frac{\mathbf{V}(j\omega) - Bl\mathbf{U}(j\omega)}{R}$$

Substituting this equivalence into the mechanical equation and accounting for the length of the coil, $l = 2\pi Nr$, the final expression for the frequency response of the loudspeaker is then given by

$$\frac{\mathbf{U}}{\mathbf{V}}(j\omega) = \frac{2\pi NBr}{Rm} \times \frac{j\omega}{(j\omega)^2 + j\omega\left(\dfrac{d + \dfrac{(2\pi)^2 B^2 N^2 r^2}{R}}{m}\right) + \dfrac{k}{m}}$$

or, numerically,

$$\frac{\mathbf{U}}{\mathbf{V}}(j\omega) = \frac{2\pi \times 47 \times 1 \times 0.05}{8(0.001)} \times \frac{j\omega}{(j\omega)^2 + j\omega\left(\frac{22.75 + \frac{(2\pi)^2(1)^2(47)^2(0.05)^2}{8}}{0.001}\right) + \frac{500{,}000}{0.001}}$$

$$\approx \frac{1{,}845\, j\omega}{\left(j\omega + 13.8 \times 10^3\right)\left(j\omega + 36.2 \times 10^3\right)}$$

$$= \frac{0.051\left(\frac{j\omega}{13.8 \times 10^3}\right)}{\left(1 + \frac{j\omega}{36.2 \times 10^3}\right)\left(1 + \frac{j\omega}{13.8 \times 10^3}\right)}$$

This frequency response shows that the speaker has a lower cutoff frequency $f_{ct} = 13{,}800/2\pi \approx 2200$ Hz and an upper cutoff frequency of $f_{ch} = 36{,}000/2\pi \approx 5800$ Hz. In practice, a loudspeaker with such a frequency response would be useful only as a midrange speaker.

Rotating Electric Machines

The range of sizes and power ratings and the different physical features of rotating machines are such that the task of explaining the operation of rotating machines in a single chapter may appear formidable at first. Some features of rotating machines, however, are common to all such devices. This introductory section is aimed at explaining the common properties of all rotating electric machines. We begin our discussion with reference to Figure 5.12.24, in which a hypothetical rotating machine is depicted in a cross-sectional view. In the figure, a box with a cross inscribed in it indicates current flowing into the page, while a dot represents current out of the plane of the page.

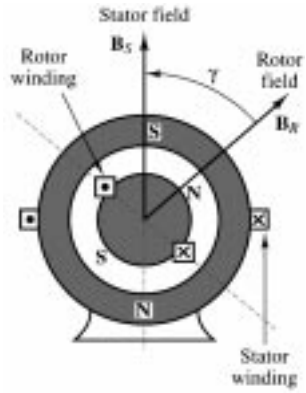


FIGURE 5.12.24 A rotating electric machine.

In Figure 5.12.24, we identify a **stator**, of cylindrical shape, and a **rotor**, which, as the name indicates, rotates inside the stator, separated from the latter by means of an air gap. The rotor stator each consist of a magnetic core, some electrical insulation, and the windings necessary to establish a magnetic flux (unless this is created by a permanent magnet). The rotor is mounted on a bearing-supported shaft, which can be connected to *mechanical loads* (if the machine is a motor) or to a *prime mover* (if the machine is a generator) by means of belts, pulleys, chains, or other mechanical couplings. The windings carry the electric currents that generate the magnetic fields and flow to the electrical loads, and also provide the closed loops in which voltages will be induced (by virtue of Faraday's law, as discussed in the previous section).

Basic Classification of Electric Machines

An immediate distinction can be made between different types of windings characterized by the nature of the current they carry. If the current serves the sole purpose of providing a magnetic field and is independent of the load, it is called a *magnetizing*, or excitation, current, and the winding is termed a **field winding**. Field currents are nearly always DC and are of relatively low power, since their only purpose is to magnetize the core (recall the important role of high-permeability cores in generating large magnetic fluxes from relatively small currents). On the other hand, if the winding carries only the load current, it is called an **armature**. In DC and AC synchronous machines, separate windings exist to carry field and armature currents. In the induction motor, the magnetizing and load currents flow in the same winding, called the *input winding*, or *primary*; the output winding is then called the *secondary*. As we shall see, this terminology, which is reminiscent of transformers, is particularly appropriate for induction motors, which bear a significant analogy to the operation of the transformers. Table 5.12.4 characterizes the principal machines in terms of their field and armature configuration.

TABLE 5.12.4 Configurations of the Three Types of Electric Machines

Machine Type	Winding	Winding Type	Location	Current
DC	Input and output	Armature	Rotor	AC (winding) DC (at brushes)
	Magnetizing	Field	Stator	DC
Synchronous	Input and output	Armature	Stator	AC
	Magnetizing	Field	Rotor	DC
Induction	Input	Primary	Stator	AC
	Output	Secondary	Rotor	AC

It is also useful to classify electric machines in terms of their energy-conversion characteristics. A machine acts as a **generator** if it converts mechanical energy from a prime mover — e.g., an internal combustion engine — to electrical form. Examples of generators are the large machines used in power-generating plants, or the common automotive alternator. A machine is classified as a **motor** if it converts electrical energy to mechanical form. The latter class of machines is probably of more direct interest to you, because of its widespread application in engineering practice. Electric motors are used to provide forces and torques to generate motion in countless industrial applications. Machine tools, robots, punches, presses, mills, and propulsion systems for electric vehicles are but a few examples of the application of electric machines in engineering.

Note that in Figure 5.12.24 we have explicitly shown the direction of two magnetic fields: that of the rotor, $\mathbf{B}_R$, and that of the stator, $\mathbf{B}_S$. Although these fields are generated by different means in differential machines (e.g., permanent magnets, AC currents, DC currents), the presence of these fields is what causes a rotating machine to turn and enables the generation of electric power. In particular, we see that in Figure 5.12.24 the north pole of the rotor field will seek to align itself with the south pole of the stator field. It is this magnetic attraction force that permits the generation of torque in an electric motor; conversely, a generator exploits the laws of electromagnetic induction to convert a changing magnetic field to an electric current.

To simplify the discussion in later sections, we shall presently introduce some basic concepts that apply to all rotating electric machines. Referring to Figure 5.12.25, we note that all machines the force on a wire is given by the expression

$$\mathbf{f} = i_w \mathbf{I} \times \mathbf{B} \tag{5.12.35}$$

where i_w is the current in the wire, $\mathbf{I}$ is a vector along the direction of the wire, and × denotes the cross product of two vectors. Then the torque for a multiturn coil becomes

$$T = KBi_w \sin\alpha \tag{5.12.36}$$

where

B = magnetic flux density caused by the stator field
K = constant depending on coil geometry
α = angle between **B** and the normal to the plane of the coil

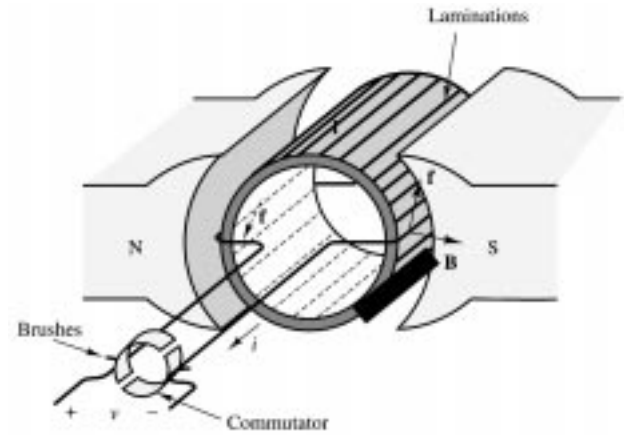


FIGURE 5.12.25 Stator and rotor fields and the force acting on a rotating machine.

Other Characteristics of Electric Machines

As already stated earlier in this chapter, electric machines are **energy-conversion devices**, and we are therefore interested in their energy-conversion **efficiency**. Typical applications of electric machines as motors or generators must take into consideration the energy losses associated with these devices. Figure 5.12.26 represents the various loss mechanisms you must consider in analyzing the efficiency of an electric machine for the case of direct-current machines. It is important for you to keep in mind this conceptual flow of energy when analyzing electric machines. The sources of loss in a rotating machine can be separated into three fundamental groups: electrical (I^2R) losses, core losses, and mechanical losses.

I^2R losses are usually computed on the basis of the DC resistance of the windings at 75°C; in practice, these losses vary with operating conditions. The difference between the nominal and actual I^2R is usually lumped under the category of *stray-load loss*. In direct-current machines, it is also necessary to account for the *brush contact loss* associated with slip rings and commutators.

Mechanical losses are due to *friction* (mostly in the bearings) and *windage,* that is, the air drag force that opposes the motion of the rotor. In addition, if external devices (e.g., blowers) are required to circulate air through the machine for cooling purposes, the energy expended by these devices is also included in the mechanical losses.

Open-circuit core losses consist of hysteresis and eddy current losses, with only the excitation winding energized. Often these losses are summed with friction and windage losses to give rise to the *no-load rotational loss*. The latter quantity is useful if one simply wishes to compute efficiency. Since open-circuit core losses do not account for the changes in flux density caused by the presence of load currents, an additional magnetic loss is incurred that is not accounted for in this term. *Stray-load losses* are used to lump the effects of nonideal current distribution in the windings and of the additional core losses just mentioned. Stray-load losses are difficult to determine exactly and are often assumed to be equal to 1.0% of the output power for DC machines; these losses can be determined by experiment in synchronous and induction machines.

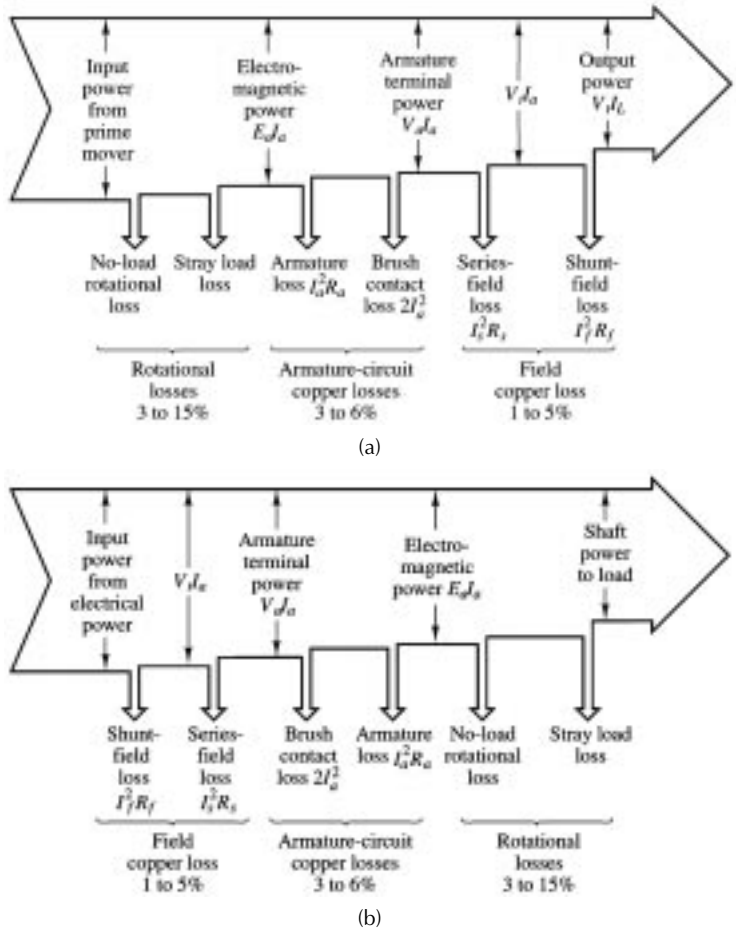


FIGURE 5.12.26 (a) Generator losses, direct current; (b) motor losses, direct current.

The performance of an electric machine can be quantified in a number of ways. In the case of an electric motor, it is usually portrayed in the form of a graphical **torque-speed characteristic.** The torque-speed characteristic of a motor describes how the torque supplied by the machine varies as a function of the speed of rotation of the motor for steady speeds. As we shall see in later sections, the torque-speed curves vary in shape with the type of motor (DC, induction, synchronous) and are very useful in determining the performance of the motor when connected to a mechanical load. Figure 5.12.27 depicts the torque-speed curve of a hypothetical motor. We shall presently describe the essential elements of such a graphical representation of motor performance, and we shall later return to analyze the typical performance curve of each type of motor we encounter in our discussion. It is quite likely that in most engineering applications, the engineer is required to make a decision regarding the performance characteristics of the motor best suited to a specified task. In this context, the torque-speed curve of a machine is a very useful piece of information.

The first feature we note of the torque-speed characteristic is that it bears a strong resemblance to the *i-v* characteristics used in earlier chapters to represent the behavior of electrical sources. It should be clear that, according to this torque-speed curve, the motor is not an ideal source of torque (if it were, the curve would appear as a horizontal line across the speed range). One can readily see, for example, that the hypothetical motor represented by the curves of Figure 5.12.27 would produce maximum torque in the range of speeds between approximately 800 and 1400 rev/min. What determines the actual speed of the motor (and therefore its output torque and power) is the torque-speed characteristic of the load

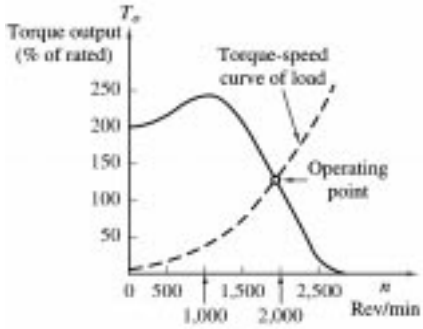


FIGURE 5.12.27 Torque-speed curve for an electric motor.

connected to it, much as a resistive load determines the current drawn from a voltage source. In the figure, we display the torque-speed curve of a load, represented by the dashed line; the operating point of the motor-load pair is determined by the intersection of the two curves.

Another important observation pertains to the fact that the motor of Figure 5.12.27 produces a nonzero torque at zero speed. This fact implies that as soon as electric power is connected to the motor, the latter is capable of supplying a certain amount of torque; this zero-speed torque is called the **starting torque**. If the load the motor is connected to requires less than the starting torque the motor can provide, then the motor can accelerate the load, until the motor speed and torque settle to a stable value, at the operating point. The motor-load pair of Figure 5.12.27 would behave in the manner just described. However, there may well be circumstances in which a motor might not be able to provide a sufficient starting torque to overcome the static load torque that opposes its motion. Thus, we see that a torque-speed characteristic can offer valuable insight into the operation of a motor. As we proceed to discuss each type of machine in greater detail, we shall devote some time to the discussion of its torque-speed curve.

The most common means of conveying information regarding electric machines is the *nameplate*. Typical information conveyed by the nameplate is

1. Type of device (e.g., DC motor, alternator)
2. Manufacturer
3. Rated voltage and frequency
4. Rated current and volt-amperes
5. Rated speed and horsepower

The **rated voltage** is the terminal voltage for which the machine was designed, and which will provide the desired magnetic flux. Operation at higher voltages will increase magnetic core losses, because of excessive core saturation. The **rated current** and **rated volt-amperes** are an indication of the typical current and power levels at the terminal that will not cause undue overheating due to copper losses (I^2R losses) in the windings. These ratings are not absolutely precise, but they give an indication of the range of excitations for which the motor will perform without overheating. Peak power operation in a motor may exceed rated torque (horsepower) or currents by a substantial factor (up to as much as six or seven times the rated value); however, continuous operation of the motor above the rated performance will cause the machine to overheat, and possibly to sustain damage. Thus, it is important to consider both peak and continuous power requirements when selecting a motor for a specific application. An analogous discussion is valid for the speed rating: while an electric machine may operate above rated speed for limited periods of time, the large centrifugal forces generated at high rotational speeds will eventually cause undesirable mechanical stresses, especially in the rotor windings, leading eventually even to self-destruction.

Another important feature of electric machines is the **regulation** of the machine speed or voltage, depending on whether it is used as a motor or as a generator, respectively. Regulation is the ability to maintain speed or voltage constant in the face of load variations. The ability to closely regulate speed in a motor or voltage in a generator is an important feature of electric machines; regulation is often

improved by means of feedback control mechanisms, some of which will be briefly introduced in this chapter. We shall take the following definitions as being adequate for the intended purpose of this chapter.

$$\text{Speed regulation} = \frac{\text{Speed at no load} - \text{Speed at rated load}}{\text{Speed at rated load}} \tag{5.12.37}$$

$$\text{Voltage regulation} = \frac{\text{Voltage at no load} - \text{Voltage at rated load}}{\text{Voltage at rated load}} \tag{5.12.38}$$

Please note that the rated value is usually taken to be the nameplate value, and that the meaning of *load* changes depending on whether the machine is a motor, in which case the load is mechanical, or a generator, in which case the load is electrical.

Basic Operation of All Rotating Machines

We have already seen how the magnetic field in electromechanical devices provides a form of coupling between electrical and mechanical systems. Intuitively, one can identify two aspects of this coupling, both of which play a role in the operation of electric machines: (1) magnetic attraction and repulsion forces generate mechanical torque, and (2) the magnetic field can induce a voltage in the machine windings (coils) by virtue of Faraday's law. Thus, we may think of the operation of an electric machine in terms of either a motor or a generator, depending on whether the input power is electrical and mechanical power is produced (motor action), or the input power is mechanical and the output power is electrical (generator action). Figure 5.12.28 illustrates the two cases graphically.

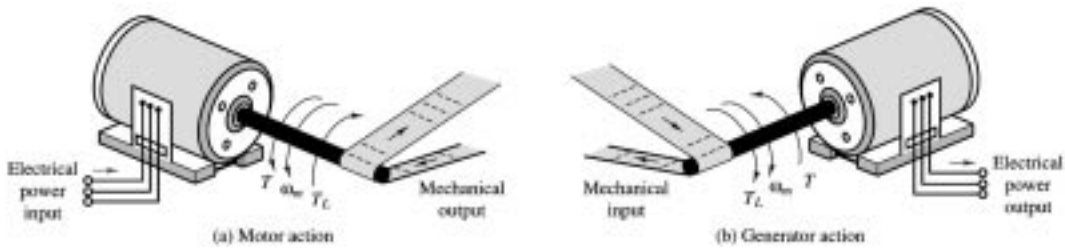


FIGURE 5.12.28 Generator and motor action in an electric machine.

The coupling magnetic field performs a dual role, which may be explained as follows. When a current i flow through conductors placed in a magnetic field, a force is produced on each conductor, according to Equation (5.12.35). If these conductors are attached to a cylindrical structure, a torque is generated, and if the structure is free to rotate, then it will rotate at an angular velocity ω_m. As the conductors rotate, however, they move through a magnetic field and cut through flux lines, thus generating an electromotive force in opposition to the excitation. This emf is also called "counter" emf; it opposes the source of the current i. If, on the other hand, the rotating element of the machine is driven by a prime mover (for example, an internal combustion engine), then an emf is generated across the coil that is rotating in the magnetic field (the armature). If a load is connected to the armature, a current i will flow to the load, and this current flow will in turn cause a reaction torque on the armature that opposes the torque imposed by the prime mover.

You see, then, that for energy conversion to take place, two elements are required: (1) a coupling field, **B**, usually generated in the field winding; and (2) an armature winding that supports the load current, i, and the emf, e.

Magnetic Poles in Electric Machines

Before discussing the actual construction of a rotating machine, we should spend a few paragraphs to illustrate the significance of **magnetic poles** in an electric machine. In an electric machine, torque is

developed as a consequence of magnetic forces of attraction and repulsion between magnetic poles on the stator and on the rotor; these poles produce a torque that accelerates the rotor and a reaction torque on the stator. Naturally, we would like a construction such that the torque generated as a consequence of the magnetic forces is continuous and in a constant direction. This can be accomplished if the number of rotor poles is equal to the number of stator poles. It is also important to observe that the number of poles must be even, since there have to be equal numbers of north and south poles.

Figure 5.12.29 depicts a two-pole machine in which the stator poles are constructed in such a way as to project closer to the rotor than to the stator structure. This type of construction is rather common, and poles constructed in this fashion are called **salient poles**. Note that the rotor could also be constructed to have salient poles.

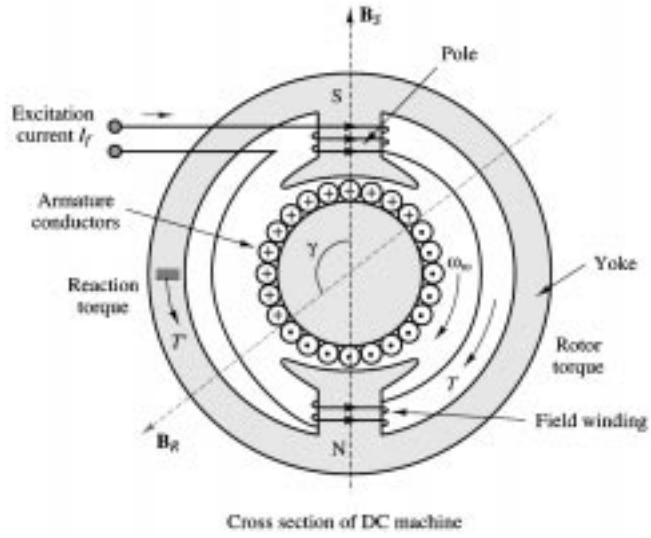


FIGURE 5.12.29 A two-pole machine with salient stator poles.

To understand magnetic polarity, we need to consider the direction of the magnetic field in a coil carrying current. Figure 5.12.30 shows how the *right-hand rule* can be employed to determine the direction of the magnetic flux. If one were to grasp the coil with the right hand, with the fingers curling in the direction of current flow, then the thumb would be pointing in the direction of the magnetic flux. Magnetic flux is by convention viewed as entering the south pole and exiting from the north pole. Thus, to determine whether a magnetic pole is north or south, we must consider the direction of the flux. Figure 5.12.31 shows a cross section of a coil wound around a pair of salient rotor poles. In this case, one can readily identify the direction of the magnetic flux and therefore the magnetic polarity of the poles by applying the right-hand rule, as illustrated in the figure.

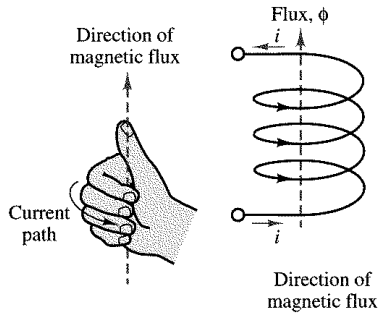


FIGURE 5.12.30 Right-hand rule.

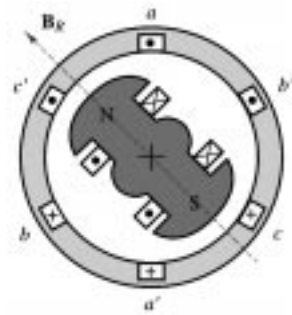


FIGURE 5.12.31 Magnetic field in a salient rotor winding.

Often, however, the coil windings are not arranged as simply as in the case of salient poles. In many machines, the windings are embedded in slots cut into the stator or rotor, so that the situation is similar to that of the stator depicted in Figure 5.12.32. This figure is a cross section in which the wire connections between "crosses" and "dots" have been cut away. In Figure 5.12.32, the dashed line indicates the axis of the stator flux according to the right-hand rule, indicating that the slotted stator in effect behaves like a pole pair. The north and south poles indicated in the figure are a consequence of the fact that the flux exits the bottom part of the structure (thus, the north pole indicated in the figure) and enters the top half of the structure (thus, the south pole). In particular, if you consider that the windings are arranged so that the current entering the right-hand side of the stator (to the right of the dashed line) flows through the back end of the stator and then flows outward from the left-hand side of the stator slots (left of the dashed line), you can visualize the windings in the slots as behaving in a manner similar to the coils of Figure 5.12.31, where the flux axis of Figure 5.12.32 corresponds to the flux axis of each of the coils of Figure 5.12.31. The actual circuit that permits current flow is completed by the front and back ends of the stator, where the wires are connected according to the pattern *a*-*a*′, *b*-*b*′, *c*-*c*′, as depicted in the figure.

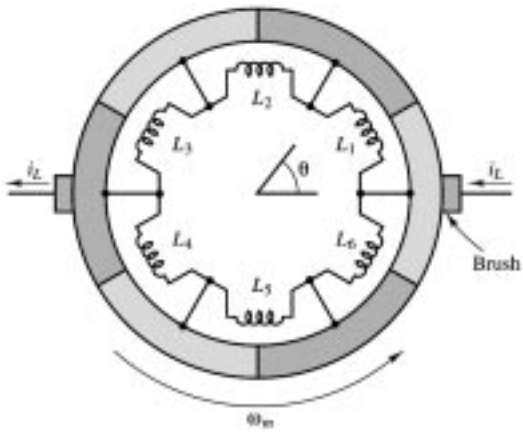


FIGURE 5.12.32 Magnetic field of stator.

Another important consideration that facilitates understanding the operation of electric machines pertains to the use of AC currents. It should be apparent by now that if the current flowing into the slotted stator is alternating, the direction of the flux will also alternate, so that in effect the two poles will reverse polarity every time the current reverses direction, that is, every half-cycle of the sinusoidal current. Further — since the magnetic flux is approximately proportional to the current in the coil — as the amplitude of the current oscillates in a sinusoidal fashion, so will the flux density in the structure. Thus, *the magnetic field developed in the stator changes both spatially and in time.*

This property is typical of AC machines, where a *rotating magnetic field* is established by energizing the coil with an alternating current. As we shall see in the next section, the principles underlying the operation of DC and AC machines are quite different: in a direct-current machine, there is no rotating

field, but a mechanical switching arrangement (the *commutator*) makes it possible for the rotor and stator magnetic fields to always align at right angles to each other.

Direct-Current Machines

As explained in the introductory section, direct-current (DC) machines are easier to analyze than their AC counterparts, although their actual construction is made rather complex by the need to have a commutator, which reverses the direction of currents and fluxes to produce a net torque. The objective of this section is to describe the major construction features and the operation of direct-current machines, as well as to develop simple circuit models that are useful in analyzing the performance of this class of machines.

Physical Structure of DC Machines

A representative DC machine was depicted in Figure 5.12.29, with the magnetic poles clearly identified for both the stator and the rotor. Note the salient pole construction of the stator and the slotted rotor. As previously stated, the torque developed by the machine is a consequence of the magnetic forces between stator and rotor poles. This torque is maximum when the angle γ between the rotor and stator poles is 90°. Also, as you can see from the figure, in a DC machine the armature is usually on the rotor, and the field winding is on the stator.

To keep this torque angle constant as the rotor spins on its shaft, a mechanical switch, called a **commutator**, is configured so the current distribution in the rotor winding remains constant and therefore the rotor poles are consistently at 90° with respect to the fixed stator poles. In a DC machine, the magnetizing current is DC, so that there is no spatial alternation of the stator poles due to time-varying currents. To understand the operation of the commutator, consider the simplified diagram of Figure 5.12.33. In the figure, the brushes are fixed, and the rotor revolves at an angular velocity ω_m; the instantaneous position of the rotor is given by the expression $\theta = \omega_m t - \gamma$.

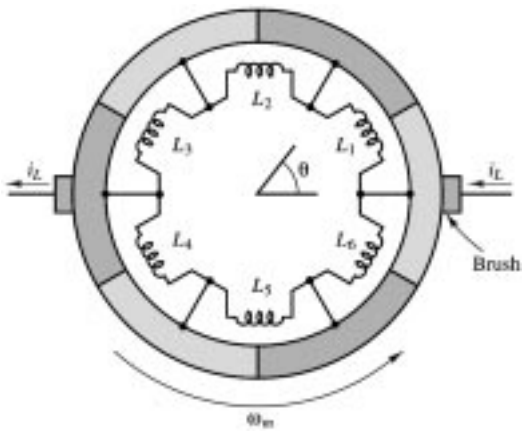


FIGURE 5.12.33 Rotor winding and commutator.

The commutator is fixed to the rotor and is made up in this example of six segments that are made of electrically conducting material but are insulated from each other. Further, the rotor windings are configured so that they form six coils, connected to the commutator segments as shown in Figure 5.12.33.

As the commutator rotates counterclockwise, the rotor magnetic field rotates with it up to $\theta = 30°$. At that point, the direction of the current changes in coils L_3 and L_6 as the brushes make contact with the next segment. Now the direction of the magnetic field is –30°. As the commutator continues to rotate, the direction of the rotor field will again change from –30° to +30°, and it will switch again when the brushes switch to the next pair of segments. In this machine, then, the torque angle, γ, is not always 90°, but can vary by as much as ±30°; the actual torque produced by the machine would fluctuate by as much as ±14%, since the torque is proportional to sin γ. As the number of segments increases, the

torque fluctuation produced by the commutation is greatly reduced. In a practical machine, for example, one might have as many as 60 segments, and the variation of γ from 90° would be only ±3°, with a torque fluctuation of less than 1%. Thus, the DC machine can produce a nearly constant torque (as a motor) or voltage (as a generator).

Configuration of DC Machines

In DC machines, the field excitation that provides the magnetizing current is occasionally provided by an external source, in which case the machine is said to be **separately excited** (Figure 5.12.34(a)). More often, the field excitation is derived from the armature voltage and the machine is said to be **self-excited**. The latter configuration does not require the use of a separate source for the field excitation and is therefore frequently preferred. If a machine is in the separately excited configuration, an additional source, V_f, is required. In the self-excited case, one method used to provide the field excitation is to connect the field in parallel with the armature, since the field winding typically has significantly higher resistance than the armature circuit (remember that it is the armature that carries the load current), this will not draw excessive current from the armature. Further, a series resistor can be added to the field circuit to provide the means for adjusting the field current independent of the armature voltage. This configuration is called a **shunt-connected** machine and is depicted in Figure 5.12.34(b). Another method for self-exciting a DC machine consists of connecting the field in series with the armature, leading to the **series-connected** machine, depicted in Figure 5.12.34(c); in this case, the field winding will support the entire armature current, and thus the field coil must have low resistance (and therefore relatively few turns). This configuration is rarely used for generators, since the generated voltage and the load voltage must always differ by the voltage drop across the field coil, which varies with the load current. Thus, a series generator would have poor (large) regulation. However, series-connected motors are commonly used in certain applications, as will be discussed in a later section.

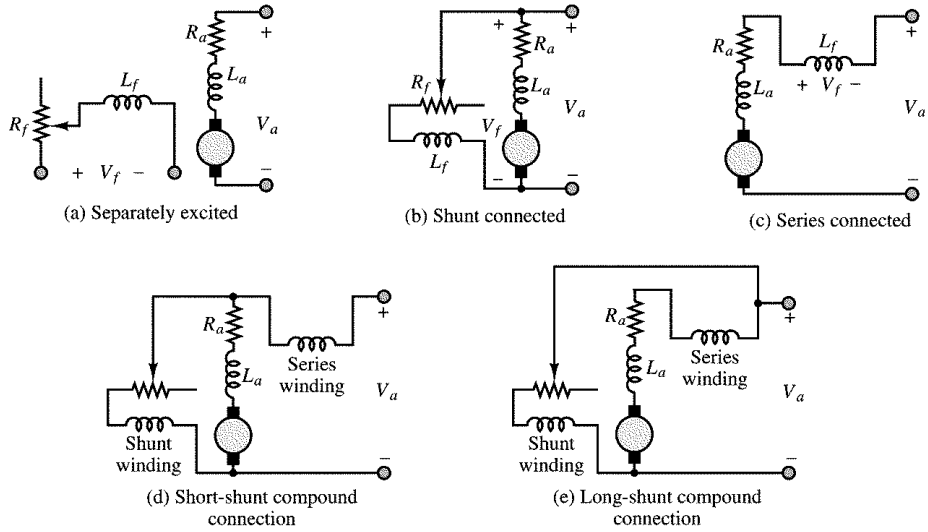


FIGURE 5.12.34

The third type of DC machine is the **compound-connected** machine, which consists of a combination of the shunt and series configurations. Figures 5.12.34(d) and (e) show the two types of connections, called the **short shunt** and the **long shunt**, respectively. Each of these configurations may be connected so that the series part of the field adds to the shunt part (**cumulative compounding**) or so that it subtracts (**differential compounding**).

DC Machine Models

As stated earlier, it is relatively easy to develop a simple model of a DC machine, which is well suited to performance analysis, without the need to resort to the details of the construction of the machine itself. This section will illustrate the development of such models in two steps. First, steady-state models relating field and armature currents and voltages to speed and torque are introduced; second, the differential equations describing the dynamic behavior or DC machines are derived.

When a field excitation is established, a magnetic flux, ϕ, is generated by the field current, I_f. From Equation (5.12.36) we know that the torque acting on the rotor is proportional to the product of the magnetic field and the current in the load-carrying wire; the latter current is the armature current, I_a (i_w, in Equation 5.12.36). Assuming that, by virtue of the cummutator, the torque angle, γ, is kept very close to 90°, and therefore $\sin\gamma = 1$, we obtain the following expression for the torque (in units of N-m) in a DC machine:

$$T = k_T \phi I_a \qquad \text{for } \gamma = 90° \tag{5.12.39}$$

You may recall that this is simply a consequence of the *Bli* law of Chapter 15. The mechanical power generated (or absorbed) is equal to the product of the machine torque and the mechanical speed of rotation, ω_m (in rad/sec), and is therefore given by

$$P_m = \omega_m T = \omega_m k_T \phi I_a \tag{5.12.40}$$

Recall now that the rotation of the armature conductors in the field generated by the field excitation causes a **back emf**, E_b, in a direction that opposes the rotation of the armature. According to the *Blu* law then, this back emf is given by the expression.

$$E_b = k_a \phi \omega_m \tag{5.12.41}$$

where k_a is called the **armature constant** and is related to the geometry and magnetic properties of the structure. The voltage E_b represents a countervoltage (opposing the DC excitation) in the case of a motor, and the generated voltage in the case of a generator. Thus, the electric power dissipated (or generated) by the machine is given by the product of the back emf and the armature current:

$$P_e = E_b I_a \tag{5.12.42}$$

The constants k_T and k_a in Equations (5.12.39) and (5.12.41) are related to geometry factors, such as the dimension of the rotor and the number of turns in the armature winding; and to properties of materials, such as the permeability of the magnetic materials. Note that in the ideal energy-conversion case, $P_m = P_e$, and therefore $k_a = k_T$. We shall, in general, assume such ideal conversion of electrical to mechanical (or vice versa) and will therefore treat the two constants as being identical: $k_a = k_T$. The constant k_a is given by

$$k_a = \frac{pN}{2\pi M} \tag{5.12.43}$$

where

p = number of magnetic poles
N = number of conductors per coil
M = number of parallel paths in armature winding

An important observation concerning the units of angular speed must be made at this point. The equality (under the no-loss assumption) between the constants k_a and k_T in Equations (5.12.39) and

(5.12.41) results from the choice of consistent units, namely, volts and amperes for the electrical quantities, and newton-meters and radians per second for the mechanical quantities. You should be aware that it is fairly common practice to refer to the speed of rotation of an electric machine in units of revolutions per minute (rev/min). In this book, we shall uniformly use the symbol n to denote angular speed in rev/min; the following relationship should be committed to memory:

$$n(\text{rev/min}) = \frac{60}{2\pi}\omega_m \ (\text{rad/sec}) \tag{5.12.44}$$

If the speed is expressed in rev/min, the armature constant changes as follows:

$$E_b = k'_a \phi n \tag{5.12.45}$$

where

$$k'_a = \frac{pN}{60M} \tag{5.12.46}$$

Having introduced the basic equations relating torque, speed, voltages, and currents in electric machines, we may now consider the interaction of these quantities in a DC machine at steady state, that is, operating at constant speed and field excitation. Figure 5.12.35 depicts the electrical circuit model of a separately excited DC machine, illustrating both motor and generator action. It is very important to note the reference direction of armature current flow and of the developed torque, in order to make a distinction between the two modes of operation. The field excitation is shown as a voltage, V_f, generating the field current, I_f, that flows through a variable resistor, R_f, and through the field coil, L_f. The variable resistor permits adjustment of the field excitation. The armature circuit, on the other hand, consists of a voltage source representing the back emf, E_b, the armature resistance, R_a, and the armature voltage, V_a. This model is appropriate both for motor and for generator action. When $V_a < E_b$, the machine acts as a generator (I_a flows out of the machine). When $V_a > E_b$, the machine acts as a motor (I_a flows into the machine). Thus, according to the circuit model of Figure 5.12.35, the operation of a DC machine at steady state (i.e., with the inductors in the circuit replaced by short circuits) is described by the following equations:

$$\begin{aligned} I_f &= \frac{V_f}{R_f} \quad \text{and} \quad V_a = R_a I_a + E_b \qquad (\text{motor action}) \\ I_f &= \frac{V_f}{R_f} \quad \text{and} \quad V_a = -R_a I_a + E_b \qquad (\text{generator action}) \end{aligned} \tag{5.12.47}$$

Equation pair (5.12.47) together with Equations (5.12.39) and (5.12.41) may be used to determine the steady-state operating condition of a DC machine.

The circuit model of Figure 5.12.35 permits the derivation of a simple set of differential equations that describe the *dynamic* analysis of a DC machine. The dynamic equations describing the behavior of a separately excited DC machine are as follows:

$$V_a(t) = I_a(t)R_a + L_a \frac{dI_a(t)}{dt} + E_b(t) \qquad (\text{armature circuit}) \tag{5.12.48a}$$

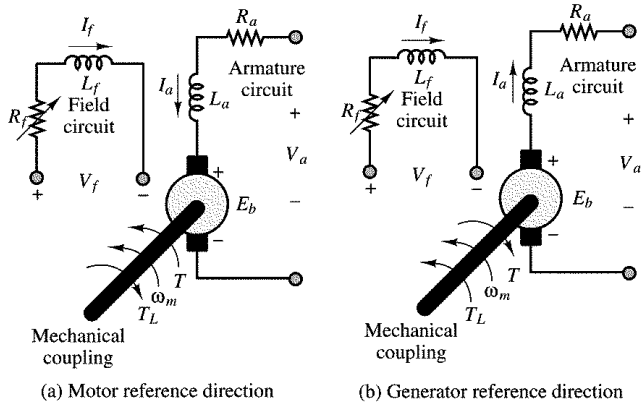


FIGURE 5.12.35 Electrical circuit model of a separately excited DC machine.

$$V_f(t) = I_f(t)R_f + L_f\frac{dI_f(t)}{dt} \qquad \text{(field circuit)} \tag{5.12.48b}$$

These equations can be related to the operation of the machine in the presence of a load. If we assume that the motor is rigidly connected to an inertial load with moment of inertia J and that the friction losses in the load are represented by a viscous friction coefficient, b, then the torque developed by the machine (in the motor mode of operation) can be written as follows:

$$T(t) = T_L + b\omega_m(t) + J\frac{d\omega_m(t)}{dt} \tag{5.12.49}$$

where T_L is the load torque. T_L is typically either constant or some function of speed, ω_m, in a motor. In the case of a generator, the load torque is replaced by the torque supplied by a prime mover, and the machine torque, $T(t)$, opposes the motion of the prime mover, as shown in Figure 5.12.35. Since the machine torque is related to the armature and field currents by Equation (5.12.39), Equations (5.12.48) and (5.12.49) are coupled to each other; this coupling may be expressed as follows:

$$T(t) = k_a\phi I_a(t) \tag{5.12.50}$$

or

$$k_a\phi I_a(t) = T_L + b\omega_m(t) + J\frac{d\omega_m(t)}{dt} \tag{5.12.51}$$

The dynamic equations described in this section apply to any DC machine. In the case of a *separately excited* machine, a further simplification is possible, since the flux is established by virtue of a separate field excitation, and therefore

$$\phi = \frac{N_f}{\mathcal{R}}I_f = k_f I_f \tag{5.12.52}$$

where N_f is the number of turns in the field coil, $\mathcal{R}$ is the reluctance of the structure, and I_f is the field current.

AC Machines

From the previous sections, it should be apparent that it is possible to obtain a wide range of performance characteristics from DC machines, as both motors and generators. A logical question at this point should be, would it not be more convenient in some cases to take advantage of the single- or multiphase AC power that is available virtually everywhere than to expend energy and use additional hardware to rectify and regulate the DC supplies required by direct-current motors? The answer to this very obvious question is certainly a resounding yes. In fact, the AC induction motor is the workhorse of many industrial applications, and synchronous generators are used almost exclusively for the generation of electric power worldwide. Thus, it is appropriate to devote a significant portion of this chapter to the study of AC machines, and of induction motors in particular. The objective if this section is to explain the basic operation of both synchronous and induction machines, and to outline their performance characteristics. In doing so, we shall also point out the relative advantages and disadvantages of these machines in comparison with direct-current machines.

Rotating Magnetic Fields

As mentioned in earlier, the fundamental principle of operation o AC machines is the generation of a rotating magnetic field, which causes the rotor to turn at a speed that depends on the speed of rotation of the magnetic field. We shall now explain how a rotating magnetic field can be generated in the stator and air gap of an AC machine by means of AC currents.

Consider the stator shown in Figure 5.12.36, which supports windings, a-a', b-b', and c-c'. The coils are geometrically spaced 120° apart, and a three-phase voltage is applied to the coils. As you may recall from the discussion of AC power in Section 5.5, the currents generated by a three-phase source are also spaced by 120°, as illustrated in Figure 5.12.37. The phase voltages referencing the neutral terminal would then be given by the expressions

$$v_a = A\cos(\omega_e t)$$

$$v_b = A\cos\left(\omega_e t - \frac{2\pi}{3}\right)$$

$$v_c = A\cos\left(\omega_e t + \frac{2\pi}{3}\right)$$

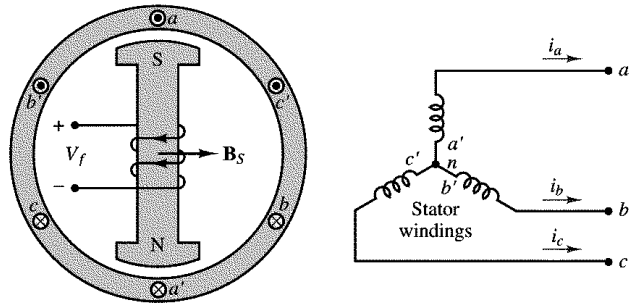


FIGURE 5.12.36 Two-pole three-phase stator.

where ω_e is the frequency of the AC supply, or line frequency. The coils in each winding are arranged in such a way that the flux distribution generated by any one winding is approximately sinusoidal. Such a flux distribution may be obtained by appropriately arranging groups of coils for each winding over the stator surface. Since the coils are spaced 120° apart, the flux distribution resulting from the sum of the contributions of the three windings is the sum of the fluxes due to the separate windings, as shown

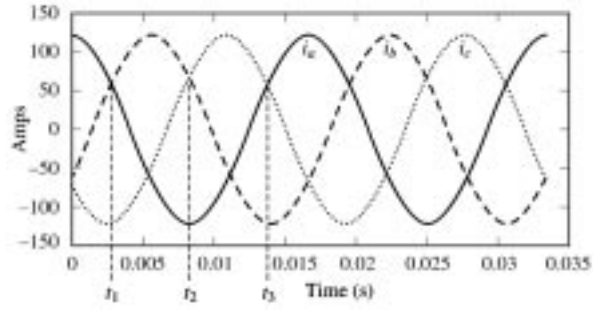


FIGURE 5.12.37 Three-phase stator winding currents.

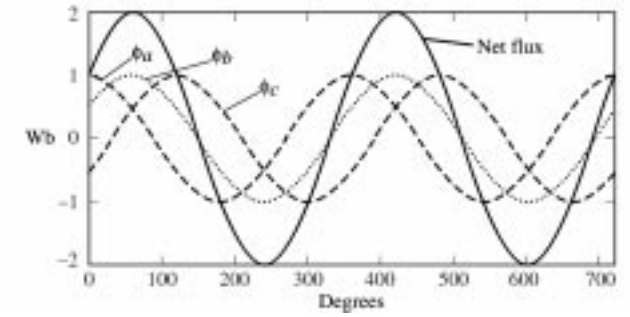


FIGURE 5.12.38 Flux distribution in a three-phase stator winding as a function of angle of rotation.

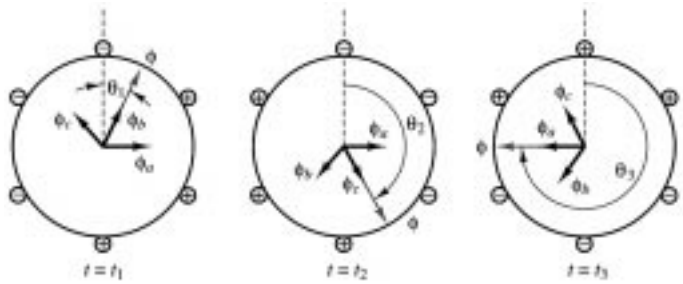


FIGURE 5.12.39 Rotating flux in a three-phase machine.

in Figure 5.12.38. Thus, the flux in a three-phase machine rotates in space according to the vector diagram of Figure 5.12.39, and is constant in amplitude. A stationary observer on the machine's stator would see a sinusoidally varying flux distribution as shown in Figure 5.12.38.

Since the resultant flux of Figure 5.12.38 is generated by the currents of Figure 5.12.37, the speed of rotation of the flux must be related to the frequency of the sinusoidal phase currents. In the case of the stator of Figure 5.12.36, the number of magnetic poles resulting from the winding configuration is two; however, it is also possible to configure the windings so that they have more poles. For example, Figure 5.12.40 depicts a simplified view of a four-pole stator.

In general, the speed of the rotating magnetic field is determined by the frequency of the excitation current, f, and by the number of poles present in the stator, p, according to the equation

$$n_s = \frac{120f}{p} \text{ rev/min}$$

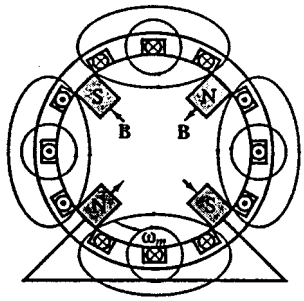


FIGURE 5.12.40 Four-pole stator.

or

$$\omega_s = \frac{2\pi n_s}{60} = \frac{2\pi \times 2f}{p} \tag{5.12.53}$$

where n_s (or ω_s) is usually called the **synchronous speed**.

The Induction Motor

The induction motor is the most widely used electric machine because of its relative simplicity of construction. The stator winding of an induction machine is similar to that of a synchronous machine; thus, the description of the three-phase winding of Figure 5.12.36 also applies to induction machines. The primary advantage of the induction machine, which is almost exclusively used as a motor (its performance as a generator is not very good), is that no separate excitation is required for the rotor. The rotor typically consists of one of two arrangements: **a squirrel cage** or a **wound rotor**. The former contains conducting bars short-circuited at the end and embedded within it; the latter consists of a multiphase winding similar to that used for the stator, but electrically short-circuited.

In either case, the induction motor operates by virtue of currents induced from the stator field in the rotor. In this respect, its operation is similar to that of a transformer, in that currents in the stator (which acts as a primary coil) induce currents in the rotor (acting as a secondary coil). In most induction motors, no external electrical connection is required for the rotor, thus permitting a simple, rugged construction, without the need for slip rings or brushes. Unlike the synchronous motor, the induction motor does not operate at synchronous speed, but at a somewhat lower speed, which is dependent on the load. Figure 5.12.41 illustrates the appearance of a squirrel-cage induction motor. The following discussion will focus mainly on this very common configuration.

You are by now acquainted with the notion of a rotating stator magnetic field. Imagine now that a squirrel-cage rotor is inserted in a stator in which such a rotating magnetic field is present. The stator field will induce voltages in the cage conductors, and if the stator field is generated by a three-phase source, the resulting rotor currents — which circulate in the bars of the squirrel cage, with the conducting path completed by the shorting rings at the end of the cage — are also three-phase, and are determined by the magnitude of the induced voltages and by the impedance of the rotor. Since the rotor currents are induced by the stator field, the number of poles and the speed of rotation of the induced magnetic field are the same as those of the stator field, *if the rotor is at rest*. Thus, when a stator field is initially applied, the rotor field is synchronous with it, and the fields are stationary with respect to each other. Thus, according to the earlier discussion, a *starting torque* is generated.

If the starting torque is sufficient to cause the rotor to start spinning, the rotor will accelerate up to its operating speed. However, an induction motor can never reach synchronous speed; if it did, the rotor would appear to be stationary with respect to the rotating stator field, since it would be rotating at the same speed. But in the absence of relative motion between the stator and rotor fields, no voltage would be induced in the rotor. Thus, an induction motor is limited to speeds somewhere below the synchronous

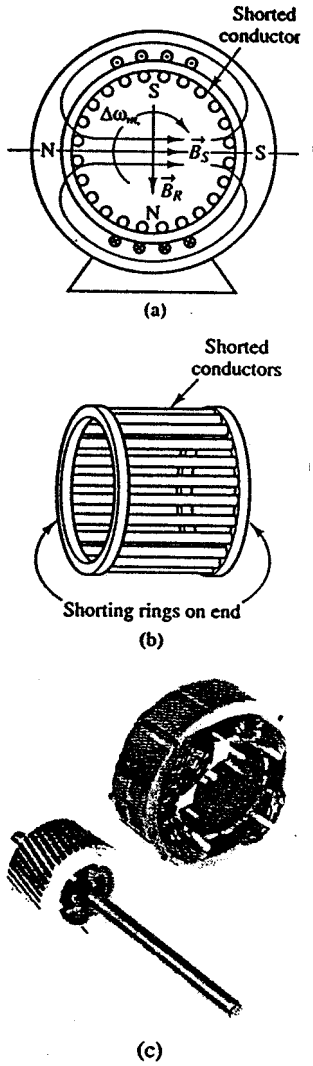


FIGURE 5.12.41 (a) Squirrel-cage induction motor; (b) conductors in rotor; (c) photo of squirrel-cage induction motor.

speed, n_s. Let the speed of rotation of the rotor be n; then, the rotor is losing ground with respect to the rotation of the stator field at a speed $(n_s - n)$. In effect, this is equivalent to backward motion of the rotor at the **slip speed**, defined by $(n_s - n)$. The **slip**, s, is usually defined as a fraction of n_s:

$$s = \frac{n_s - n}{n_s} \tag{5.12.54}$$

which leads to the following expression for the motor speed:

$$n = n_s(1 - s) \tag{5.12.55}$$

The slip, s, is a function of the load, and the amount of slip in a given motor is dependent on its construction and rotor type (squirrel cage or wound rotor). Since there is a relative motion between the stator and rotor fields, voltages will be induced in the rotor at a frequency called the **slip frequency**, related to the relative speed of the two fields. This gives rise to an interesting phenomenon: the rotor

field travels relative to the rotor at the slip speed sn_s, but the rotor is mechanically traveling at the speed $(1 - s)n_s$, so that the net effect is that the rotor field travels at the speed

$$sn_s + (1 - s)n_s = n_s \tag{5.12.56}$$

that is, at synchronous speed. The fact that the rotor field rotates at synchronous speed — although the rotor itself does not — is extremely important, because it means that the stator and rotor fields will continue to be stationary with respect to each other, and therefore a net torque can be produced.

As in the case of DC and synchronous motors, important characteristics of induction motors are the starting torque, the maximum torque, and the torque-speed curve.

Performance of Induction Motors

The performance of induction motors can be described by torque-speed curves similar to those already used for DC motors. Figure 5.12.42 depicts an induction motor torque-speed curve, with five torque ratings marked *a* through *e*. Point *a* is the *starting torque*, also called **breakaway torque**, and is the torque available with the rotor "locked", that is, in a stationary position. At this condition, the frequency of the voltage induced in the rotor is highest, since it is equal to the frequency of rotation of the stator field; consequently, the inductive reactance of the rotor is greatest. As the rotor accelerates, the torque drops off, reaching a maximum value called the **pull-up torque** (point *b*); this typically occurs somewhere between 25 and 40% of synchronous speed. As the rotor speed continues to increase, the rotor reactance decreases further (since the frequency of the induced voltage is determined by the relative speed of rotation of the rotor with respect to the stator field). The torque becomes a maximum when the rotor inductive reactance is equal to the rotor resistance; maximum torque is also called **breakdown torque** (point *c*). Beyond this point, the torque drops off, until it is zero at synchronous speed, as discussed earlier. Also marked on the curve are the *150% torque* (point *d*), and the *rated torque* (point *e*).

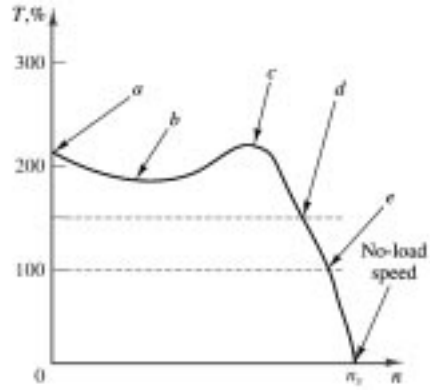


FIGURE 5.12.42 Performance curve for induction motor.

A general formula for the computation of the induction motor steady-state torque-speed characteristic is

$$T = \frac{1}{\omega_e} \frac{mV_S^2 R_R / s}{\left[\left(R_S + \dfrac{R_R}{s}\right)^2 + \left(X_S + X_R\right)^2\right]} \tag{5.12.57}$$

where m is the number of phases.

Different construction arrangements permit the design of induction motors with different torque-speed curves, thus permitting the user to select the motor that best suits a given application. Figure 5.12.43 depicts the four basic classifications, classes A, B, C, and D, as defined by NEMA. The determining features in the classification are the locked-rotor torque and current, the breakdown torque, the pull-down torque, and the percent slip. Class A motors have a higher breakdown torque than class B motors, and a slip of 5% or less. Motors in this class are often designed for a specific application. Class B motors are general-purpose motors; this is the most commonly used type of induction motor, with typical values of slip of 3 to 5%. Class C motors have a high starting torque for a given starting current, and a low slip. These motors are typically used in applications demanding high starting torque but having relatively normal running loads, once running speed has been reached. Class D motors are characterized by high starting torque, high slip, low starting current, and low full-load speed. A typical value of slip is around 13%.

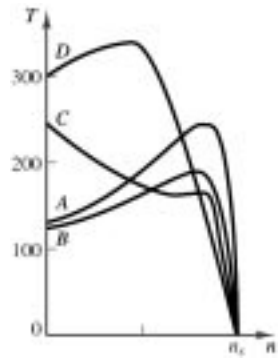


FIGURE 5.12.43 Induction motor classification.

Factors that should be considered in the selection of an AC motor for a given application are the *speed range*, both minimum and maximum, and the speed variation. For example, it is important to determine whether constant speed is required; what variation might be allowed, either in speed or in torque; or whether variable-speed operation is required, in which case a variable-speed drive will be needed. The torque requirements are obviously important as well. The starting and running torque should be considered; they depend on the type of load. Starting torque can vary from a small percentage of full-load to several times full-load torque. Furthermore, the excess torque available at start-up determines the *acceleration characteristics* of the motor. Similarly, *deceleration characteristics* should be considered, to determine whether external braking might be required.

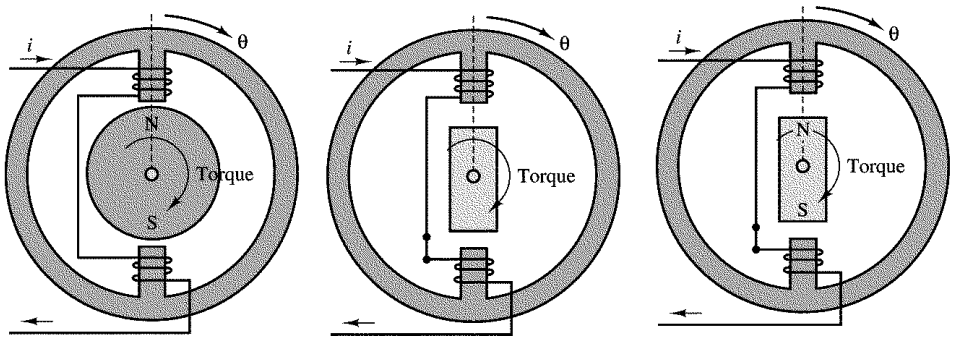
Another factor to be considered is the *duty cycle* of the motor. The duty cycle, which depends on the nature of the application, is an important consideration when the motor is used in repetitive, noncontinuous operation, such as is encountered in some types of machine tools. If the motor operates at zero or reduced load for periods of time, the duty cycle — that is, the percentage of the time the motor is loaded — is an important selection criterion. Last, but by no means least, are the *heating properties* of a motor. Motor temperature is determined by internal losses and by ventilation; motors operating at a reduced speed may not generate sufficient cooling, and forced ventilation may be required.

Stepping Motors

Stepping, or **stepper, motors** are motors that convert digital information to mechanical motion. The principles of operation of stepping motors have been known since the 1920s; however, their application has seen a dramatic rise with the increased use of digital computers. Stepping motors, as the name suggests, rotate in distinct steps, and their position can be controlled by means of logic signals. Typical applications of stepping motors are line printers, positioning of heads in magnetic disks drives, and any other situation where continuous or stepwise displacements are required.

Stepping motors can generally be classified in one of three categories: variable-reluctance, permanent-magnet, and hybrid types. It will soon be shown that the principles of operation of each of these devices bear a definite resemblance to those of devices already encountered in this book. Stepping motors have a number of special features that make them particularly useful in practical applications. Perhaps the most important feature of a stepping motor is that the angle of rotation of the motor is directly proportional to the number of input pulses; further, the angle error per step is very small and does not accumulate. Stepping motors are also capable of rapid response to starting, stopping, and reversing commands, and can be driven directly by digital signals. Another important feature is a self-holding capability that makes it possible for the rotor to be held in the stopped position without the use of brakes. Finally, a wide range of rotating speeds — proportional to the frequency of the pulse signal — may be attained in these motors.

Figure 5.12.44 depicts the general appearance of three types of stepping motors. The **permanent-magnet-rotor stepping motor**, Figure 5.12.44(a), permits a nonzero holding torque when the motor is not energized. Depending on the construction of the motor, it is typically possible to obtain step angles of 7.5, 11.25, 15, 18, 45, or 90°. The angle of rotation is determined by the number of stator poles, as will be illustrated shortly in an example. The **variable-reluctance stepping motor,** Figure 5.12.44(b), has an iron multipole rotor and a laminated wound stator, and rotates when the teeth on the rotor are attracted to the electromagnetically energized stator teeth. The rotor inertia of a variable-reluctance stepping motor is low, and the response is very quick, but the allowable load inertia is small. When the windings are not energized, the static torque of this type of motor is zero. Generally, the step angle of the variable-reluctance stepping motor is 15°.



(a) Permanent-magnet stepping motor (b) Variable-reluctance stepping motor (c) Hybrid stepping motor

FIGURE 5.12.44 Stepping motor configurations.

The **hybrid stepping motor**, Figure 5.12.44(c), is characterized by multitoothed stator and rotor, the rotor having an axially magnetized concentric magnet around its shaft. It can be seen that this configuration is a mixture of the variable-reluctance and permanent-magnet types. This type of motor generally has high accuracy and high torque and can be configured to provide a step angle as small as 1.8°.

For any of these configurations, the principle of operation is essentially the same: when the coils are energized, magnetic poles are generated in the stator, and the rotor will align in accordance with the direction of the magnetic field developed in the stator. By reversing the phase of the currents in the coils or by energizing only some of the coils (this is possible in motors with more than two stator poles), the alignment of the stator magnetic field can take one of a discrete number of positions; if the currents in the coils are pulsed in the appropriate sequence, the rotor will advance in a step-by-step fashion. Thus, this type of motor can be very useful whenever precise incremental motion must be attained. As mentioned earlier, typical applications are printer wheels, computer disk drives, and plotters. Other applications are found in the control of the position of valves (e.g., control of the throttle valve in an

engine or of a hydraulic valve in a fluid power-system), and in drug-dispensing apparatus for clinical applications.

The Universal Motor

If it were possible to operate a DC motor from a single-phase AC supply, a wide range of simple applications would become readily available. Recall that the direction of the torque produced by a DC machine is determined by the direction of current flow in the armature conductors and by the polarity of the field; torque is developed in a DC machine because the commutator arrangement permits the field and armature currents to remain in phase, thus producing torque in a constant direction. A similar result can be obtained by using an AC supply, and by connecting the armature and field windings in series, as shown in Figure 5.12.45. A series DC motor connected in this configuration can therefore operate on a single-phase AC supply, and is referred to as a **universal motor**. An additional consideration is that, because of the AC excitation, it is necessary to reduce AC core losses by laminating the stator; thus, the universal motor differs from the series DC motor in its construction features. Typical torque-speed curves for AC and DC operation of a universal motor are shown in Figure 5.12.46. As shown in Figure 5.12.45, the load current is sinusoidal and therefore reverses direction each half-cycle; however, the torque generated by the motor is always in the same direction, resulting in a pulsating torque, with nonzero averaging value.

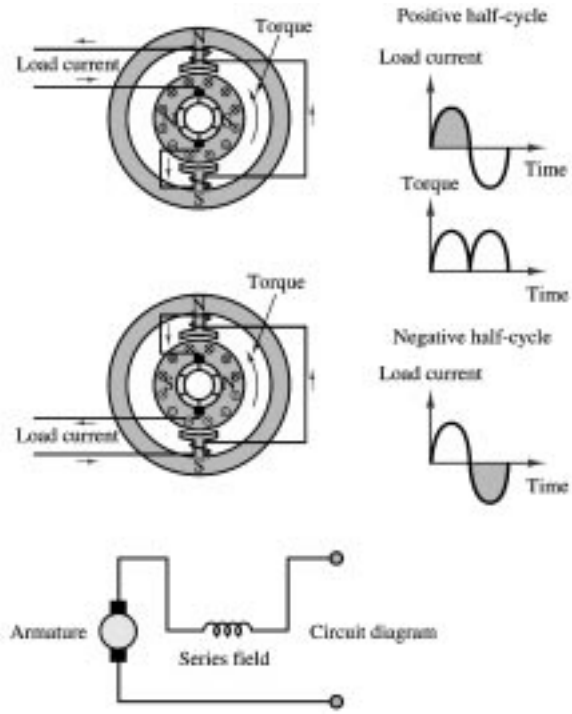


FIGURE 5.12.45 Operation and circuit diagram of a universal motor.

As in the case of a DC series motor, the best method for controlling the speed of a universal motor is to change its (rms) input voltage. The higher the rms input voltage, the greater the resulting speed of the motor. Approximate torque-speed characteristics of a universal motor as a function of voltage are shown in Figure 5.12.47.

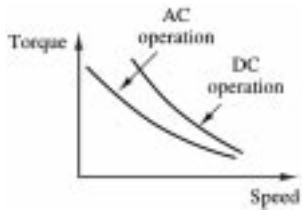


FIGURE 5.12.46 Torque-speed curve of a universal motor.

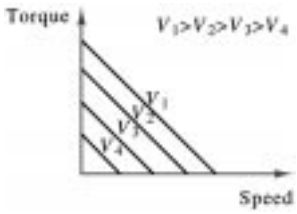


FIGURE 5.12.47 Torque-speed characteristics of a universal motor.

Single-Phase Induction Motors

Thus far, we have not mentioned how the initial starting torque can be provided to a single-phase motor. In practice, single-phase motors are classified by their starting and running characteristics, and several methods exist to provide nonzero starting torque. The aim of this section is to classify single-phase motors by describing their configuration on the basis of the method of starting. For each class of motor, a torque-speed characteristic will also be described.

Split-Phase Motors

Split-phase motors are constructed with two separate stator windings, called **main** and **auxiliary windings**; the axes of the two windings are actually at 90° with respect to each other, as shown in Figure 5.12.48. The auxiliary winding current is designed to be out of phase with the main winding current, as a result of different reactances of the two windings. Different winding reactances can be attained by having a different ratio of resistance to inductance — for example, by increasing the resistance of the auxiliary winding. In particular, the auxiliary winding current, $\mathbf{I}_{aux}$, leads the main winding current, $\mathbf{I}_{main}$. The net effect is that the motor sees a two-phase (unbalanced) current that results in a rotating magnetic field, as in any polyphase stator arrangement. Thus, the motor has a nonzero starting torque, as shown in Figure 5.12.49. Once the motor, a centrifugal switch is used to disconnect the auxiliary winding, since a single winding is sufficient to sustain the motion of the rotor. The switching action permits the use of relatively high-resistance windings, since these are not used during normal operation and therefore one need not be concerned with the losses associated with a higher-resistance winding. Figure 5.12.49 also depicts the combined effect of the two modes of operation of the split-phase motor.

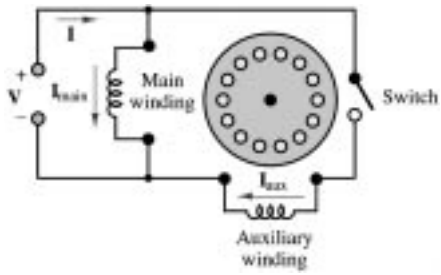


FIGURE 5.12.48 Split-phase motor.

Split-phase motors have appropriate characteristics (at very low cost) for fans, blowers, centrifugal pumps, and other applications in the range of 1/20 to 1/2 hp.

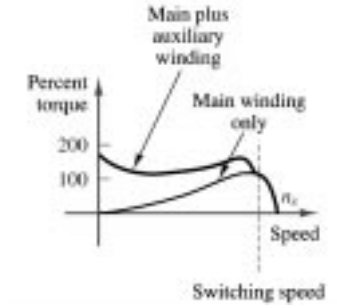


FIGURE 5.12.49 Torque-speed curve of split-phase motor.

Capacitor-Type Motors

Another method for obtaining a phase difference between currents that will give rise to a rotating magnetic field is by the addition of a capacitor. Motors that use this arrangement are termed **capacitor-type motors**. These motors make different use of capacitors to provide starting or running capabilities, or a combination of the two. The **capacitor-start motor** is essentially identical to the split-phase motor, except for the addition of a capacitor in series with the auxiliary winding, as shown in Figure 5.12.50. The addition of the capacitor changes the reactance of the auxiliary circuit in such a way as to cause the auxiliary current to lead the main current. The advantage of using the capacitor as a means for achieving a phase split is that greater starting torque may be obtained than with the split-phase arrangement. A centrifugal switching arrangement is used to disconnect the auxiliary winding above a certain speed, in the neighborhood of 75% of synchronous speed.

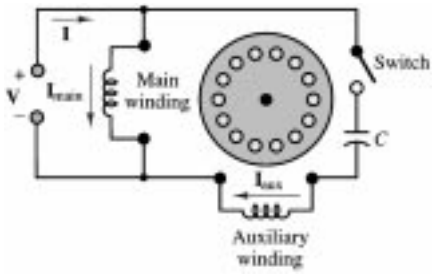


FIGURE 5.12.50 Capacitor-start motor.

Figure 5.12.51 depicts the torque-speed characteristic of a capacitor-start motor. Because of their higher starting torque, these motors are very useful in connection with loads that present a high static torque. Examples of such loads are compressors, pumps, and refrigeration and air-conditioning equipment.

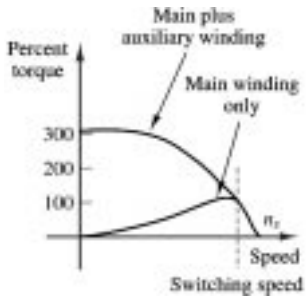


FIGURE 5.12.51 Torque-speed curve for a capacitor-start motor.

It is also possible to use the capacitor-start motor without the centrifugal switch, leading to a simpler design. Motors with this design are called **permanent split-capacitor motors**; they offer a compromise between running and starting characteristics. A typical torque-speed curve is shown in Figure 5.12.52.

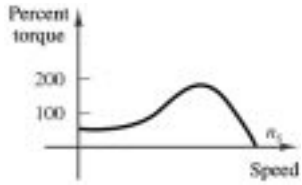


FIGURE 5.12.52 Torque-speed curve for a permanent split-capacitor motor.

A further compromise can be achieved by using two capacitors, one to obtain a permanent phase split and the resulting improvement in running characteristics, the other to improve the starting torque. A small capacitance is sufficient to improve the running performance, while a much larger capacitor provides the temporary improvement in starting torque. A motor with this design is called a **capacitor-start capacitor-run motor**; its schematic diagram is shown in Figure 5.12.53. Its torque-speed characteristic is similar to that of a capacitor-start motor.

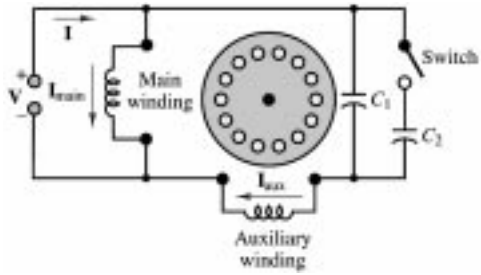


FIGURE 5.12.53 Capacitor-start capacitor-run motor.

Shaded-Pole Motors

The last type of single-phase induction motor discussed in this chapter is the **shaded-pole motor**. This type of motor operates on a different principle from the motors discussed thus far. The stator of a shaded-pole motor has a salient pole construction, as shown in Figure 5.12.54, that includes a shading coil consisting of a copper band wound around part of each pole. The flux in the shaded portion of the pole lags behind the flux in the unshaded part, achieving an effect similar to a rotation of the flux in the direction of the shaded part of the pole. This flux rotation in effect produces a rotating field that enables the motor to have a starting torque. This construction technique is rather inexpensive and is used in motors up to about 1/20 hp.

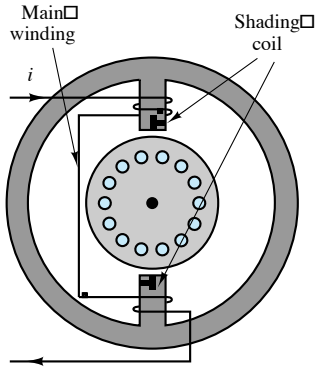


FIGURE 5.12.54 Shaded-pole motor.

A typical torque-speed characteristic for a shaded-pole motor is given in Figure 5.12.55.

Summary of Single-Phase Motor Characteristics

For basic classes of single-phase motors are commonly used:

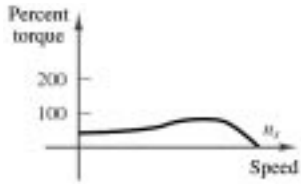


FIGURE 5.12.55 Torque-speed curve of a shaded-pole motor.

1. Single-phase induction motors are used for the larger home and small business tasks, such as furnace oil burner pumps, or hot water or hot air circulators. Refrigerator compressors, lathes, and bench-mounted circular saws are also powered with induction motors.
2. Shaded-pool motors are used in the smaller sizes for quite, low-cost applications. The size range is for 1/30 hp (24.9 W) to 1/2 hp (373 W), particularly for fans and similar drives in which the starting torque is low.
3. Universal motors will operate on any household AC frequency or on DC without modification or adjustment. They can develop very high speed while loaded, and very high power for their size. Vacuum cleaners, sewing machines, kitchen food mixers, portable electric drills, portable circular saws, and home motion-picture projectors are examples of applications of universal motors.
4. The capacitor-type motor finds its widest field of application at low speeds (below 900 rev/min) and in ratings from 3/4 hp (0.5595 kW) to 3 hp (2.238 kW) at all speeds, especially in fan drives.

References

Section 2

Irwin, J.D., 1989. *Basic Engineering Circuit Analysis,* 3rd ed. Macmillan, New York.
Nilsson, J.W., 1989. *Electric Circuits,* 3rd ed. Addison-Wesley, Reading, MA.
Rizzoni, G., 1966. *Principles and Applications of Electrical Engineering,* 2nd ed. Richard D. Irwin, Burr Ridge, IL.
Smith, R.J. and Dorf, R.C., 1992. *Circuits, Devices and Systems,* 5th ed. John Wiley & Sons, New York.
1993. *The Electrical Engineering Handbook,* CRC Press, Boca Raton, FL.

Section 3

Irwin, J.D., 1989. *Basic Engineering Circuit Analysis,* 3rd ed. Macmillan, New York.
Nilsson, J.W., 1989. *Electric Circuits,* 3rd ed. Addison-Wesley, Reading, MA.
Rizzoni, G., 1966. *Principles and Applications of Electrical Engineering,* 2nd ed. Richard D. Irwin, Burr Ridge, IL.
Smith, R.J. and Dorf, R.C., 1992. *Circuits, Devices and Systems,* 5th ed. John Wiley & Sons, New York.
1993. *The Electrical Engineering Handbook,* CRC Press, Boca Raton, FL.

Section 4

Budak, A., *Passive and Active Network Analysis and Synthesis*, Houghton Mifflin, Boston.
Irwin, J.D., 1989. *Basic Engineering Circuit Analysis,* 3rd ed. Macmillan, New York.
Nilsson, J.W., 1989. *Electric Circuits,* 3rd ed. Addison-Wesley, Reading, MA.
Rizzoni, G., 1966. *Principles and Applications of Electrical Engineering,* 2nd ed. Richard D. Irwin, Burr Ridge, IL.
Smith, R.J. and Dorf, R.C., 1992. *Circuits, Devices and Systems,* 5th ed. John Wiley & Sons, New York.
Van Valkenburg, M.E., 1982, *Analog Filter Design,* Holt, Rinehart & Winston, New York.
1993. *The Electrical Engineering Handbook,* CRC Press, Boca Raton, FL.

Section 5

Del Toro, V., 1992. *Electric Power Systems,* Prentice-Hall, Englewood Cliffs, NJ.
Nilsson, J.W., 1989. *Electric Circuits,* 3rd ed. Addison-Wesley, Reading, MA.
Rizzoni, G., 1966. *Principles and Applications of Electrical Engineering,* 2nd ed. Richard D. Irwin, Burr Ridge, IL.
Smith, R.J. and Dorf, R.C., 1992. *Circuits, Devices and Systems,* 5th ed. John Wiley & Sons, New York.
1993. *The Electrical Engineering Handbook,* CRC Press, Boca Raton, FL.

Section 6

Irwin, J.D., 1989. *Basic Engineering Circuit Analysis,* 3rd ed. Macmillan, New York.
Nilsson, J.W., 1989. *Electric Circuits,* 3rd ed. Addison-Wesley, Reading, MA.
Rizzoni, G., 1966. *Principles and Applications of Electrical Engineering,* 2nd ed. Richard D. Irwin, Burr Ridge, IL.
Smith, R.J. and Dorf, R.C., 1992. *Circuits, Devices and Systems,* 5th ed. John Wiley & Sons, New York.
1993. *The Electrical Engineering Handbook,* CRC Press, Boca Raton, FL.

Section 7

Horowitz, P. and Hill, W., 1989. *The Art of Electronics,* 2nd ed. Cambridge University Press, New York.
Millman, J. and Grabel, A., 1987. *Microelectronics*. McGraw-Hill, New York.
Sedra, A.S. and Smith, K.C., 1991. *Microelectronic Circuits,* 3rd ed. W.B. Saunders, Philadelphia.
Neamen, D.A., 1994. *Semiconductor Physics and Devices*. Richard D. Irwin, Burr Ridge, IL.
Rizzoni, G., 1966. *Principles and Applications of Electrical Engineering,* 2nd ed. Richard D. Irwin, Burr Ridge, IL.
Streetman, B.G., 1990. *Solid State Electronic Devices,* 3rd ed. Prentice-Hall, Englewood Cliffs, NJ.
Sze, S.M., 1981. *Physics of Semiconductor Devices,* 2nd ed. Wiley, New York.
1993. *The Electrical Engineering Handbook,* CRC Press, Boca Raton, FL.

Section 8

Bose, B.K., 1986. *Power Electronics and AC Drives,* Prentice-Hall, Englewood Cliffs, NJ.
Mohan, N., Undeland, T.M., and Robbins, P., *Power Electronics*, Van Nostrand, New York.
Rashid, M.H., 1988. *Power Electronics,* Prentice-Hall, Englewood Cliffs, NJ.
Rizzoni, G., 1966. *Principles and Applications of Electrical Engineering,* 2nd ed. Richard D. Irwin, Burr Ridge, IL.
1993. *The Electrical Engineering Handbook,* CRC Press, Boca Raton, FL.

Section 9

Horowitz, P. and Hill, W., 1989. *The Art of Electronics,* 2nd ed. Cambridge University Press, New York.
Millman, J. and Grabel, A., 1987. *Microelectronics*. McGraw-Hill, New York.
Rizzoni, G., 1966. *Principles and Applications of Electrical Engineering,* 2nd ed. Richard D. Irwin, Burr Ridge, IL.
Kennedy, E.J., 1988. *Operational Amplifier Circuits,* Holt, Rinehart & Winston, New York.
1993. *The Electrical Engineering Handbook,* CRC Press, Boca Raton, FL.

Section 10

Breeding, K.J., 1992. *Digital Design Fundamentals,* 2nd ed. Prentice-Hall, Englewood Cliffs, NJ.
Horowitz, P. and Hill, W., 1989. *The Art of Electronics,* 2nd ed., Cambridge University Press, New York.
Mano, M.M., 1988. *Computer Engineering Hardware Design,* Prentice-Hall, Englewood Cliffs, NJ.

Rizzoni, G., 1966. *Principles and Applications of Electrical Engineering,* 2nd ed. Richard D. Irwin, Burr Ridge, IL.
Sandige, R.S., 1990. *Modern Digital Design*. McGraw-Hill, New York.
1993. *The Electrical Engineering Handbook,* CRC Press, Boca Raton, FL.

Section 11

Doebelin, E.O., 1990. *Measurement Systems: Application and Design,* 4th ed. McGraw-Hill, New York.
Annino, R. and Driver, R., 1986. *Scientific and Engineering Applications with Personal Computers*. Wiley-Interscience, New York.
Horowitz, P. and Hill, W., 1989. *The Art of Electronics,* 2nd ed., Cambridge University Press, New York.
Nachitgal, C. (ed.), 1990. *Instrumentation and Control: Fundamentals and Applications*. John Wiley & Sons, New York.
Rizzoni, G., 1966. *Principles and Applications of Electrical Engineering,* 2nd ed. Richard D. Irwin, Burr Ridge, IL.
Webster, J.G., 1992. *Medical Instrumentation: Application and Design,* 2nd ed. Houghton Mifflin, Boston.
1993. *The Electrical Engineering Handbook,* CRC Press, Boca Raton, FL.

Section 12

Del Toro, V., 1990. *Basic Electric Machines*. Prentice-Hall, Englewood Cliffs, NJ.
Fitzgerald, A.E., Kingsley, C., and Umans, S., 1990. *Electric Machinery,* 5th ed. McGraw-Hill, New York.
Krause, P. and Wasynczuk, O., 1989. *Electromechanical Motion Devices*. McGraw-Hill, New York.
Rizzoni, G., 1966. *Principles and Applications of Electrical Engineering,* 2nd ed. Richard D. Irwin, Burr Ridge, IL.
1993. *The Electrical Engineering Handbook,* CRC Press, Boca Raton, FL.

Kreider, J.F.; Curtiss, P.S.; et. al. "Mechanical System Controls"
Mechanical Engineering Handbook
Ed. Frank Kreith
Boca Raton: CRC Press LLC, 1999

Mechanical System Controls

Jan F. Kreider
University of Colorado

Peter S. Curtiss
Architectural Energy Corporation

Thomas B. Sheridan
Massachusetts Institute of Technology

Shou-Heng Huang
Raytheon Co. Appliance Tech Center

Ron M. Nelson
Iowa State University

6.1 Human-Machine Interaction

Thomas B. Sheridan

Over the years machines of all kinds have been improved and made more reliable. However, machines typically operate as components of larger systems, such as transportation systems, communication systems, manufacturing systems, defense systems, health care systems, and so on. While many aspects of such systems can be and have been automated, the human operator is retained in many cases. This may be because of economics, tradition, cost, or (most likely) capabilities of the human to perceive patterns of information and weigh subtle factors in making control decisions which the machine cannot match.

Although the public as well as those responsible for system operation usually demand that there be a human operator, “human error” is a major reason for system failure. And aside from prevention of

error, getting the best performance out of the system means that human and machine must be working together effectively — be properly "impedance matched." Therefore, the performance capabilities of the human relative to those of the machine must be taken into account in system design.

Efforts to "optimize" the human-machine interaction are meaningless in the mathematical sense of optimization, since most important interactions between human and machine cannot be reduced to a mathematical form, and the objective function (defining what is good) is not easily obtained in any given context. For this reason, engineering the human-machine interaction, much as in management or medicine, remains an art more than a science, based on laboratory experiments and practical experience.

In the broadest sense, engineering the human-machine interface includes all of *ergonomics* or *human factors engineering*, and goes well beyond design of displays and control devices. Ergonomics includes not only questions of sensory physiology, whether or not the operator can see the displays or hear the auditory warnings, but also questions of *biomechanics*, how the body moves, and whether or not the operator can reach and apply proper force to the controls. It further includes the fields of operator selection and training, human performance under stress, human factors in maintenance, and many other aspects of the relation of the human to technology. This section focuses primarily on human-machine interaction in control of systems.

The human-machine interactions in control are considered in terms of Figure 6.1.1. In Figure 6.1.1a the human directly controls the machine; i.e., the control loop to the machine is closed through the physical sensors, displays, human senses (visual, auditory, tactile), brain, human muscles, control devices, and machine actuators. Figure 6.1.1b illustrates what has come to be called a *supervisory control system*, wherein the human intermittently instructs a computer as to goals, constraints, and procedures, then turns a task over to the computer to perform automatic control for some period of time.

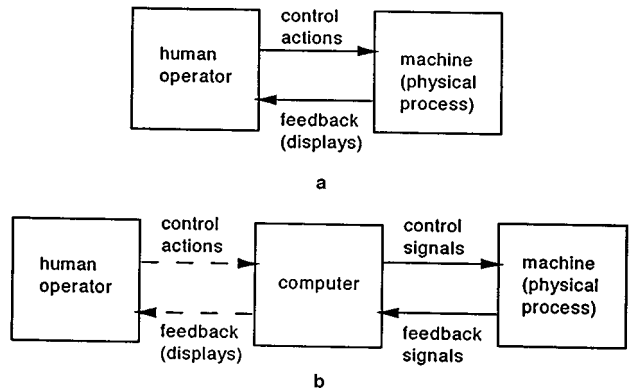


FIGURE 6.1.1 Direct manual control (a) and supervisory control (b).

Displays and control devices can be *analogic* (movement signal directions and extent of control action, isomorphic with the world, such as an automobile steering wheel or computer mouse controls, or a moving needle or pictorial display element). Or they can be *symbolic* (dedicated buttons or general-purpose keyboard controls, icons, or alarm light displays). In normal human discourse we use both speech (symbolic) and gestures (analogic) and on paper we write alphanumeric text (symbolic) and draw pictures (analogic). The system designer must decide which type of displays or controls best suits a particular application, and/or what mix to use. The designer must be aware of important criteria such as whether or not, for a proposed design, changes in the displays and controls caused by the human operator correspond in a natural and common-sense way to "more" or "less" of some variable as expected by that operator and correspond to cultural norms (such as reading from left to right in western countries), and whether or not the movement of the display elements correspond geometrically to movements of the controls.

Direct Manual Control

In the 1940s aircraft designers appreciated the need to characterize the transfer function of the human pilot in terms of a differential equation. Indeed, this is necessary for any vehicle or controlled physical process for which the human is the controller, see Figure 6.1.2. In this case both the human operator *H* and the physical process *P* lie in the closed loop (where *H* and *P* are Laplace transforms of the component transfer functions), and the *HP* combination determines whether the closed-loop is inherently stable (i.e., the closed loop characteristic equation 1 + HP = 0 has only negative real roots).

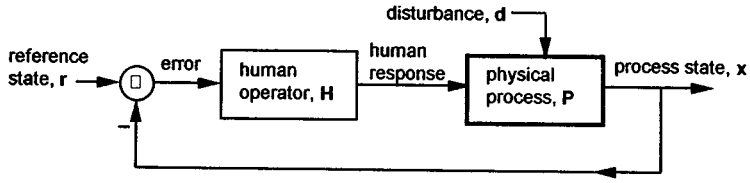


FIGURE 6.1.2 Direct manual control-loop analysis.

In addition to the stability criterion are the criteria of rapid response of process state *x* to a desired or reference state *r* with minimum overshoot, zero "steady-state error" between *r* and output *x*, and reduction to near zero of the effects of any disturbance input *d*. (The latter effects are determined by the closed-loop transfer functions $x = HP/(1 + HP)r + 1/(1 + HP)d$, where if the magnitude of *H* is large enough $HP/(1 + HP)$ approaches unity and $1/(1 + HP)$ approaches 0. Unhappily, there are ingredients of *H* which produce delays in combination with magnitude and thereby can cause instability. Therefore, *H* must be chosen carefully by the human for any given *P*.)

Research to characterize the pilot in these terms resulted in the discovery that the human adapts to a wide variety of physical processes so as to make $HP = K(1/s)(e^{-sT})$. In other words, the human adjusts *H* to make *HP* constant. The term *K* is an overall amplitude or gain, $(1/s)$ is the Laplace transform of an integrator, and (e^{-sT}) is a delay *T* long (the latter time delay being an unavoidable property of the nervous system). Parameters *K* and *T* vary modestly in a predictable way as a function of the physical process and the input to the control system. This model is now widely accepted and used, not only in engineering aircraft control systems, but also in designing automobiles, ships, nuclear and chemical plants, and a host of other dynamic systems.

Supervisory Control

Supervisory control may be defined by the analogy between a supervisor of subordinate staff in an organization of people and the human overseer of a modern computer-mediated semiautomatic control system. The supervisor gives human subordinates general instructions which they in turn may translate into action. The supervisor of a computer-controlled system does the same.

Defined strictly, *supervisory control* means that one or more human operators are setting initial conditions for, intermittently adjusting, and receiving high-level information from a computer that itself closes a control loop in a well-defined process through artificial sensors and effectors. For some time period the computer controls the process automatically.

By a less strict definition, *supervisory control* is used when a computer transforms human operator commands to generate detailed control actions, or makes significant transformations of measured data to produce integrated summary displays. In this latter case the computer need not have the capability to commit actions based upon new information from the environment, whereas in the first it necessarily must. The two situations may appear similar to the human supervisor, since the computer mediates both human outputs and human inputs, and the supervisor is thus removed from detailed events at the low level.

A supervisory control system is represented in Figure 6.1.3. Here the human operator issues commands to a *human-interactive* computer capable of understanding high-level language and providing integrated summary displays of process state information back to the operator. This computer, typically located in a control room or cockpit or office near to the supervisor, in turn communicates with at least one, and probably many (hence the dotted lines), *task-interactive* computers, located with the equipment they are controlling. The task-interactive computers thus receive subgoal and conditional branching information from the human-interactive computer. Using such information as reference inputs, the task-interactive computers serve to close low-level control loops between artificial sensors and mechanical actuators; i.e., they accomplish the low-level automatic control.

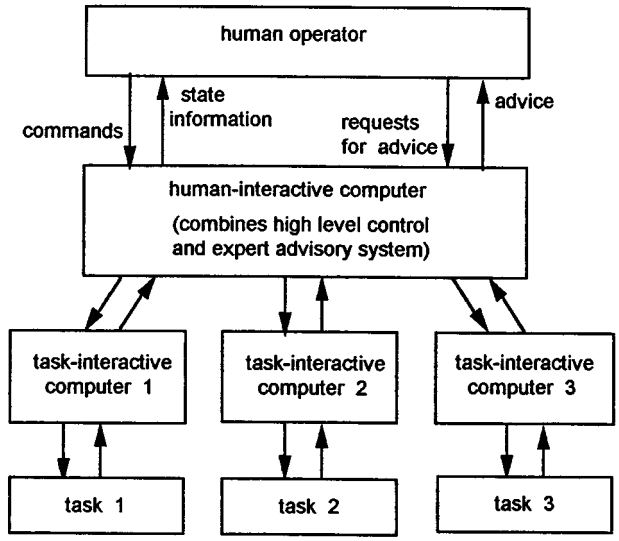


FIGURE 6.1.3 Supervisory control.

The low-level task typically operates at some physical distance from the human operator and his human-friendly display-control computer. Therefore, the communication channels between computers may be constrained by multiplexing, time delay, or limited bandwidth. The task-interactive computer, of course, sends analog control signals to and receives analog feedback signals from the controlled process, and the latter does the same with the environment as it operates (vehicles moving relative to air, sea, or earth, robots manipulating objects, process plants modifying products, etc.).

Supervisory command and feedback channels for process state information are shown in Figure 6.1.3 to pass through the left side of the human-interactive computer. On the right side are represented decision-aiding functions, with requests of the computer for advice and displayed output of advice (from a database, expert system, or simulation) to the operator. There are many new developments in computer-based decision aids for planning, editing, monitoring, and failure detection being used as an auxiliary part of operating dynamic systems. Reflection upon the nervous system of higher animals reveals a similar kind of supervisory control wherein commands are sent from the brain to local ganglia, and peripheral motor control loops are then closed locally through receptors in the muscles, tendons, or skin. The brain, presumably, does higher-level planning based on its own stored data and "mental models," an internalized expert system available to provide advice and permit trial responses before commitment to actual response.

Theorizing about supervisory control began as aircraft and spacecraft became partially automated. It became evident that the human operator was being replaced by the computer for direct control responsibility, and was moving to a new role of monitor and goal-constraint setter. An added incentive was the U.S. space program, which posed the problem of how a human operator on Earth could control a

manipulator arm or vehicle on the moon through a 3-sec communication round-trip time delay. The only solution which avoided instability was to make the operator a supervisory controller communicating intermittently with a computer on the moon, which in turn closed the control loop there. The rapid development of microcomputers has forced a transition from manual control to supervisory control in a variety of industrial and military applications (Sheridan, 1992).

Let us now consider some examples of human-machine interaction, particularly those which illustrate supervisory control in its various forms. First, we consider three forms of vehicle control, namely, control of modern aircraft, "intelligent" highway vehicles, and high-speed trains, all of which have both human operators in the vehicles as well as humans in centralized traffic-control centers. Second, we consider telerobots for space, undersea, and medical applications.

Advanced Control of Commercial Aircraft

Flight Management Systems

Aviation has appreciated the importance of human-machine interaction from its beginning, and today exemplifies the most sophisticated forms of such interaction. While there have been many good examples of display and control design over the years, the current development of the flight management systems (FMS) is the epitome. It also provides an excellent example of supervisory control, where the pilot flies the aircraft by communicating in high-level language through a computer intermediary. The FMS is a centralized computer which interacts with a great variety of sensors, communication from the ground, as well as many displays and controls within the aircraft. It embodies many functions and mediates most of the pilot information requirements shown in Figure 6.1.4. Gone are the days when each sensor had its own display, operating independently of all other sensor-display circuits. The FMS, for example, brings together all of the various autopilot modes, from long-standing low-level control modes, wherein the aircraft is commanded to go to and hold a commanded altitude, heading, and speed, to more-sophisticated modes where the aircraft is instructed to fly a given course, consisting of a sequence of way points (latitudes and longitudes) at various altitudes, and even land automatically at a given airport on a given runway.

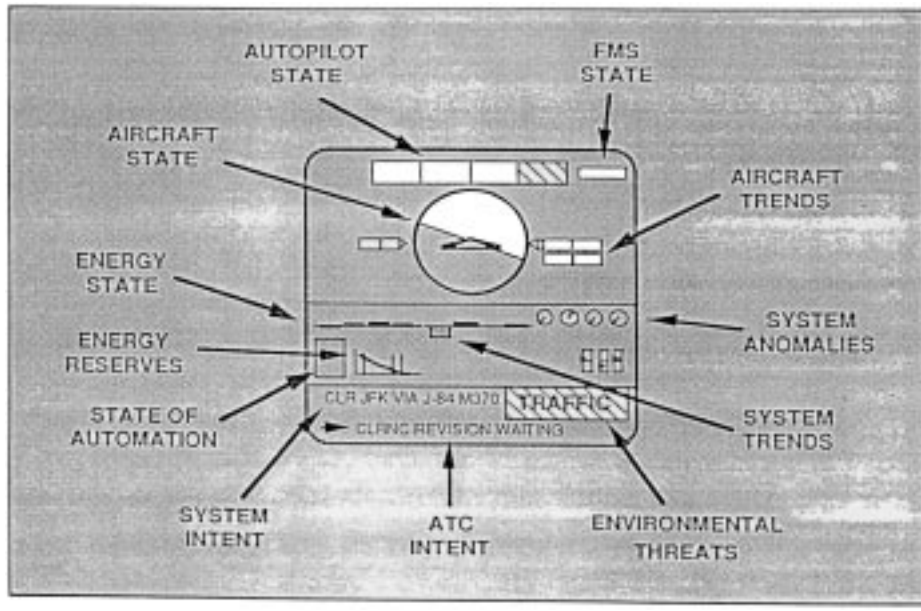


FIGURE 6.1.4 Pilot information requirements. (From Billings, 1991.)

Figure 6.1.5 illustrates one type of display mediated by the FMS, in this case integrating many formerly separate components of information. Mostly it is a multicolor plan-view map showing position and orientation of important objects relative to one's own aircraft (the triangle at the bottom). It shows heading (compass arc at top, present heading 175°), ground speed plus wind speed and wind direction (upper left), actual altitude relative to desired altitude (vertical scale on right side), programmed course connecting various way points (OPH and FLT), salient VOR radar beacons to the right and left of present position/direction with their codes and frequencies (lower left and right corners), the location of key VORs along the course (three-cornered symbols), the location of weather to be avoided (two gray blobs), and a predicted trajectory based on present turn rate, showing that the right turn is appropriately getting back on course.

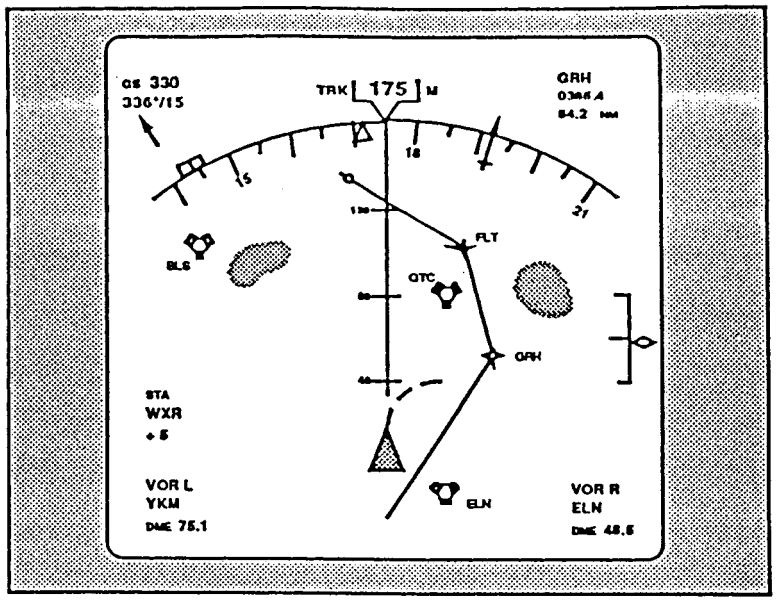


FIGURE 6.1.5 Integrated aircraft map display. (From Billings, 1991.)

Programming the FMS is done through a specialized keyboard and text display unit (Figure 6.1.6) having all the alphanumeric keys plus a number of special function keys. The displays in this case are specialized to the different phases of a flight (taxi, takeoff, departure, enroute approach, land, etc.), each phase having up to three levels of pages.

The FMS makes clear that designing displays and controls is no longer a matter of what can be built — the computer allows essentially any conceivable display/control to be realized. The computer can also provide a great deal of real-time advice, especially in emergencies, based on its many sensors and stored knowledge about how the aircraft operates. But pilots are not sure they need all the information which aircraft designers would like to give them, and have an expression "killing us with kindness" to refer to this plethora of available information. The question is what should be designed based on the needs and capabilities of the pilot.

Boeing, McDonnell Douglas, and Airbus have different philosophies for designing the FMS. Airbus has been the most aggressive in automating, intending to make piloting easier and safer for pilots from countries with less well established pilot training. Unfortunately, it is these most-automated aircraft which have had the most accidents of the modern commercial jets — a fact which has precipitated vigorous debate about how far to automate.

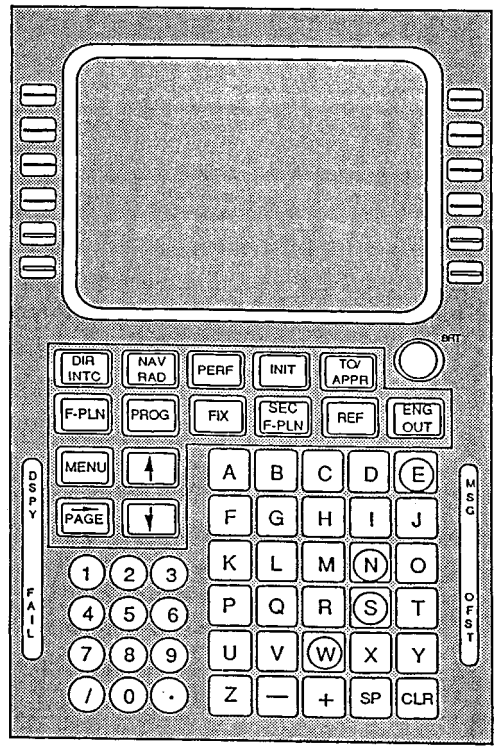


FIGURE 6.1.6 Flight management system control and display unit. (From Billings, 1991.)

Air Traffic Control

As demands for air travel continue to increase, so do demands for air traffic control. Given what are currently regarded as safe separation criteria, air space over major urban areas is already saturated, so that simply adding more airports is not acceptable (in addition to which residents do not want more airports, with their noise and surface traffic). The need is to reduce separations in the air, and to land aircraft closer together or on parallel runways simultaneously. This puts much greater demands on air traffic controllers, particularly at the terminal area radar control centers (TRACONs), where trained operators stare at blips on radar screens and verbally guide pilots entering the terminal airspace from various directions and altitudes into orderly descent and landing patterns with proper separation between aircraft.

Currently, many changes are being introduced into air traffic control which have profound implications for human-machine interaction. Where previously communication between pilots and air traffic controllers was entirely by voice, now digital communication between aircraft and ground (a system called *datalink)* allows both more and more reliable two-way communication, so that weather and runway and wind information, clearances, etc. can be displayed to pilots visually. But pilots are not so sure they want this additional technology. They fear the demise of the “party line” of voice communications with which they are so familiar and which permits all pilots in an area to listen in on each other’s conversations.

New aircraft-borne radars allow pilots to detect air traffic in their own vicinity. Improved ground-based radars detect microbursts or wind shear which can easily put an aircraft out of control. Both types of radars pose challenges as to how best to warn the pilot and provide guidance as to how to respond. But they also pose a cultural change in air traffic control, since heretofore pilots have been dependent upon air traffic controllers to advise them of weather conditions and other air traffic. Furthermore, because of the new weather and collision-avoidance technology, there are current plans for radically altering the rules whereby high-altitude commercial aircraft must stick to well-defined traffic lanes. Instead, pilots

will have great flexibility as to altitude (to find the most favorable winds and therefore save fuel) and be able to take great-circle routes straight to their destinations (also saving fuel). However, air traffic controllers are not sure they want to give up the power they have had, becoming passive observers and monitors, to function only in emergencies.

Intelligent Highway Vehicles

Vehicle Guidance and Navigation Systems

The combination of GPS (global positioning system) satellites, high-density computer storage of map data, electronic compass, synthetic speech synthesis, and computer-graphic displays allows cars and trucks to know where they are located on the Earth to within 100 m or less, and can guide a driver to a programmed destination by a combination of a map display and speech. Some human factor challenges are in deciding how to configure the map (how much detail to present, whether to make the map north-up with a moving dot representing one's own vehicle position or current-heading-up and rapidly changing with every turn). The computer graphics can also be used to show what turns to anticipate and which lane to get in. Synthetic speech can reinforce these turn anticipations, can caution the driver if he is perceived to be headed in the wrong direction or off course, and can even guide him or her how to get back on course. An interesting question is what the computer should say in each situation to get the driver's attention, to be understood quickly and unambiguously but without being an annoyance. Another question is whether or not such systems will distract the driver's attention from the primary tasks, thereby reducing safety. The major vehicle manufacturers have developed such systems, they have been evaluated for reliability and human use, and they are beginning to be marketed in the United States, Europe, and Japan.

Smart Cruise Control

Standard cruise control has a major deficiency in that it knows nothing about vehicles ahead, and one can easily collide with the rear end of another vehicle if not careful. In a smart cruise control system a microwave or optical radar detects the presence of a vehicle ahead and measures that distance. But there is a question of what to do with this information. Just warn the driver with some visual or auditory alarm (auditory is better because the driver does not have to be looking in the right place)? Can a warning be too late to elicit braking, or surprise the driver so that he brakes too suddenly and causes a rear-end accident to his own vehicle. Should the computer automatically apply the brakes by some function of distance to obstacle ahead, speed, and closing deceleration. If the computer did all the braking would the driver become complacent and not pay attention, to the point where a serious accident would occur if the radar failed to detect an obstacle, say, a pedestrian or bicycle, or the computer failed to brake? Should braking be some combination of human and computer braking, and if so by what algorithm? These are human factor questions which are currently being researched.

It is interesting to note that current developmental systems only decelerate and downshift, mostly because if the vehicle manufacturers sell vehicles which claim to perform braking they would be open to a new and worrisome area of litigation.

The same radar technology that can warn the driver or help control the vehicle can also be applied to cars overtaking from one side or the other. Another set of questions then arises as to how and what to communicate to the driver and whether or not to trigger some automatic control maneuver in certain cases.

Advanced Traffic Management Systems

Automobile congestion in major cities has become unacceptable, and advanced traffic management systems are being built in many of these cities to measure traffic flow at intersections (by some combination of magnetic loop detectors, optical sensors, and other means), and regulate stoplights and message signs. These systems can also issue advisories of accidents ahead by means of variable message signs or radio, and give advice of alternative routings. In emergencies they can dispatch fire, police,

ambulances, or tow trucks, and in the case of tunnels can shut down entering traffic completely if necessary. These systems are operated by a combination of computers and humans from centralized control rooms. The operators look at banks of video monitors which let them see the traffic flow at different locations, and computer-graphic displays of maps, alarm windows, and textual messages. The operators get advice from computer-based expert systems, which suggest best responses based on measured inputs, and the operator must decide whether to accept the computer's advice, whether to seek further information, and how to respond.

High-Speed Train Control

With respect to new electronic technology for information sensing, storage, and processing, railroad technology has lagged behind that of aircraft and highway vehicles, but currently is catching up. The role of the human operator in future rail systems is being debated, since for some limited right-of-way trains (e.g., in airports) one can argue that fully automatic control systems now perform safely and efficiently. The train driver's principal job is speed control (though there are many other monitoring duties he must perform), and in a train this task is much more difficult than in an automobile because of the huge inertia of the train — it takes 2 to 3 km to stop a high-speed train. Speed limits are fixed at reduced levels for curves, bridges, grade crossings, and densely populated areas, while wayside signals temporarily command lower speeds if there is maintenance being performed on the track, if there are poor environmental conditions such as rock slides or deep snow, or especially if there is another train ahead. The driver must obey all speed limits and get to the next station on time. Learning to maneuver the train with its long time constants can take months, given that for the speed control task the driver's only input currently is an indication of current speed.

The author's laboratory has proposed a new computer-based display which helps the driver anticipate the future effects of current throttle and brake actions. This approach, based on a dynamic model of the train, gives an instantaneous prediction of future train position and speed based on current acceleration, so that speed can be plotted on the display assuming the operator holds to current brake-throttle settings. It also plots trajectories for maximum emergency braking and maximum service braking. In addition, the computer generates a speed trajectory which adheres at all (known) future speed limits, gets to the next station on time, and minimizes fuel/energy. Figure 6.1.7 shows the laboratory version of this display, which is currently being evaluated.

Telerobots for Space, Undersea, and Medicine

When nuclear power was first adopted in the late 1940s engineers began the development of master-slave remote manipulators, by which a human operator at one location could position and orient a device attached to his hand, and a servomechanism-controlled gripper would move in correspondence and handle objects at another location. At about the same time, remotely controlled wheeled vehicles, submarines, and aircraft began to be developed. Such manipulators and vehicles remotely controlled by humans are called *teleoperators*. Teleoperator technology got a big boost from the industrial robot technology, which came in a decade or so later, and provided improved vision, force, and touch sensors, actuators, and control software. Large teleoperators were developed for rugged mining and undersea tasks, and small teleoperators were developed for delicate tasks such as eye surgery. Eventually, teleoperators have come to be equipped with sensitive force feedback, so that the human operator not only can see the objects in the remote environment, but also can feel them in his grasp.

During the time of the Apollo flights to the moon, and stimulated by the desire to control lunar manipulators and vehicles from Earth and the fact that the unavoidable round-trip time delays of 3 sec (speed of light from Earth to moon and back) would not permit simple closed loop control, supervisory controlled teleoperators were developed. The human could communicate a subgoal to be reached and a procedure for getting there, and the teleoperator would be turned loose for some short period to perform automatically. Such a teleoperator is called a *telerobot.*

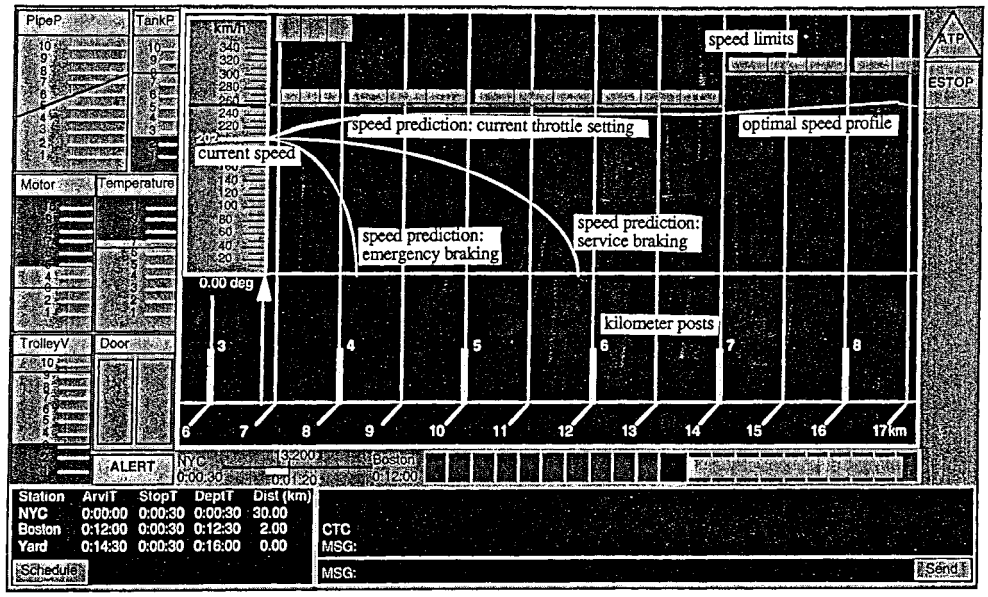


FIGURE 6.1.7 Prototype of computer-generated display for high speed trains. (From Askey, 1995.)

Figure 6.1.8 shows a the Flight Telerobotic Servicer (FTS) developed by Martin Marietta for the U.S. Space Station Freedom. It has two seven-degree of freedom (DOF) arms (including gripper) and one five-DOF "leg" for stabilizing itself while the arms work. It has two video "eyes" to present a stereoimage to its human operator. It can be configured either as a master-slave teleoperator (under direct human control) or as a telerobot (able to execute small programmed tasks using its own eyes and force sensors). Unfortunately, the FTS project was canceled by Congress.

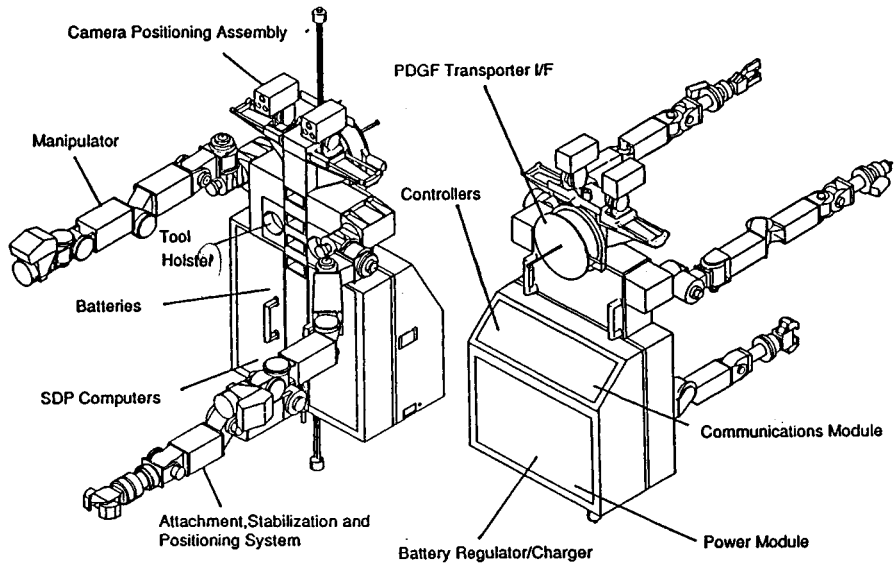


FIGURE 6.1.8 Flight Telerobotic Servicer prototype design. (Courtesy of NASA.)

Figure 6.1.9 shows the remotely operated submersible *Jason* developed by Woods Hole Oceanographic Institution. It is the big brother of *Jason Junior*, which swam into the interior of the ship *Titanic* and made a widely viewed video record when the latter was first discovered. It has a single manipulator arm, sonar and photo sensors, and four thrusters which can be oriented within limited range and which enable it to move in any direction. It is designed for depths up to 6000 m — rather severe pressures! It too, can be operated either in direct teleoperator mode or as a telerobot.

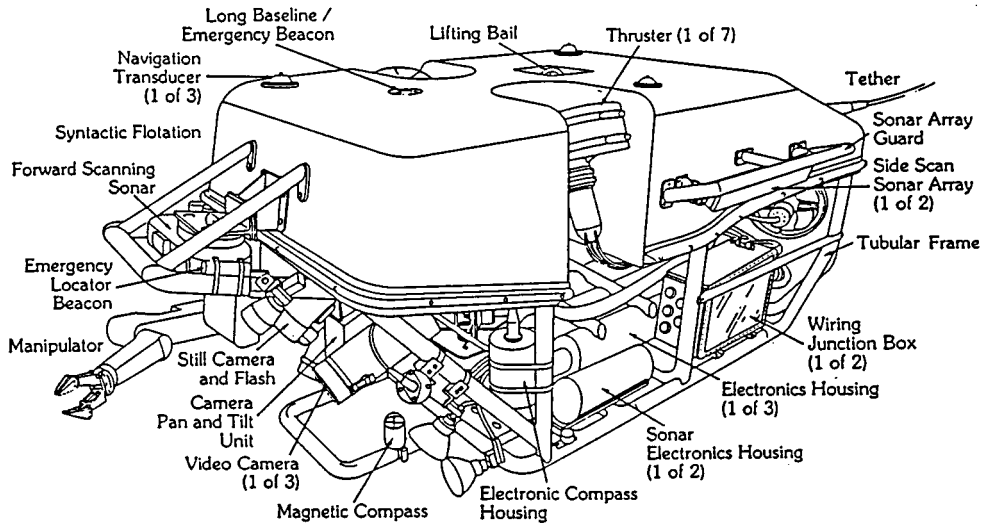


FIGURE 6.1.9 Deep ocean submersible *Jason*. (Courtesy of Woods Hole Oceanographic Institution.)

Common Criteria for Human Interface Design

Design of operator control stations for teleoperators poses the same types of problems as design of controls and displays for aircraft, highway vehicles, and trains. The displays must show the important variables unambiguously to whatever accuracy is required, but more than that must show the variables in relation to one another so as to clearly portray the current "situation" (situation awareness is currently a popular test of the human operator in complex systems). Alarms must get the operator's attention, indicate by text, symbol, or location on a graphic display what is abnormal, where in the system the failure occurred, what is the urgency, if response is urgent, and even suggest what action to take. (For example, the ground-proximity warning in an aircraft gives a loud "Whoop, whoop!" followed by a distinct spoken command "Pull up, pull up!") Controls — whether analogic joysticks, master-arms, or knobs — or symbolic special-purpose buttons or general-purpose keyboards — must be natural and easy to use, and require little memory of special procedures (computer icons and windows do well here). The placement of controls and instruments and their mode and direction of operation must correspond to the desired direction and magnitude of system response.

Human Workload and Human Error

As noted above, new technology allows combination, integration, and simplification of displays compared to the intolerable plethora of separate instruments in older aircraft cockpits and plant control rooms. The computer has taken over more and more functions from the human operator. Potentially these changes make the operator's task easier. However, it also allows for much more information to be presented, more extensive advice to be given, etc.

These advances have elevated the stature of the human operator from providing both physical energy and control, to providing only continuous control, to finally being a supervisor or a robotic vehicle or

system. Expert systems can now answer the operator's questions, much as does a human consultant, or whisper suggestions in his ear even if he doesn't request them. These changes seem to add many cognitive functions that were not present at an earlier time. They make the operator into a monitor of the automation, who is supposed to step in when required to set things straight. Unfortunately, people are not always reliable monitors and interveners.

Mental Workload

Under such complexity it is imperative to know whether or not the mental workload of the operator is too great for safety. Human-machine systems engineers have sought to develop measures of mental workload, the idea being that as mental load increases, the risk of error increases, but presumably measurable mental load comes before actual lapse into error.

Three approaches have been developed for measuring mental workload:

1. The first and most used is the subjective rating scale, typically a ten-level category scale with descriptors for each category from no load to unbearable load.
2. The second approach is use of physiological indexes which correlate with subjective scales, including heart rate and the variability of heart rate, certain changes in the frequency spectrum of the voice, electrical resistance of the skin, diameter of the pupil of the eye, and certain changes in the evoked brain wave response to sudden sound or light stimuli.
3. The third approach is to use what is called a secondary task, an easily measurable additional task which consumes all of the operator's attention remaining after the requirements of the primary task are satisfied. This latter technique has been used successfully in the laboratory, but has shortcomings in practice in that operators may refuse to cooperate.

Such techniques are now routinely applied to critical tasks such as aircraft landing, air traffic control, certain planned tasks for astronauts, and emergency procedures in nuclear power plants. The evidence suggests that supervisory control relieves mental load when things are going normally, but when automation fails the human operator is subjected rapidly to increased mental load.

Human Error

Human error has long been of interest, but only in recent decades has there been serious effort to understand human error in terms of categories, causation, and remedy. There are several ways to classify human errors. One is according to whether it is an error of *omission* (something not done which was supposed to have been done) or *commission* (something done which was not supposed to have been done). Another is *slip* (a correct intention for some reason not fulfilled) vs. a *mistake* (an incorrect intention which was fulfilled). Errors may also be classified according to whether they are in sensing, perceiving, remembering, deciding, or acting. There are some special categories of error worth noting which are associated with following procedures in operation of systems. One, for example, is called a *capture error*, wherein the operator, being very accustomed to a series of steps, say, A, B, C, and D, intends at another time to perform E, B, C, F. But he is "captured" by the familiar sequence B, C and does E, B, C, D.

As to effective therapies for human error, proper design to make operation easy and natural and unambiguous is surely the most important. If possible, the system design should allow for error correction before the consequences become serious. Active warnings and alarms are necessary when the system can detect incipient failures in time to take such corrective action. Training is probably next most important after design, but any amount of training cannot compensate for an error-prone design. Preventing exposure to error by guards, locks, or an additional "execute" step can help make sure that the most critical actions are not taken without sufficient forethought. Least effective are written warnings such as posted decals or warning statements in instruction manuals, although many tort lawyers would like us to believe the opposite.

Trust, Alienation, and How Far to Go with Automation

Trust

If operators do not trust their sensors and displays, expert advisory system, or automatic control system, they will not use it or will avoid using it if possible. On the other hand, if operators come to place too much trust in such systems they will let down their guard, become complacent, and, when it fails, not be prepared. The question of operator trust in the automation is an important current issue in human-machine interface design. It is desirable that operators trust their systems, but it is also desirable that they maintain alertness, situation awareness, and readiness to take over.

Alienation

There is a set of broader social effects that the new human-machine interaction can have, which can be discussed under the rubric of *alienation.*

1. People worry that computers can do some tasks much better than they themselves can, such as memory and calculation. Surely, people should not try to compete in this arena.
2. Supervisory control tends to make people remote from the ultimate operations they are supposed to be overseeing — remote in space, desynchronized in time, and interacting with a computer instead of the end product or service itself.
3. People lose the perceptual-motor skills which in many cases gave them their identity. They become "deskilled", and, if ever called upon to use their previous well-honed skills, they could not.
4. Increasingly, people who use computers in supervisory control or in other ways, whether intentionally or not, are denied access to the knowledge to understand what is going on inside the computer.
5. Partly as a result of factor 4, the computer becomes mysterious, and the untutored user comes to attribute to the computer more capability, wisdom, or blame than is appropriate.
6. Because computer-based systems are growing more complex, and people are being “elevated” to roles of supervising larger and larger aggregates of hardware and software, the stakes naturally become higher. Where a human error before might have gone unnoticed and been easily corrected, now such an error could precipitate a disaster.
7. The last factor in alienation is similar to the first, but all-encompassing, namely, the fear that a “race” of machines is becoming more powerful than the human race.

These seven factors, and the fears they engender, whether justified or not, must be reckoned with. Computers must be made to be not only “human friendly” but also not alienating with respect to these broader factors. Operators and users must become computer literate at whatever level of sophistication they can deal with.

How Far to Go with Automation

There is no question but that the trend toward supervisory control is changing the role of the human operator, posing fewer requirements on continuous sensory-motor skill and more on planning, monitoring, and supervising the computer. As computers take over more and more of the sensory-motor skill functions, new questions are being raised regarding how the interface should be designed to provide the best cooperation between human and machine. Among these questions are: To what degree should the system be automated? How much “help” from the computer is desirable? What are the points of diminishing returns?

Table 6.1.1 lists ten levels of automation, from 0 to 100% computer control. Obviously, there are few tasks which have achieved 100% computer control, but new technology pushes relentlessly in that direction. It is instructive to consider the various intermediate levels of Table 6.1.1 in terms not only of how capable and reliable is the technology but what is desirable in terms of safety and satisfaction of the human operators and the general public.

TABLE 6.1.1 Scale of Degrees of Automation

1.	The computer offers no assistance; the human must do it all.
2.	The computer offers a complete set of action alternatives, and
3.	Narrows the selection down to a few, or
4.	Suggests one alternative, and
5.	Executes that suggestion if the human approves, or
6.	Allows the human a restricted time to veto before automatic execution, or
7.	Executes automatically, then necessarily informs the human, or
8.	Informs the human only if asked, or
9.	Informs the human only if it, the computer, decides to
10.	The computer decides everything and acts autonomously, ignoring the human.
	The current controversy about how much to automate large commercial transport aircraft is often couched in these terms

Source: Sheridan 1987. With permission.

6.2 The Need for Control of Mechanical Systems

Peter S. Curtiss

Process control typically involves some mechanical system that needs to be operated in such a fashion that the output of the system remains within its design operating range. The objective of a process control loop is to maintain the process at the set point under the following dynamic conditions:

- The set point is changed;
- The load on the process is changed;
- The transfer function of the process is changed or a disturbance is introduced.

The Classical Control System Representation

Feedback-Loop System. A *feedback* (or *closed-loop*) system contains a process, a sensor and a controller. Figure 6.2.1 below shows some of the components and terms used when discussing feedback loop systems.

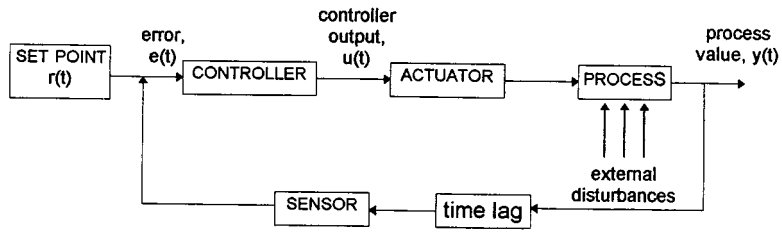


FIGURE 6.2.1 Typical feedback control schematic diagram.

Process. A *process* is a system that produces a motion, a temperature change, a flow, a pressure, or many other actions as a function of the actuator position and external inputs. The output of the process is called the process value. If a positive action in the actuator causes an increase in the process value then the process is called *direct acting*. If positive action in the actuator decreases the process value, it is called *reverse acting*.

Sensor. A *sensor* is a pneumatic, fluidic, or electronic or other device that produces some kind of signal indicative of the process value.

Set Point. The *set point* is the desired value for a process output. The difference between the set point and the process value is called the *process error*.

Controller. A *controller* sends signals to an actuator to effect changes in a process. The controller compares the set point and the process value to determine the process error. It then uses this error to adjust the output and bring the process back to the set point. The controller *gain* dictates the amount that the controller adjusts its output for a given error.

Actuator. An *actuator* is a pneumatic, fluidic, electric, or other device that performs any physical action that will control a process.

External Disturbances. An *external disturbance* is any effect that is unmeasured or unaccounted for by the controller.

Time Constants. The *time constant* of a sensor or process is a quantity that describes the dynamic response of the device or system. Often the time constant is related to the mass of an object or other dynamic effect in the process. For example, a temperature sensor may have a protective sheath around

it that must first be warmed before the sensor registers a change of temperature. Time constant can range from seconds to hours.

Dead Time. The *dead time* or *lag time* of a process is the time between the change of a process and the time this change arrives at the sensor. The delay time is not related to the time constant of the sensor, although the effects of the two are similar. Large dead times must be properly treated by the control system to prevent unstable control.

Hysteresis. *Hysteresis* is a characteristic response of positioning actuators that results in different positions depending on whether the control signal is increasing or decreasing.

Dead Band. The *dead band* of a process is that range of the process value in which no control action is taken. A dead band is usually used in two-position control to prevent "chattering" or in split-range systems to prevent sequential control loops from fighting each other.

Control Point. The *control point* is the actual, measured value of a process (i.e., the set point + steady-state offset + compensation).

Direct/Reverse Action. A *direct-acting* process will increase in value as the signal from the controller increases. A *reverse-acting* process will decrease in value as the signal from the controller increases.

Stability. The *stability* of a feedback control loop is an indication of how well the process is controlled or, alternatively, how controllable the process is. The stability is determined by any number of criteria, including overshoot, settling time, correction of deviations due to external disturbances, etc.

Electric Control. *Electric control* is a method of using low voltages (typically, 24 VAC) or line voltages (110 VAC) to measure values and effect changes in controlled variables.

Electronic Control. *Electronic controls* use solid-state, electronic components used for measurement and amplification of measured signals and the generation of proportional control signals.

Pneumatic Control. *Pneumatic controls* use compressed air as the medium for measuring and controlling processes.

Open-Loop Systems. An *open-loop system* is one in which there is no feedback. A whole-house attic fan in an example. It will continue to run even though the house may have already cooled off. Also, timed on/off devices are open loops.

Examples

Direct-Acting Feedback Control. A classic control example is a reservoir in which the fluid must be maintained at a constant level. Figure 6.2.2 shows this process schematically. The key features of this direct-acting system are labeled. We will refer to the control action of this system shortly after defining some terms.

Cascaded (Master-Slave) Control Loops. If a process consists of several subprocesses, each with a relatively different transfer function, it is often useful to use cascaded control loops. For example, consider a building housing a manufacturing line in which 100% outside air is used but which must also have very strict control of room air temperature. The room temperature is controlled by changing the position of a valve on a coil at the main air-handling unit that supplies the zone. Typically, the time constant of the coil will be much smaller than the time constant of the room. A single feedback loop would probably result in poor control since there is so much dead time involved with both processes. The solution is to use two controllers: the first (the master) compares the room temperature with the thermostat setting and sends a signal to the second (the slave) that uses that signal as its own set point for controlling the coil valve. The slave controller measures the output of the coil, not the temperature of the room. The controller gain on the master can be set lower than that of the slave to prevent excessive cycling.

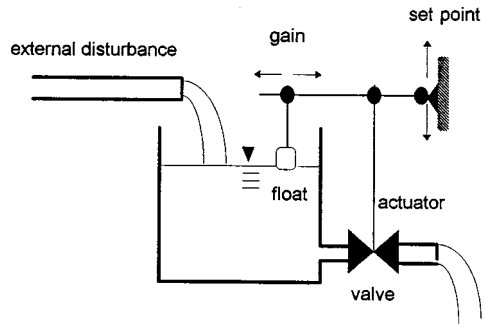


FIGURE 6.2.2 Example of a controlled process.

Sequential Control Loops. Sometimes control action is needed at more than one point in a process. An example of this is an air-handling unit that contains both heating and cooling coils in order to maintain a fixed outlet air temperature no matter the season. Typically, a *sequential* (or *split-range*) system in an air-handling unit will have three temperature ranges of operation, the first for heating mode, the last for cooling mode, and a middle dead-band region where neither the cooling nor heating coils are operating. Most sequential loops are simply two different control loops acting from the same sensor. The term *sequential* refers to the fact that in most of these systems the components are in series in the air or water stream.

Combined Feed-Forward/Feedback Loops. As pointed out earlier, feed-forward loops can be used when the effects of an external disturbance on a system are known. An example of this is outside air temperature reset control used to modify supply air temperatures. The control loop contains both a discharge air temperature sensor (the *primary* sensor) and an outdoor air temperature sensor (the *compensation* sensor). The designer should have some idea about the influence of the outside temperature on the heating load, and can then assign an *authority* to the effect of the outside air temperature on the controller set point. As the outdoor temperature increases, the control point decreases, and vice versa, as shown in Figure 6.2.3.

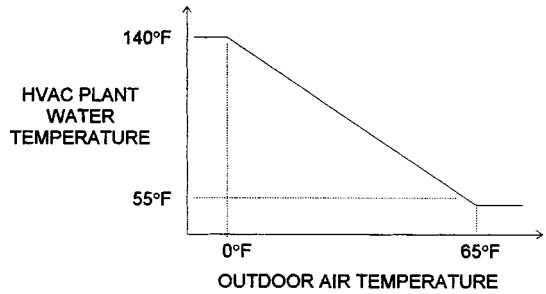


FIGURE 6.2.3 Example of the effect of compensation control.

Predictive Control. *Predictive control* uses a model of the process to predict what the process value will be at some point in the future based upon the current and past conditions. The controller then specifies a control action to be taken at the present that will reduce the future process error.

Adaptive Control. *Adaptive controllers* modify their gains dynamically so to adapt to current process conditions.

Supervisory Controllers. *Supervisory controllers* are used to govern the operation of an entire plant and/or control system. These may be referred to as *distributed control systems* (DCSs) which can be

used to govern the control of individual feedback loops and can also be used to ensure some kind of optimal performance of the entire plant. The controller will vary setpoints and operating modes in an attempt to minimize a cost function. A basic diagram of a supervisory controller in Figure 6.2.4.

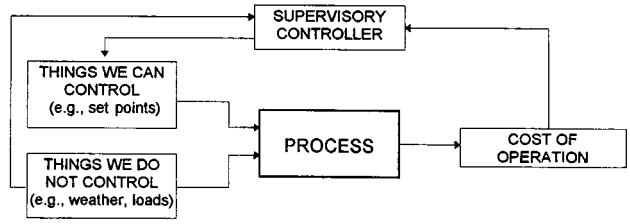


FIGURE 6.2.4 Typical supervisory controller.

6.3 Control System Analysis

Peter S. Curtiss

The Linear Process Approximation

To design controllers it is necessary to have both a dynamic process and control system representation. This section describes the key points of the most common such representation, that of linear processes and their controls. A process is basically a collection of mechanical equipment in which an input is changed or transformed somehow to produce an output. Many processes will be at near-steady-state, while others may be in a more or less constant state of change. We use building control systems as an illustration.

Steady-State Operation

The true response of a seemingly simple process can be, in fact, quite complex. It is very difficult to identify and quantify every single input because of the stochastic nature of life. However, practically any process can be approximated by an equation that takes into account the known input variables and produces a reasonable likeness to the actual process output.

It is convenient to use differential equations to describe the behavior of processes. For this reason, we will denote the "complexity" of the function by the number of terms in the corresponding differential equation (i.e., the *order* or *degree* of the differential equation). In a linear system analysis, we usually consider a step change in the control signal and observe the response. The following descriptions will assume a step input to the function, as shown in Figure 6.3.1. Note that a step change such as this is usually unlikely in most fields of control outside of electronic systems and even then can only be applied to a digital event, such as a power supply being switched on or a relay being energized. Zero-order system output has a one-to-one correspondence to the input,

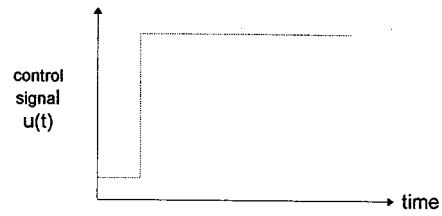


FIGURE 6.3.1 Step change in control signal.

$$y(t) = a_0 \cdot u(t)$$

First-order functions will produce a time-varying output with a step change as input,

$$\frac{dy(t)}{dt} + a_1 \cdot y(t) = b_1 \cdot u(t)$$

and higher-order functions will produce more complex outputs.

The function that relates the process value to the controller input is called the *transfer function* of the process. The time between the application of the step change, t_0, and the time at which the full extent of the change in the process value has been achieved is called the *transfer period*. A related phenomenon is process dead time. If there is a sufficient physical distance between the process output and the sensor assigned to measuring it, then one observes dead time during which the process output is not affected by the control signal (see Figure 6.3.2). The *process gain* (or *static gain*) is the ratio of the percentage

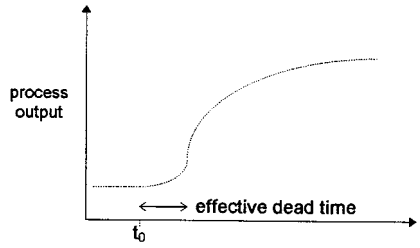


FIGURE 6.3.2 Effective dead time of a process subjected to a step change in controlled signal.

change of the process output to the corresponding percentage change of the control signal for a given response. For example, the gain can be positive (as in a heating coil) or negative (as in a cooling coil).

Dynamic Response

In practice, there are very few processes controlled in a stepwise fashion. Usually, the control signal is constantly modulating much the way that one makes small changes to the steering wheel of a car when driving down the highway. We now consider the dynamic process of level control in buckets filled with water (see Figure 6.3.3). Imagine that the level of water in the bucket on the left of Figure 6.3.3 is the control signal and the level of water in the bucket on the right is the process value. It is obvious that a step change in the control signal will bring about a first-order response of the process value.

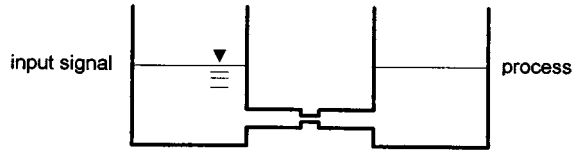


FIGURE 6.3.3 Connected water containers used for example of dynamic response.

Suppose, however, that a periodic signal is applied to the level of the bucket on the left. If the frequency of the signal is small enough, we see a response in the level in the bucket on the right that varies as a function of this driving force, but with a delay and a decrease in the amplitude.

Here the *dynamic process gain* is less than one even though the static process gain is one. There is no dead time in this process; as soon as we begin to increase the control signal the process value will also begin to increase. The dynamic process gain, therefore, can be defined similarly to that of the static gain — it is the ratio of the amplitude of the two signals, comparable with the normalized ranges used in the static gain definition.

The dynamic gain, as its name suggests, is truly dynamic. It will change not only according to the transfer function, but also to the frequency of the control signal. As the frequency increases, the output will lag even farther behind the input and the gain will continue to decrease. At one point, the frequency may be exactly right to cancel any past effects of the input signal (i.e., the phase shift is 180°) and the dynamic gain will approach zero. If the frequency rises further, the process output may decrease as the control signal increases (this can easily be the case with a building cooling or heating coil due to the mass effects) and the dynamic gain will be negative!

At this point it is convenient to define a feedback loop mathematically. A general feedback loop is shown in Figure 6.3.4. The controller, actuator, and process have all been combined into the *forward transfer function* (or *open-loop transfer function*) G and the sensor and dead time have all been combined into the *feedback path* transfer function H. The overall *closed-loop transfer function* is defined as

$$\frac{C}{R} = \frac{C}{1 + G \cdot H}$$

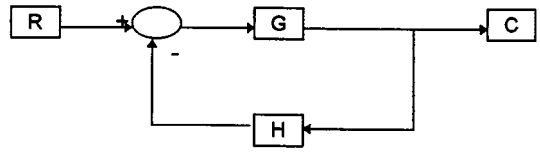


FIGURE 6.3.4 Generalized feedback loop.

The right-hand side of this equation is usually a ratio of two polynomials when using Laplace or *z*-transforms. The roots of the numerator are called the *zeros* of the transfer function and the roots of the denominator are called the *poles* (Shinners, 1978).

The denominator of the closed loop transfer function, $1 + G \cdot H$, is called the *characteristic function.* If we set the characteristic function equal to zero we have the *characteristic equation*

$$1 + G \cdot H = 0$$

The characteristic equation can be used to assess process control stability during system design.

Representation of Processes in *t*, *s*, and *z* Domains

We cannot hope to ever know how a process truly behaves. The world is an inherently stochastic place and any model of a system is going to approximate at best. Nonetheless, we will need to choose some kind of representation in order to perform any useful analysis.

This section will consider three different domains: the continuous-time domain, the frequency domain, and the discrete-time domain. The frequency domain is useful for certain aspects of controller design, whereas the discrete-time domain is used in digital controllers.

Continuous-Time-Domain Representation of a Process

In the time domain we represent a process by a differential equation, such as

$$\frac{d^n y}{dt^n} + a_1 \frac{d^{n-1} y}{dt^{n-1}} + a_2 \frac{d^{n-2} y}{dt^{n-2}} + \cdots + a_{n-1} \frac{dy}{dt} + a_n y = b_0 \frac{d^m u}{dt^m} + b_1 \frac{d^{m-1} u}{dt^{m-1}} + \cdots + b_{m-1} \frac{du}{dt} + b_m u$$

This is just a generalization of the first-order system equation described earlier.

Frequency-Domain Representation of a Process — Laplace Transforms

The solution of higher-order system models, closed-form solution is difficult in the time domain. For this reason, process transfer functions are often written using Laplace transforms. A Laplace transform is a mapping of a continuous-time function to the frequency domain and is defined as

$$F(s) = \int_0^\infty f(t) e^{-st}\, dt$$

Laplace transforms are treated in Section 19. This formulation allows us to greatly simplify problems involving ordinary differential equations that describe the behavior of systems. A transformed differential equation becomes purely algebraic and can be easily manipulated and solved. These solutions are not of great interest in themselves in modern control system design but the transformed system (+) controller differential equation is very useful in assessing control stability. This is the single key aspect of Laplace transforms that is of most interest. Of course, it is possible just to solve the governing differential equation for the system directly and explore stability in that fashion.

The Laplace transform of the previous differential equation is

$$s^n Y(s) + A_1 s^{n-1} Y(s) + \ldots + A_{n-1} s Y(s) + A_n Y(s) = B_0 s^m U(s) + B_1 s^{n-1} U(s) + \ldots + B_{n-1} s U(s) + B_n U(s)$$

This equation can be rewritten as

$$Y(s) \cdot \left(s^n + A_1 s^{n-1} + \ldots + A_{n-1} s + A_n\right) = U(s) \cdot \left(B_0 s^m + B_1 s^{n-1} + \ldots + B_{n-1} s + B_n\right)$$

so that the transfer function is found from

$$\frac{Y(s)}{U(s)} = \frac{s^m + B_1 s^{m-1} + \ldots + B_{m-1} s + A_m}{s^n + A_1 s^{n-1} + \ldots + A_{n-1} s + A_n}$$

This is the expression that is used for stability studies.

Discrete-Time-Domain Representation of a Process. A process in the discrete time domain is described (Radke and Isermann, 1989) by

$$y(k) = a_1 y(k-1) + a_2 y(k-2) + a_3 y(k-3) + \ldots + b_1 u(k-1) + b_2 u(k-2) + b_3 u(k-3) + \ldots$$

This representation is of use when one is designing and analyzing the performance of direct digital control (DDC) systems. Note that the vectors **a** and **b** are *not* the same as for the continuous-time domain equation. The *z*-transform uses the backward shift operator and therefore the *z*-transform of the discrete-time equation is given by

$$y\left(1 - a_1 z^{-1} - a_2 z^{-2} - a_3 z^{-3} + \ldots\right) = u\left(b_1 z^{-1} - b_2 z^{-2} - b_3 z^{-3} + \ldots\right)$$

The transfer function can now be found:

$$\frac{y}{u} = \frac{b_1 z^{-1} - b_2 z^{-2} - b_3 z^{-3} + \ldots}{1 - a_1 z^{-1} - a_2 z^{-2} - a_3 z^{-3} + \ldots}$$

z-Transform Details. Because *z*-transforms are important in modern control design and are not treated elsewhere in this handbook, some basics of their use are given below. More and more control applications are being turned over to computers and DDC systems. In such systems, the sampling is not continuous, as required for a Laplace transform. The control loop schematic is shown in Figure 6.3.5.

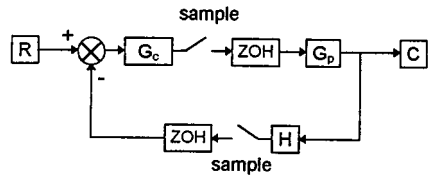


FIGURE 6.3.5 Sampled feedback loop.

It would be prohibitively expensive to include a voltmeter or ohmmeter on each loop; therefore, the controller employs what is called a *zero-order hold*. This basically means that the value read by the controller is "latched" until the next value is read in. This discrete view of the world precludes the use of Laplace transforms for analyses and makes it necessary, therefore, to find some other means of

simplifying the simulation of processes and controllers. The following indicates briefly how *z*-transforms of controlled processes can be derived and how they are used in a controls application. In the design section of this chapter we will use the *z*-transform to assess controller stability.

Recall that the Laplace transform is given as

$$\mathcal{L}\{f(t)\} = \int_0^\infty f(t)e^{-st}\,dt$$

Now suppose we have a process that is sampled at a discrete, constant time interval *T*. The index *k* will be used to count the intervals,

$$\text{at time } t = 0,\ k = 0,$$

$$\text{at time } t = T,\ k = 1,$$

$$\text{at time } t = 2T,\ k = 2,$$

$$\text{at time } t = 3T,\ k = 3,$$

and so forth. The equivalent Laplace transform of a process that is sampled at a constant interval *T* can be represented as

$$\mathcal{L}\{f^*(t)\} = \sum_{k=0}^{\infty} f(kT)e^{-s\cdot kT}$$

By substituting the *backward-shift operator* z for e^{Ts}, we get the definition of the *z*-transform:

$$Z\{f(t)\} = \sum_{k=0}^{\infty} f(kT)z^{-k}$$

Example of Using z-Transfer Functions. Suppose we have a cylindrical copper temperature sensor in a fluid stream with material properties as given. We wish to establish its dynamic characteristics for the purpose of including it in a controlled process model using both Laplace and *z*-transforms. Figure 6.3.6 shows the key characteristics of the sensor. The sensor measures 0.5 cm in diameter and is 2 cm long. For the purposes of this example we will assume that the probe is solid copper. The surface area of the sensor is then

$$A_s = 2\cdot\pi\cdot(0.25\text{ cm})^2 + \pi\cdot(0.5\text{ cm})\cdot(2\text{ cm}) \approx 3.5\text{ cm}^2$$

and the mass is

$$M_s = (9\text{ g/cm}^3)\cdot\left[(2\text{ cm})\cdot\pi\cdot(0.25\text{ cm})^2\right] \approx 3.5\text{ g}$$

The thermal capacitance of the sensor is found from the product of the mass and the heat capacity,

$$C_s = M_s\cdot c_p = 3.5\text{ g}\cdot 0.4\text{ J/g}\cdot\text{K} = 1.4\text{ J/K}$$

and the total surface heat transfer rate is the product of the area and the surface heat transfer coefficient,

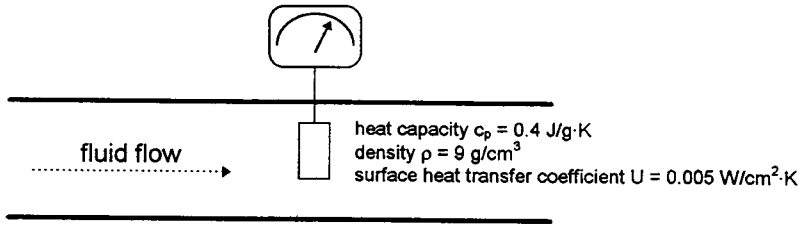


FIGURE 6.3.6 Fluid temperature sensor used in example.

$$UA_s = 0.005\ \mathrm{W/cm^2 \cdot K} \cdot 3.5\ \mathrm{cm^2} = 0.018\ \mathrm{W/K}$$

Now we can perform an energy balance on the sensor by setting the sum of the energy flow into the sensor and the energy stored in the sensor equal to zero:

$$C_s = \frac{dT_s}{dt} + \left(T_s - T_a\right) \cdot UA_s = 0$$

where T_s is the temperature of the sensor and T_a is the ambient fluid temperature. This relationship is a nonhomogeneous first-order differential equation:

$$\frac{dT_s}{dt} + \frac{UA_s}{C_s} T_s = \frac{UA_s}{C_s} T_a$$

The *time constant* of the sensor is defined as

$$\tau = \frac{C_s}{UA_s} = \frac{1.4\ \mathrm{J/K}}{0.018\ \mathrm{W/K}} \approx 80\ \mathrm{sec}$$

The differential equation that describes this sensor is

$$\frac{dT_s}{dt} + \frac{1}{\tau} T_s = \frac{1}{\tau} T_a$$

This example will find the response of the sensor when the fluid temperature rises linearly by 30°C from time t = 0 to time t = 200 seconds and then remains constant. That is, the driving function is

$$T_a(t) = \frac{30°\mathrm{C}}{200\ \mathrm{sec}} t = 0.15°\mathrm{C/sec}\ t \qquad 0 \le t < 200\ \mathrm{sec}$$

$$T_a(t) = 30°\mathrm{C} \qquad t \ge 200\ \mathrm{sec}$$

Assuming an initial condition of T_s = 0 at t = 0, we find that

$$T_s(t) = 0.15t - 12 + 12e^{-0.0125t} \qquad \text{for } t < 200\ \mathrm{sec}$$

and

$$T_s(t) = 30 - 134.2e^{-t/\tau} \qquad \text{for } t > 200\ \mathrm{sec}$$

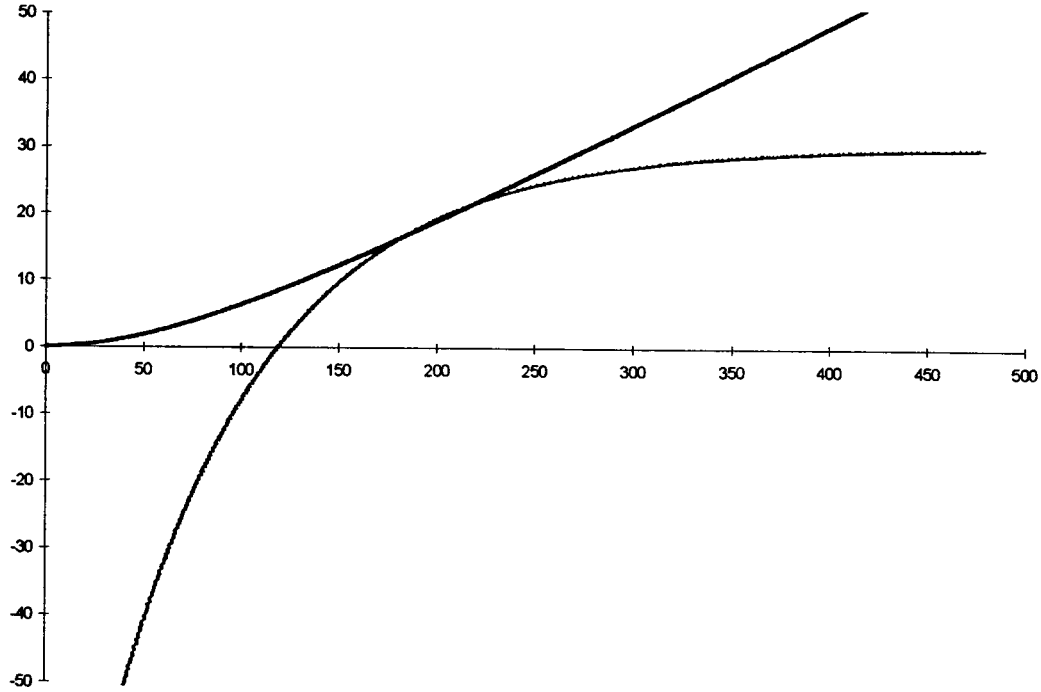


FIGURE 6.3.7 Time domain solution of example.
First line: $T = 0.15t - 12 + 12e^{-0.0125t}$. Second line: $T = 30 - 134.2e^{-t/r}$.

A graph of the entire process (rise time and steady state) is shown in Figure 6.3.7. The two lines intersect at t = 200 seconds.

Solution of example in frequency domain

This same process can be solved using Laplace transforms. First we will solve this problem for the first 200 sec. The Laplace transform of a ramp function can be found from the tables (see Section 19) to get

$$T_s(s) = \frac{1}{\tau(s + 1/\tau)} \frac{0.15}{s^2}$$

This can be expanded by partial fractions to get

$$T_s(s) = 0.15\left[\frac{\tau}{s} + \frac{1}{s^2} + \frac{\tau}{s + 1/\tau}\right]$$

Using the table for the inverse Laplace transforms gives

$$T_s(t) = 0.15\left[-\tau + t + \tau e^{-t/\tau}\right]$$

Substituting τ = 80,

$$T_s(t) = -12 + 0.15t + 12e^{-0.0125t}$$

This is exactly the same as the time-domain solution above. For the steady-state driving force of T_a = 30 for t > 200 sec we also find the same result.

Solution of example in the discrete-time domain

We consider the same problem in the discrete-time domain that a DDC system might use. Recall that the transfer function in the frequency domain was given by

$$\frac{T_s(s)}{T_a(s)} = \frac{\dfrac{1}{\tau}}{s + \dfrac{1}{\tau}}$$

We look to the table in this book's appendix and find that the discrete-time equivalent is

$$\frac{T_s(z)}{T_a(z)} = \frac{1}{\tau} \cdot \frac{z}{z - e^{-T/\tau}}$$

where T is the sampling frequency in seconds. The driving function T_a is a given as

$$T_a(t) = \alpha t$$

where α is the rate of change of the temperature as above (0.15°C/sec). Note that to put this into the discrete-time domain we must correct this rate by the sampling interval,

$$T_a(z) = 0.15 \cdot T \cdot \frac{Tz}{(z-1)^2}$$

using α = (0.15 · T) ºC/sampling interval.

We can now express the response of the process as

$$T_s(z) = \frac{1}{\tau} \cdot \left(\frac{z}{z - e^{-T/\tau}} \right) \cdot (0.15 \cdot T) \cdot \left(\frac{Tz}{(z-1)^2} \right)$$

since the z operator acts on the sensed temperature by performing a backward shift of the time index. In other words, the previous equation can be rewritten as

$$T_{s,k} - \left(2 + e^{-T/\tau}\right) T_{s,k-1} + \left(1 + 2e^{-T/\tau}\right) T_{s,k-2} - e^{-T/\tau} T_{s,k-3} = \frac{0.15}{\tau} \cdot T^2 z^{-1}$$

So the current temperature is determined by the previous three temperature measurements,

$$T_{s,k} = \left(2 + e^{-T/\tau}\right) T_{s,k-1} - \left(1 + 2e^{-T/\tau}\right) T_{s,k-2} + e^{-T/\tau} T_{s,k-3} + \frac{0.15}{\tau} \cdot T^2 z^{-1} \qquad \text{for } kT < 200 \text{ sec}$$

Regarding the last term in the equation, recall that the inverse transform of z^{-k} is given as 1 when $t = k$, and zero otherwise. This term provides the initial "jump-start" of the progression. Table 6.3.1 shows the first few time steps for the z-domain solution using time steps of 0.1 and 1.0 seconds. In general, the accuracy of the z-domain solution increases as the time step grows smaller. The solution for kT > 200 seconds is similar to that for the Laplace transforms.

TABLE 6.3.1 Initial Time Steps for *z*-Domain Solution of Example

Initial 10 Steps for T = 0.1 sec				Initial 10 Steps for T = 1.0 sec			
k	**Time**	$T_{s,\text{exact}}$	$T_{s,z\ \text{transform}}$	**k**	**Time**	$T_{s,\text{exact}}$	$T_{s,z\ \text{transform}}$
0	0.00	0.00000	0.00000	0	0.00	0.0000	0.0000
1	0.10	0.00001	0.00002	1	1.00	0.0009	0.0019
2	0.20	0.00004	0.00006	2	2.00	0.0037	0.0056
3	0.30	0.00008	0.00011	3	3.00	0.0083	0.0112
4	0.40	0.00015	0.00019	4	4.00	0.0148	0.0185
5	0.50	0.00023	0.00028	5	5.00	0.0230	0.0277
6	0.60	0.00034	0.00039	6	6.00	0.0329	0.0386
7	0.70	0.00046	0.00052	7	7.00	0.0446	0.0512
8	0.80	0.00060	0.00067	8	8.00	0.0580	0.0656
9	0.90	0.00076	0.00084	9	9.00	0.0732	0.0816
10	1.00	0.00093	0.00103	10	10.00	0.0900	0.0994

The next three figures show the effect of using different time intervals in the *z*-domain solution. The values shown here are for the example outlined in this section. If one could use an infinitesimally small time step, the *z*-domain solution would match the exact solution. Of course, this would imply a much larger computational effort to simulate even a small portion of the process. In practice, a time interval will be chosen that reflects a compromise between accuracy and speed of calculation.

Each graph shows two lines, one for the exact solution of the first 200 sec of the example and the other for the *z*-domain solution. Figures 6.3.8 to 6.3.10 give the *z*-domain solution using time intervals of 0.1, 1.0 and 10.0 seconds, respectively. Notice that there is not much difference between the first two graphs even though there is an order of magnitude difference between the time intervals used. The latter two graphs show significant differences.

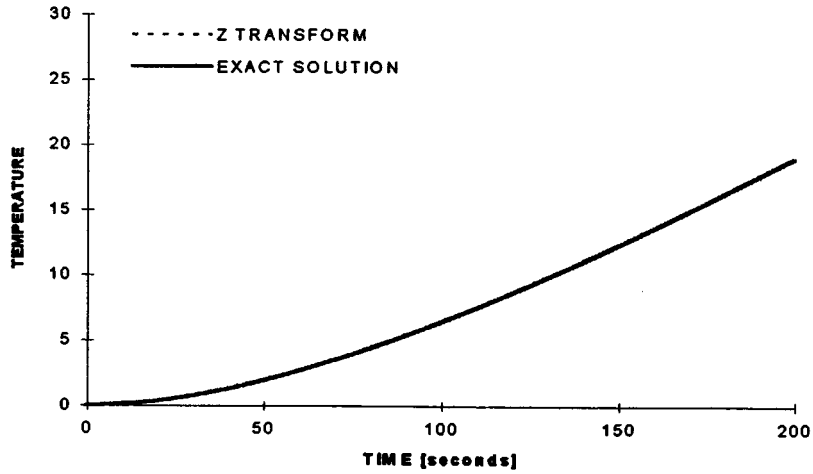


FIGURE 6.3.8 Result of *z*-transform when T = 0.1 sec.

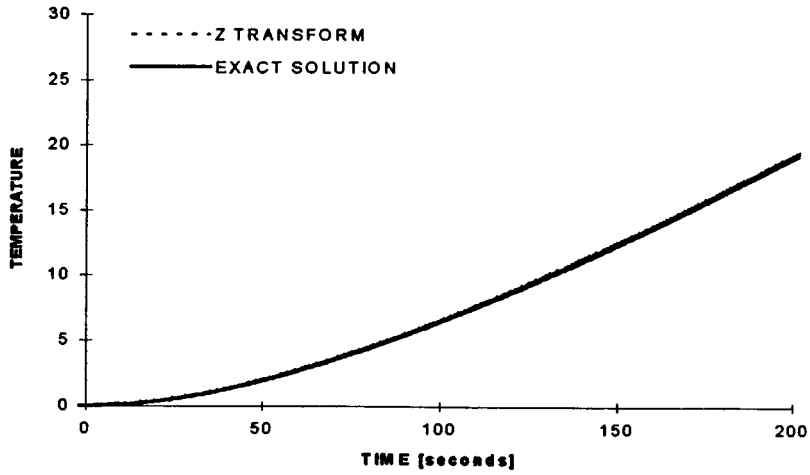


FIGURE 6.3.9 Results of *z*-transform when $T = 1.0$ sec.

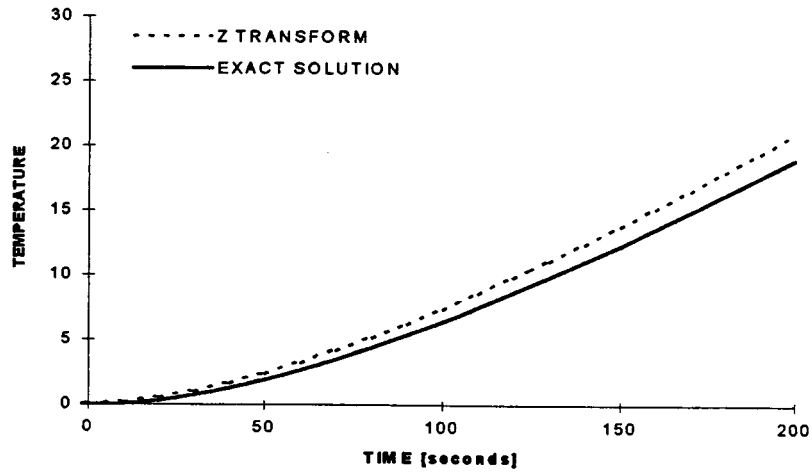


FIGURE 6.3.10 Results of *z*-transform when $T = 10.0$ sec.

6.4 Control System Design and Application

Peter S. Curtiss

Controllers

Controllers are akin to processes in that they have gains and transfer functions. Generally, there is no dead time in a controller or it is so small as to be negligible.

Steady-State Effects of Controller Gain

Recall that the process static gain can be viewed as the total change in the process value due to a 100% change in the controller output. A proportional controller acts like a multiplier between an *error signal* and this process gain. Under stable conditions, therefore, there must be some kind of error to yield any controller output. This is called the steady-state or static *offset*.

Dynamic Effects of Controller Gain

Ideally, a controller gain value is chosen that compensates for the dynamic gain of the process under normal operating conditions. The total loop dynamic gain can be considered as the product of the process, feedback, and controller gains. If the total dynamic loop gain is one, the process will oscillate continuously at the natural frequency of the loop with no change in amplitude of the process value. If the loop gain is greater than one, the amplitude will increase with each cycle until the limits of the controller or process are reached or until something fails. If the dynamic loop gain is less than one, the process will eventually settle down to stable control.

Controller Bias

The controller bias is a constant offset applied to the controller output. It is the output of the controller if the error is zero,

$$u = K \cdot e + M$$

where M is the bias. This is useful for processes that become nonlinear at the extremes or for process in which the normal operating conditions are at a nonzero controller output.

PID Controllers

Many mechanical systems are controlled by proportional-integral-derivative (PID) controllers. There are many permutations of such controllers which use only certain portions of the PID controllers or use variations of this kind of controller. In this section we consider this very common type of controller.

Proportional Control

Proportional control results in action that is linear with the error (recall the error definition in Figure 6.2.1) The proportional term, $K_p \cdot e$, has the greatest effect when the process value is far from the desired setpoint. However, very large values of K_p will tend to force the system into oscillatory response. The proportional gain effect of the controller goes to zero as the process approaches set point. Purely proportional control should therefore only be used when

- The time constant of the process is small and hence a large controller gain can be used;
- The process load changes are relatively small so that the steady-state offset is limited;
- The steady-state offset is within an acceptable range.

Integral Control

Integral control makes a process adjustment based on the cumulative error, not its current value. The integral term K_i is the reciprocal of the reset time, T_r, of the system. The reset time is the duration of

each error-summing cycle. Integral control can cancel any steady-state offsets that would occur when using purely proportional control. This is sometimes called *reset* control.

Derivative Control

Derivative control makes a process adjustment based on the current rate of change of the process control error. Derivative control is typically used in cases where there is a large time lag between the controlled device and the sensor used for the feedback. This term has the overall effect of preventing the actuator signal from going too far in one direction or another, and can be used to limit excessive overshoot.

PID Controller in Time Domain

The PID controller can be represented in a variety of ways. In the time domain, the output of the controller is given by

$$u(t) = K_p\left[e(t) + K_i\int_0^t e(t)\,dt + K_d\frac{de(t)}{dt}\right]$$

PID Controller in the *s* Domain

It is relatively straightforward to derive the Laplace transform of the time-domain PID equation. The transfer function of the controller is

$$\frac{U(s)}{E(s)} = \left[K_p + \frac{K_pK_i}{s} + K_pK_ds\right]$$

This controller transfer function can be multiplied by the process transfer function to yield the overall forward transfer function **G** of an *s*-domain process model. The criteria described earlier can then be used to assess overall system stability.

PID Controller in the *z* Domain

Process data are measured discretely at time intervals Δt, and the associated PID controller can be represented by

$$u(k) = K_p\left[e(k) + K_i\,\Delta t\sum_{i=0}^{k} e(i) + K_d\frac{e(k)-e(k-1)}{\Delta t}\right]$$

The change of the output from one time step to the next is given by $u(k) - u(k-1)$, so the PID *difference equation* is

$$u(k)-u(k-1) = K_p\left[\left(1+\frac{K_d}{\Delta t}\right)e(k) + \left(K_i\,\Delta t - 1 - 2\frac{K_d}{\Delta t}\right)e(k-1) + \left(\frac{K_d}{\Delta t}\right)e(k-2)\right]$$

and can be simplified as

$$u(k)-u(k-1) = q_0e(k) + q_1e(k-1) + q_2e(k-2)$$

where

$$q_0 = K_p\left(1+\frac{K_d}{\Delta t}\right);\qquad q_1 = K_p\left(K_i\,\Delta t - 1 - 2\frac{K_d}{\Delta t}\right);\qquad q_2 = K_p\left(\frac{K_d}{\Delta t}\right)$$

Note that we can write this as

$$u\left(1-z^{-1}\right)=e\left(q_0+q_1z^{-1}+q_2z^{-2}\right)$$

The *z*-domain transfer function of the PID controller is then given as

$$\frac{u(z)}{e(z)}=\frac{q_0+q_1z^{-1}+q_2z^{-2}}{1-z^{-1}}=\frac{q_0z^2+q_1z+q_2}{z^2-z}$$

Controller Performance Criteria and Stability

Performance Indexes

Obviously, in feedback loops we wish to reduce the process error quickly and stably. The control systems engineer can use different cost functions in the design of a given controller depending on the criteria for the controlled process. Some of these cost functions (or *performance indexes*) are listed here:

ISE	Integral of the square of the error	$\int^2 e$
ITSE	Integral of the time and the square of the error	$\int e^2$
ISTAE	Integral of the square of the time and the absolute error	$\int^2 \lvert e\rvert$
ISTSE	Integral of the square of the time and the square of the error	$\int^2 e^2$

These indexes are readily calculated with DDC systems and can be used to compare the effects of different controller settings, gains, and even control methods.

Stability

Stability in a feedback loop means that the feedback loop will tend to converge on a value as opposed to exhibiting steady-state oscillations or divergence. Recall that the closed-loop transfer function is given by

$$\frac{C}{R}=\frac{G}{1+GH}$$

and that the denominator, 1 + *GH*, when equated to zero, is called the characteristic equation. Typically, this equation will be a polynomial in *s* or *z* depending on the method of analysis of the feedback loop. Two necessary conditions for stability are that all powers of *s* must be present in the characteristic equation from zero to the highest order and that all coefficients in the characteristic equation must have the same sign. Note that the process may still be unstable even when these conditions are satisfied.

Roots of the Characteristic Equation. The roots of the characteristic equation play an important role in determining the stability of a process. These roots can be real and/or imaginary and can be plotted as shown in Figure 6.4.1. In the *s*-domain, if all the roots are in the left half-plane (i.e., to the left of the imaginary axis), then the feedback loop is guaranteed to be asymptotically stable and will converge to a single output value. If one or more roots are in the right half-plane, then the process is unstable. If one or more roots lie on the imaginary axis and none are in the right half-plane, then the process is considered to be marginally stable. In the *z*-domain, if all the roots lie within the unit circle about the origin then the feedback loop is asymptotically stable and will converge. If one or more roots lie outside

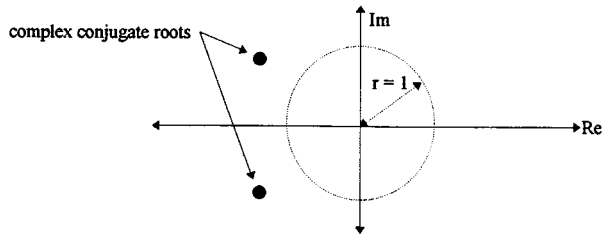


FIGURE 6.4.1 Placement of roots in the imaginary plane (showing unit circle).

the unit circle then the process is unstable. If one or more roots lie on the unit circle and none are outside the unit circle, then the process is marginally stable.

Root locus example

Consider the feedback loop shown in Figure 6.4.2. The characteristic equation is given by $1 + GH = 0$ or

$$1 + K\left(\frac{s}{s+\alpha}\right)\left(\frac{1}{s+\beta}\right) = 0$$

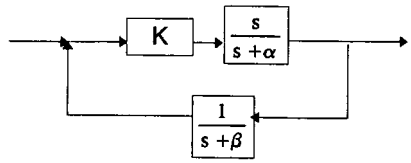


FIGURE 6.4.2 Simple feedback control loop.

For different values of K we can plot the roots of this equation. The graph in Figure 6.4.3 shows an example plot when the characteristic equation is given by $s^2 + (1.25 + K)s + 1.25 = 0$. The plot shows that a system described by this characteristic demonstrates stable response for a process gain of $0.0 \leq K \leq 10.0$. For gains greater than 10, there exists at least one root in the right half-plane and the process is not under stable control.

Note that the root locus plot is always symmetric about the real axis and that the number of separate segments of the locus is equal to the number of roots of the characteristic equation (i.e., the number of poles of the closed-loop transfer function).

Routh-Hurwitz Stability Criteria. The Routh-Hurwitz method is an tabular manipulation of the characteristic equation in the frequency domain and is used to assess stability. If the characteristic equation is given by

$$a_0 s^n + a_1 s^{n-1} + \ldots + a_{n-1} s + a_n = 0$$

then the Routh-Hurwitz method constructs a table from the coefficients as follows:

$$\begin{array}{ccccc} s^n & a_0 & a_2 & a_4 & \ldots \\ s^{n-1} & a_1 & a_3 & a_5 & \ldots \\ s^{n-2} & X_1 & X_2 & X_3 & \ldots \\ s^{n-3} & Y_1 & Y_2 & Y_3 & \ldots \\ \vdots & \vdots & \vdots & \vdots & \ldots \end{array}$$

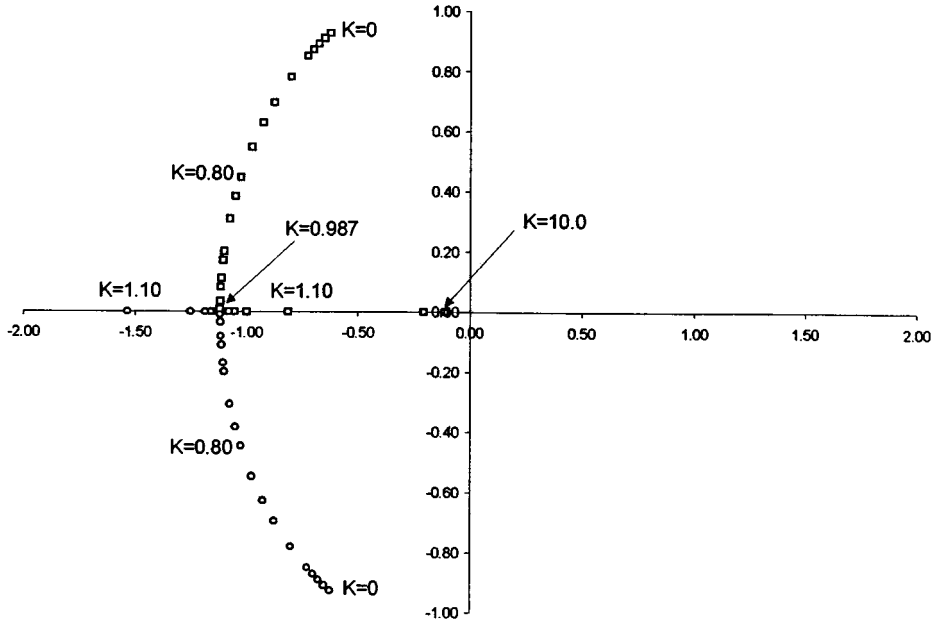


FIGURE 6.4.3 Root locus of $s^2 + (1.25 + K)s + 1.25 = 0$.

where

$$X_1 = \frac{a_1a_2 - a_0a_3}{a_1}; \qquad X_2 = \frac{a_1a_4 - a_0a_5}{a_1}; \qquad X_3 = \frac{a_1a_6 - a_0a_7}{a_1}\cdots$$

$$Y_1 = \frac{X_1a_3 - a_1X_2}{X_1}; \qquad Y_2 = \frac{X_1a_5 - a_1X_3}{X_1}\cdots$$

and so forth. The number of roots in the right-hand plane of the *s*-domain is equal to the number of sign changes in the first column, i.e., the column containing a_0, a_1, X_1, Y_1, etc. In other words, if all the elements in the first column have the same sign, then there are no roots in the right-hand plane and the process is stably controlled. Also, for special cases of the characteristic equation,

- If the first element of any row is zero but the remaining elements are not, then use some small value ε and interpret the final results as $\varepsilon \to 0$.
- If one of the rows before the final row is entirely zeros, then (1) there is at least one pair of real roots of equal magnitude but opposite signs, or (2) there is at least one pair of imaginary roots that lie on the imaginary axis, or (3) there are complex roots symmetric about the origin.

Field Commissioning — Installation, Calibration, Maintenance

Tuning of Feedback Loops

The *tuning* of a controller involves finding controller gains that will ensure at least a critically damped response of the process to a change in set point or process disturbance. A good starting point for PID constants is that derived during the design phase by the stability assessment approaches described above. However, real processes do not necessarily behave as their models would suggest and actual field tuning of controls is needed during the system-commissioning process.

Pole-Zero Cancellation. One method of obtaining the desired critically damped response of a process is to determine the closed-loop transfer function in the form

$$\frac{C}{R} = \frac{(s + A_1)(s + A_2)\ldots(s + A_m)}{(s + B_1)(s + B_2)\ldots(s + B_n)}$$

The coefficients *A* and *B* will depend on both the process characteristics and the controller gains. The objective of pole-zero cancellation is to find values for the controller gains that will set some numerator coefficients equal to those in the denominator, effectively canceling terms. As can be imagined, however, this can be a very difficult exercise, particularly when working with complex roots of the equations. This method can only be used with very simple system models.

Reaction Curve Techniques. Often it is advisable to test a feedback loop *in situ*. Several techniques have been developed that allow for the derivation of "good" PID constants for a given open-loop response. Consider the process response shown in Figure 6.4.4 where Δ_c is the change of process output, Δ_u is the change of controller, *L* is the time between change and intersection, and *T* is the time between lower intersection and upper intersection. We can define the following variables: $A = \Delta_u/\Delta_c$, $B = T/L$, and $R = L/T$. These values can be used with the equations given in Table 6.4.1 to estimate "decent" control constants. The users of these constants should be aware, however, that these constants are based on the typical response of second-order systems and may not provide good values for all processes.

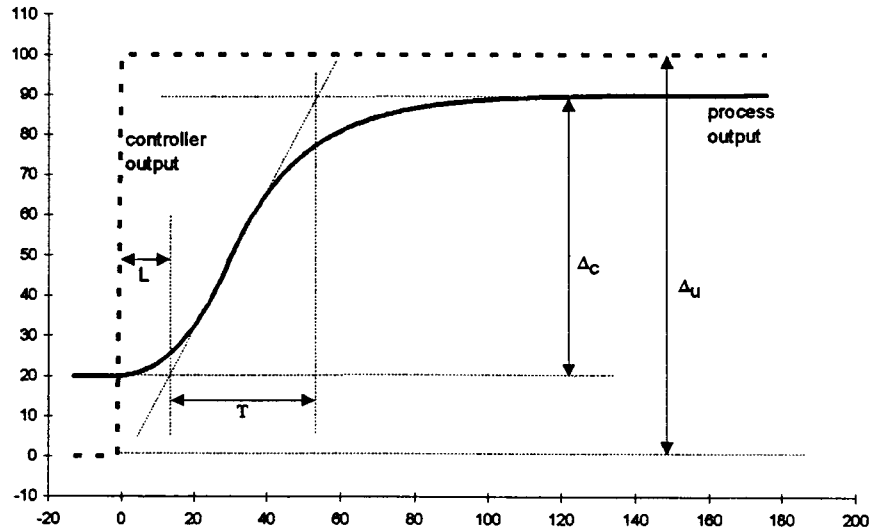


FIGURE 6.4.4 Reaction curve components.

Ultimate Frequency. The ultimate frequency test involves increasing the proportional gain of a process until it begins steady-state oscillations. K_p^* is defined as the proportional gain that results in steady oscillations of the controlled system and T^* is the period of the oscillations. The desired controller gains are given in Table 6.4.2. Note that the use of the ultimate period test is not always easy to do in practice and may be prohibited in certain cases by the a process operations manager.

TABLE 6.4.1 Equations for Finding PID Constants Using the Zeigler-Nichols and Cohen and Coon Reaction Curve Tests

	Zeigler-Nichols			Cohen and Coon		
Controller Components	K_p	$\frac{K_p}{K_i}$	$\frac{K_d}{K_p}$	K_p	$\frac{K_p}{K_i}$	$\frac{K_d}{K_p}$
P	AB	—	—	$AB\left(1+\frac{R}{3}\right)$	—	—
P + I	$0.9AB$	$3.3L$	—	$AB\left(1.1+\frac{R}{12}\right)$	$L\frac{30+3R}{9+20R}$	—
P + D	—	—	—	$AB\left(1.25+\frac{R}{6}\right)$	—	$L\frac{6-2R}{22+3R}$
P + I + D	$1.2AB$	$2L$	$0.5L$	$AB\left(1.33+\frac{R}{4}\right)$	$L\frac{32+6R}{13+8R}$	$L\frac{4}{11+2R}$

TABLE 6.4.2 Equations for Estimating PID constants Using the Ultimate Frequency Test

Controller Components	K_p	$\frac{K_p}{K_i}$	$\frac{K_d}{K_p}$
P	$0.5K_p^*$	—	—
P + I	$0.45K_p^*$	$0.8T^*$	—
P + I + D	$0.6K_p^*$	$0.5T^*$	$0.125T^*$

6.5 Advanced Control Topics

Peter S. Curtiss, Jan Kreider, Ronald M. Nelson, and Shou-Heng Huang

Neural Network-Based Predictive/Adaptive Controllers

Neural networks are powerful modeling tools used for predicting nonlinear behavior of processes and require a minimum of knowledge about the physical system involved. This approach can be used to predict the behavior of a process and can calculate the future value of the process variables. The effects of current modifications to the future value of the controlled process can be easily quantified and used to obtain the desired process response.

Overview of Neural Networks

The artificial neural network attempts to mimic a few aspects of the behavior of biological neural networks. Inputs to a biological nerve cell are carried along the dendrites of that cell. These inputs come from the positive impulse signals of other cells but may be converted to negative signals by the chemical interactions at the synapse between the cells. All of the inputs are then carried to the soma where they add or subtract from the overall potential difference between the interior of the soma and the surrounding fluid. Once the cell potential rises above a certain level, the cell "fires" and sends signals to other cells along its axon.

The artificial cell behaves in much the same way, except that the output signal is analog instead of digital. Signals from sending cells are passed along to a receiving cell through a series of connections. Each connection has an associated weighting factor that acts as a multiplier on the signal from the sending cell. All the inputs to a cell are summed (along with a cell bias, if included) and the resulting value is used to generate the output of the receiving cell. The output of the cell is referred to as the cell *activation* and the function that uses the net input to generate the cell activation is called the *activation function*. The activation function can theoretically be of any form, although linear and sigmoidal functions are frequently used. Figure 6.5.1 shows a comparison between a biological cell and an artificial cell.

When many different cells are combined together into a richly connected network (Figure 6.5.2), the result can behave mathematically like a nonlinear regression engine capable of mapping inputs to outputs for complex relationships. The trick is to find a series of weights *W* that allow the network to provide the desired outputs using specific inputs.

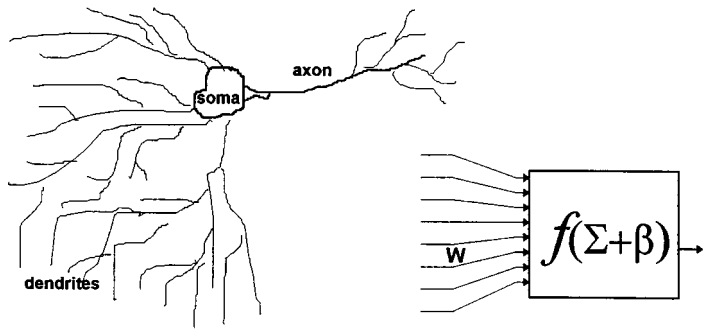


FIGURE 6.5.1 Biological cell vs. artificial cell.

Training Neural Networks

The net input to a cell is given by

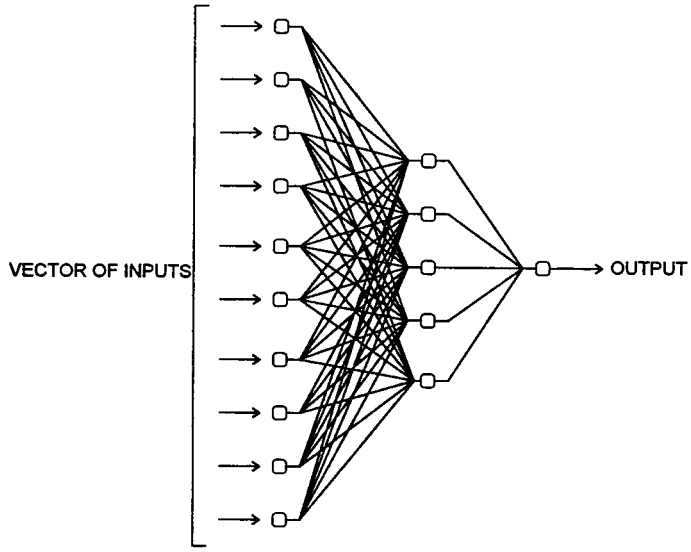


FIGURE 6.5.2 Artificial neural network consisting of several layers of mathematical models of biological neurons.

$$I_{\mathrm{NET},i} = \sum_{j=f}^{1} A_j w_{j\to i}$$

where $I_{\mathrm{NET,i}}$ is the net input of node i due to input from nodes f through I (as in *first* through *last*), A_j is the output of cell j, and $w_{j\to i}$ is the weighting factor associated with the connection between the sending node j and the receiving node i. The output (*activation*) of cell i is then going to be a function of the net input to the cell. Typically, the sigmoid function is used to relate activation to cell inputs

$$A_i = \frac{1}{1 + e^{-I_{\mathrm{NET},i}}}$$

Practically any function can be used as the activation function of each node; the sigmoid function is usually chosen because the derivative, used during the training process, is easy to calculate. Traditional training of neural networks seeks to minimize the error function

$$E = \sum_{p=1}^{N_p} \sum_{i=1}^{N_o} \left(t_{p,i} - A_{p,i}\right)^2$$

where N_P is the number of input/output pairs of data, N_o is the number of outputs of the network, and $t_{p,i}$ is the desired (*target*) output value for a given set of inputs. This error function is minimized by adjusting the values of the weights proportionally to the negative of the derivative of the error with respect to each weight,

$$\Delta w_{j\to i} = -\varepsilon \frac{\partial E}{\partial w_{j\to i}}$$

where ε is the *learning rate* of the network. In a multilayered network the training is initiated by stimulating the network with a specific vector of inputs. The outputs of all the cells are then calculated

in order, starting with the input layer and ending with the output layer. The output is then compared with the known target output value, and any errors are compensated for by adjusting the weights from the output layer back toward the input layer. This method of training is, therefore. called *back-propagation*. With such a method, the previous equation for Δw can be rewritten as

$$\Delta w_{j\to i} = -\varepsilon\delta_i A_j$$

where δ_i represents the effect of a change in the net input to cell j on the output of cell i. For cells in the output layer,

$$\delta_i = \left(t_i - A_i\right) f'\left(I_{\text{NET},i}\right)$$

where f' is the derivative of the sigmoid activation function. For nodes in the hidden layers,

$$\delta_i = f'\left(I_{\text{NET},i}\right)\sum_{j=f}^{1}\left(\delta_i w_{j\to i}\right)$$

The product d_iA_j is something called the *weight-error derivative* (WED). To prevent excessive oscillation of the weights during training, the change of weights can be restricted according to

$$\Delta w(k) = \varepsilon \cdot \text{WED} + \mu \cdot \Delta w(k-1)$$

where μ is the *momentum* of the network. Finally, to gradually reduce the rate at which the weights change, the learning rate ε can be made subject to exponential decay during the training process. This is sometimes called *simulated annealing*.

Bias nodes are like stand-alone cells that connect to each "normal" cell of the network. The output activation of each bias node is always unity and the "weight" connecting the bias node to the normal cell acts as a threshold function that suppresses or augments the output of the cell.

Using Networks for Controlling Feedback Loop Processes

Neural networks offer the potential for and have demonstrated improved control of processes through predictive techniques. The concept is fairly simple: train a network to predict the dynamic behavior of a process and then use these predictions to modify the controller output to place the process at a desired set point $R(t)$ at some time in the future. Initial results from computer simulations of such a controller are presented in Curtiss et al. (1993 a,b,c). Anderson (1989) described a computer simulation in which a network was trained to recognize the dynamic properties of an inverted pendulum (e.g., a broom balanced on an open palm). A control system was developed in which the angle and position of the pendulum were used to move the supporting base in order to maintain the pendulum upright. A neural network-based predictive controller is outlined in the classic discussion by Nguyen and Widrow (1989) on the "truck backer-upper" problem in which a tractor-trailer is backed into position at a loading dock.

Properly tuned fixed-gain controllers will usually work over a relatively wide range of process operation provided that the external perturbations and influences are small or time invariant. With nonlinear processes, however, a conventional control algorithm can lead to unstable control if the gains were chosen for a range different from the current operating conditions.

Architecture of the Network

With the neural network approach it is possible to overcome these problems by using as many additional salient inputs (the *auxiliary* inputs) as necessary and by incorporating an inherently nonlinear model to accomplish the control objectives. The network is trained using examples of the time-dependent relationship between a value of the feedback and previous values of the feedback, the controller output and

the auxiliary inputs. An example of the network architecture required for this is shown in Figure 6.5.3. In practice there does not need to be a limit on the number of previous measurements of any of the inputs, although the final size of the network and the corresponding training time and memory requirements need to be taken into consideration.

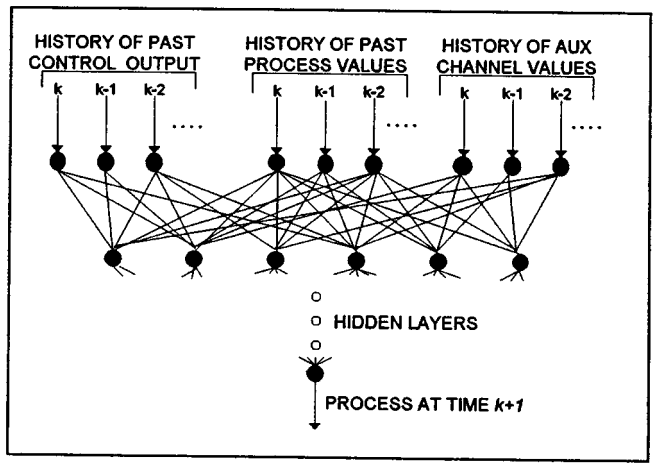


FIGURE 6.5.3 Network used for process prediction.

The network, once trained, can predict the future feedback value for any controller output. The trick is to find the controller output that causes the future process value to match the set point. This is accomplished by finding the derivative of the future error with respect to the current controller signal. Starting with the current process conditions, the feedback value is predicted at each time step into the future over a preset time window (). During each step of the prediction the values for the controller output and auxiliary inputs are held constant. This simulation is performed twice: the first time with a small increase in the controller output and the second time with a small decrease. This allows for the calculation of the change of the future process value (and hence the change of the future error) as a function of the change in the current controller output. The controller output is then modified by

$$\Delta U(t) = -G_{\text{net}} \cdot E_f(t) \frac{\partial E_f(t)}{\partial U(t)}$$

where E_f is the future error and G_{net} is the network controller gain. For a multiple-output controller, the additional outputs are simply added as more outputs of the network and the future predictions repeated several times to find the correct partial derivatives.

Many different variations on this theme are possible, for example, using the sum of the absolute values of all the errors over the prediction window (or the sum of the square of the errors, etc.) instead of simply the future error. Computer-simulated results of such tests are provided by Curtiss et al. (1993).

Estimating the Size of the Prediction Time Window. It is possible to use the network model to determine the size of the time window by estimating the amount of time required for the process to reach some future steady state after a simulated change in the controller output. An example of such an open-loop response it is shown in Figure 6.5.5. Here the network is simulating the response of a reverse-acting process after a decrease in actuator position at time step 0. About 70% (ln 2) of total rise time is achieved after 15 time steps. This kind of calculation can be performed during the control sequence and should indicate the proper time window size.

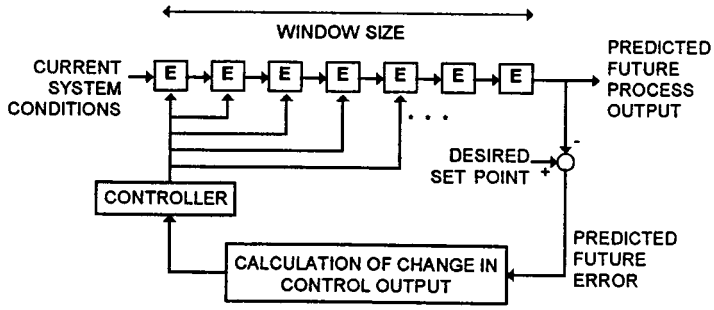


FIGURE 6.5.4 Schematic of procedure for determining future process value and error.

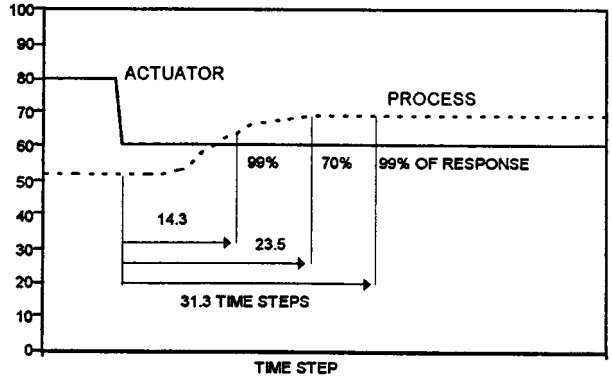


FIGURE 6.5.5 Example of computer-simulated process step change (used to determine size of time window).

Example of PID vs. Network Controller

Figure 6.5.6 shows an example of a process under PID control that demonstrates nonlinearity at different ranges of actuator position. Figure 6.5.7 shows the same process under the influence of a predictive neural network controller that had been trained on the process. Note that the network-based controller does not show the same problems of unstable control in certain actuator ranges. The size of the time window (15 time steps) was determined using the method discussed in the previous section. .

Using Networks as Supervisory Controllers

The previous section discussed the use of neural networks to minimize a predicted error of a feedback-loop process. It is possible to apply a similar methodology for supervisory plant control to optimize the process according to some cost function. A network is first trained to predict the cost function under a wide range of operating conditions. This network is then used to predict what will happen with different control strategies Figure 6.5.8. shows a schematic of this technique. The left side of the figure shows the training mode, where the network is attempting to associate the various plant inputs with the cost function output. There can be multiple inputs, including uncontrollable variables (e.g., ambient conditions, plant loads, etc.) and controlled variables (i.e., the various process set points.)

Once the network is sufficiently trained, it is used to find values for the set points under any set of uncontrolled variables. The technique for doing so is similar to the back-propagation training technique of the network. The inputs corresponding to the controlled variables are replaced with *virtual nodes* whose outputs are always unity. These nodes are connected to the predictor network through adjustable weights. The optimization occurs by finding values for these weights that allow the model to predict a desired output. These weights can be found through any number of search methods, including the gradient descent technique used in back-propagation training. In this case, the predictor network is "trained"

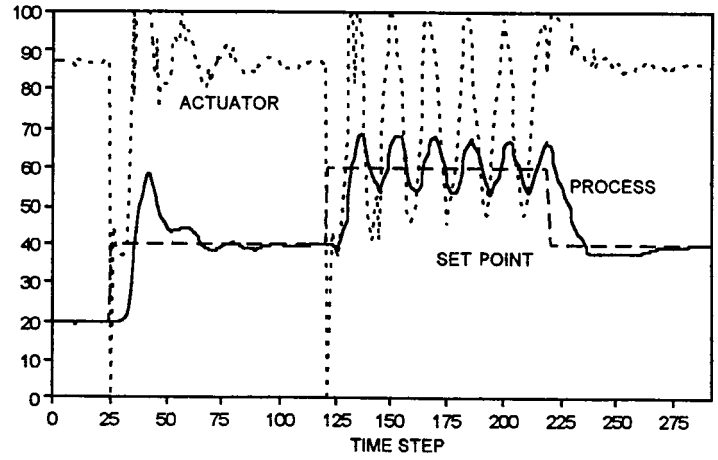


FIGURE 6.5.6 Example of computer simulation using a PID controller.

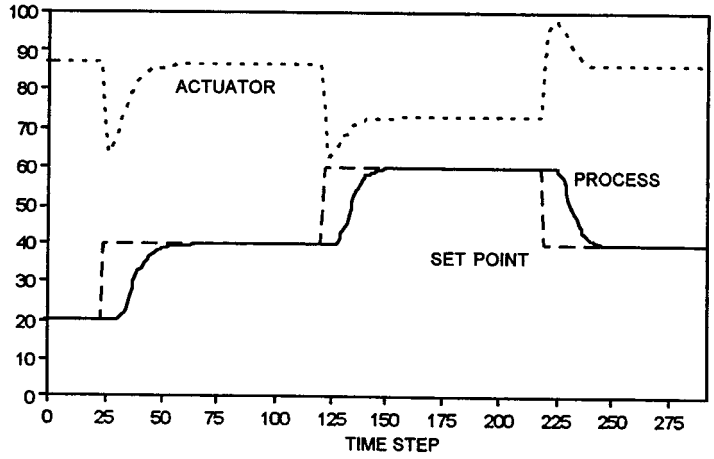


FIGURE 6.5.7 Example of computer simulation using a neural network controller.

normally, except that all weights in the network are static except those connected to the virtual nodes. Once weights have been found that produce the desired output, the set points can be found from direct interpretation of these weights. Constraints can be imposed on the weights either through physical limitations (e.g., freezing points) or from predictions from local-loop neural network controllers.

Fuzzy Logic Controllers

Fuzzy logic controllers (FLC) use conditional relationships to analyze one or more inputs. That is, the inputs are subject to a series of *if…then* queries to produce some intermediate values. An example would be something like a simple cruise control on an automobile:

- *If* vehicle speed = much lower than set point, *then* need to increase speed = large
- *If* vehicle speed = slightly lower than set point, *then* need to increase speed = small

These intermediate values are then used to determine the actual change of speed in the car:

- *If* need to increase speed = large, *then* increase of throttle position = 10%
- *If* need to increase speed = small, *then* increase of throttle position = 3%

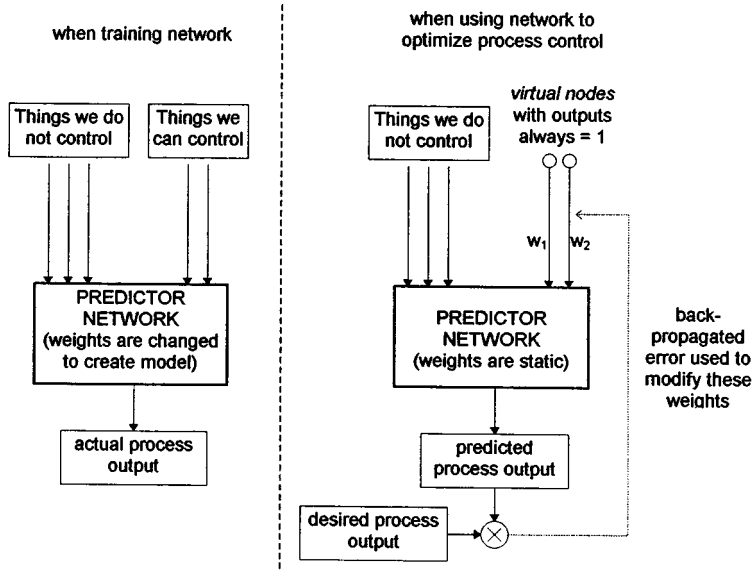


FIGURE 6.5.8 Using network model to optimize process control.

In fuzzy control, the satisfaction of a particular *if* statement may not lead to or be restricted by a true or false response. A range of weighting coefficients is assigned to a particular conditional, with the coefficients generally decreasing as the certainty of a specific condition decreases. In the example above, "need to increase speed = large" may be assigned a certainty of 1 when the speed of the vehicle is less than 50% of the desired speed of the car. This certainty will decrease as the speed of the car increases, so that "need to increase speed = large" may be 0 when the speed of the car is greater than 90% of the desired speed, but the certainty of "need to increase speed = small" will be large. It is possible that for a given speed of the car, two or more conditions may be satisfied with different magnitudes of certainty. These conditions can then be applied to the output rules along with their respective certainties to determine the actual increase (or decrease) in the controller output. When the speed of the car is below the desired speed, the initial rules may yield, for example,

- Need to increase speed = large with certainty = 0.3
- Need to increase speed = small with certainty = 0.7

the actual output would then be

- Increase of throttle position = $(0.3 \times 10\% + 0.7 \times 3\%)/(0.3 + 0.7) = 5.1\%$

The following section formalizes some of these ideas and includes a detailed example.

Section 19 Mathematics contains the formalism underlying fuzzy set theory and fuzzy logic. The reader is referred to that section and the one that follows for the technical basis for FLCs.

Fuzzy Logic Controllers for Mechanical Systems

Introduction

In the last decade, FLCs have been receiving more attention (Leigh and Wetton 1983; Daley and Gill, 1985; Yasunobu and Miyamoto, 1985; Xu, 1989), not only in test cases, but also in real industrial process control applications, including building mechanical systems (Sakai and Ohkusa, 1985; Ono et al., 1989; Togai and Maski, 1991; Huang and Nelson, 1991; Meijer, 1992). The basic idea of this approach is to incorporate the experience of human operators in the design of controllers. From a set of linguistic rules describing operators' control strategies, a control algorithm can be constructed (Ralston and Ward, 1985).

Computer simulations and experiments have shown that FLCs may have better performance than those obtained by conventional controllers. In particular, FLCs appear very useful when the processes are too complex for analysis using conventional control algorithms or when the available information is qualitative, inexact, or uncertain. Thus, fuzzy logic control may be viewed as a compromise between conventional precise mathematical control and humanlike decision making, as indicated by Gupta (Gupta and Tsukamoto, 1980).

However, fuzzy logic controllers sometimes fail to obtain satisfactory results with the initial rule set drawn from the operators' experience. This is because there still are some differences between the way a plant is operated by an experienced operator and by an FLC using the rules based directly on his or her experience. It is often difficult to express human experience exactly using linguistic rules in a simple form. Sometimes there is no experience that could be used to construct control rules for FLCs. In these cases it is necessary to design, develop, and modify control rules for FLCs to obtain optimal performance. There have been few discussions about rule development and adjustment strategies for FLCs (Sheridah, 1984; Scharf and Mandic, 1985; Wakileh and Gill, 1988; Ollero and Williams, 1989).

The Basic Aspects of an FLC

An FLC includes three parts: fuzzifier, fuzzy reasoning unit, and defuzzifier. The fuzzifier converts ordinary inputs into their fuzzy counterparts, the fuzzy reasoning unit creates fuzzy control signals based on these fuzzy variables, and the defuzzifier converts the fuzzy control signals into the real control outputs. The block diagram of a fuzzy control system is shown in Figure 6.5.9, where e, d, and u are tracking error, derivative error, and output control action; $\tilde{e}$, $\tilde{d}$, and $\tilde{u}$ are their fuzzy counterparts, respectively; y is the controlled parameter; and r is the set point for y. K_p is the scale factor for e, K_d is the scale factor for d, and K_o is the output gain.

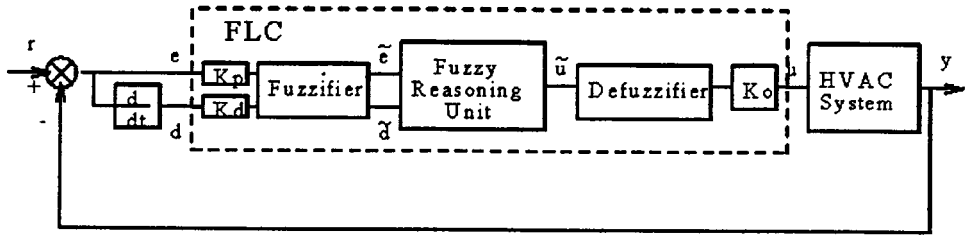


FIGURE 6.5.9 The block diagram of a fuzzy control system.

The control rules expressed in natural language are expressed in the following form:

IF (e is A) AND (d is B) THEN (u is C)

where A, B, and C are fuzzy subsets defined on the universes of discourse of e, d, and u, respectively. Every rule is interpreted into a fuzzy reasoning matrix:

$$R_k = \left[A_k(e) \otimes B_k(d)\right]^{\Theta} \otimes C_k(k) \qquad k = (1, N)$$

where N is the number of rules, the symbol $\otimes$ denotes aggregation operator, and the symbol Θ denotes an align-turning operator (see Section 19). The general fuzzy relation matrix R can be constructed as the union of the individual rules:

$$R = \bigcup_{k=1}^{N} R_k$$

This matrix represents the relationship between the fuzzy inputs and the fuzzy control output. The fuzzy control output can then be calculated from the known fuzzy input $\tilde{e}$ and $\tilde{d}$ by

$$\tilde{u} = \left[\tilde{e} \otimes \tilde{d}\right]^{\Theta} \circ R$$

where the symbol ° denotes the max-min composition operator (see Section 19).

The input universe of discourse for tracking error *e* or derivative error *d* is divided into several degrees connected with a number of fuzzy subsets by membership functions. In this study, *e* and *d* can each range from –6 to +6, and 13 degrees are used:

$$-6, -5, -4, -3, -2, -1, 0, 1, 2, 3, 4, 5, 6$$

Also, seven fuzzy subsets are defined as:

$$\text{NL, NM, NS, ZZ, PS, PM, PL.}$$

where the first letters N and P mean negative and positive, the second letters L, M, and S mean large, middle, and small, and ZZ means zero. These degrees and fuzzy subsets are shown in Table 6.5.1 which uses a 1.0–0.8–0.5–0.1 distribution. For example, if *e* = 3, then its membership in PL is 0.1, its membership in PM is 0.8, etc.

TABLE 6.5.1 The Membership Function of Input of an FLC

***A*(*e*), *B*(*d*)**	**–6**	**–5**	**–4**	**–3**	**–2**	**–1**	**0**	**1**	**2**	**3**	**4**	**5**	**6**
PL	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.5	0.8	1.0
PM	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.5	0.8	1.0	0.8	0.5
PS	0.0	0.0	0.0	0.0	0.0	0.1	0.5	0.8	1.0	0.8	0.5	0.1	0.0
ZZ	0.0	0.0	0.0	0.1	0.5	0.8	1.0	0.8	0.5	0.1	0.0	0.0	0.0
NS	0.0	0.1	0.5	0.8	1.0	0.8	0.5	0.1	0.0	0.0	0.0	0.0	0.0
NM	0.5	0.8	1.0	0.8	0.5	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
NL	1.0	0.8	0.5	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

A similar analysis is given to the outputs for the control action indicated in Table 6.5.2 which uses a 1.0–0.7–0.2 distribution and where the abbreviations mean that the output control actions are Very Strong (Level 7), STrong (Level 6), SUbstrong (Level 5), MEdium (Level 4), Slightly Small (Level 3), SMall (Level 2), and TIny (Level 1).

TABLE 6.5.2 The Membership Function of Output of an FLC

***C*(*u*)**	**–6**	**–5**	**–4**	**–3**	**–2**	**–1**	**0**	**1**	**2**	**3**	**4**	**5**	**6**
VS (Level 7)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.7	1.0
ST (Level 6)	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.7	1.0	0.7	0.2
SU (Level 5)	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.7	1.0	0.7	0.2	0.0	0.0
ME (Level 4)	0.0	0.0	0.0	0.0	0.2	0.7	1.0	0.7	0.2	0.0	0.0	0.0	0.0
SS (Level 3)	0.0	0.0	0.2	0.7	1.0	0.7	0.2	0.0	0.0	0.0	0.0	0.0	0.0
SM (Level 2)	0.2	0.7	1.0	0.7	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
TI (Level 1)	1.0	0.7	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

The fuzzifier converts ordinary inputs into their fuzzy counterparts. In this study, a fuzzy singleton is used as a fuzzification strategy, which interprets an input, *e* (or *d*), into a fuzzy value, $\tilde{e}$ (or $\tilde{d}$), with membership function (μ) equal to zero except at the element nearest to the real input, where μ = 1.0. For example, if *e* = 3.2, the nearest element is 3, then the fuzzy singleton will be

$$\tilde{d} = (0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 1,\ 0,\ 0,\ 0)$$

This fuzzy singleton has membership function $\mu = 1.0$ at the point of element $e = 3$. The defuzzifier converts the fuzzy control output created by the rule-based fuzzy reasoning unit into a real control action. In this study, weighted combination method is used as defuzzification strategy, which can be explained by the following example, if

$$\tilde{u} = (0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0.2,\ 0.4,\ 0.8,\ 0.7,\ 0.5,\ 0.1)$$

then

$$u = \left[0.2(1) + 0.4(2) + 0.8(3) + 0.7(4) + 0.5(5) + 0.1(6)\right] / \left[0.2 + 0.4 + 0.8 + 0.7 + 0.5 + 0.1\right] = 3.4$$

Rule Refinement. An FLC is characterized by a set of linguistic statements which are usually in the form of *if-then* rules. The initial set of rules is usually constructed based on the operators' experience, or sometimes by analyzing the dynamic process of the controlled plant. Both approaches require modifying the initial set of rules to obtain an optimal rule set. This is called *rule refinement.*

Figure 6.5.10 shows an initial rule set analyzed on a "linguistic plane". The horizontal axis expresses the fuzzy subsets defined on the universe of discourse for the tracking error (e), and the vertical axis expresses the fuzzy subsets defined on the universe of discourse for the derivative error (d). Both have seven fuzzy "values": NL, NM, NS, ZZ, PS, PM, PL. On the cross points of these fuzzy values there are output control action levels, which are also fuzzy subsets having seven "values" from Level 1 (TIny) to Level 7 (Very Strong). For example, the cross point of e = NM and d = PM indicates u = Level 3. This corresponds to the rule:

IF (e is NM) AND (d is PM) THEN (u is Level 3)

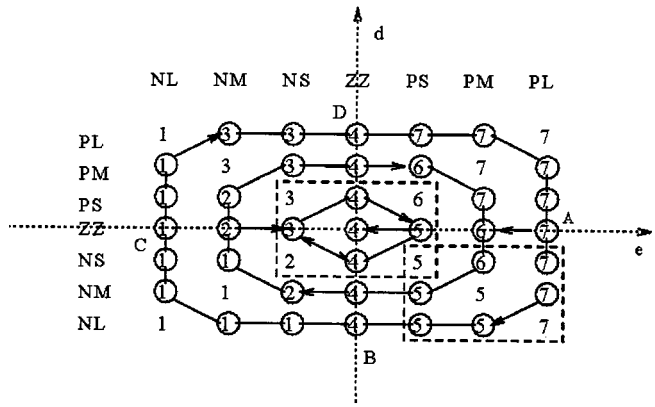


FIGURE 6.5.10 The initial rule set and performance trajectory on the linguistic plane.

For example, the initial rule set could be based on the following control strategies. First, it tries to keep a proportional relationship between the control action (u) and the tracking error (e). Note that if the derivative error (d) is ZZ, then the output control action (u) increases from Level 1 to Level 7 when the tracking error (e) changes from NL to PL. Second, the influence of derivative error (d) is considered such that, if it is positive, then increase the control action (u) a little bit, and if it is negative, then decrease the control action (u). For example, if the tracking error (e) keeps PM, the control action (u)

increases from Level 6 to Level 7 when the derivative error (*d*) is positive, and it decreases from Level 6 to Level 5 when the derivative error (*d*) is negative.

Consider a second-order plant with a transfer function:

$$H(s) = \frac{1.0}{s^2 + 0.1s + 1.0}$$

that is controlled using the initial rule set to respond to a step input for computer simulation. The performance trajectory of the FLC is shown by the arrows in Figure 6.5.10 and the dynamic process of the normalized controlled parameter (CP) is shown in Figure 6.5.11 where the horizontal axis indicates the number of sample period (SP). The dynamic process can be divided into two stages. At the first stage, there is a strong oscillation with a higher frequency, and, at the second stage, there is a moderate swing with a smaller frequency. Looking at the performance trajectory in the linguistic plane, we can see that the stronger oscillation occurs at the out-cycle (points further from the center). As time increases, the state moves to the in-cycle near the center of the plane and becomes moderate. This shows that FLCs have the desirable property of a structure-variable controller. The rules at the out-cycle belong to one kind of structure for the first stage, and the rules at the in-cycle belong to another structure for the second stage.

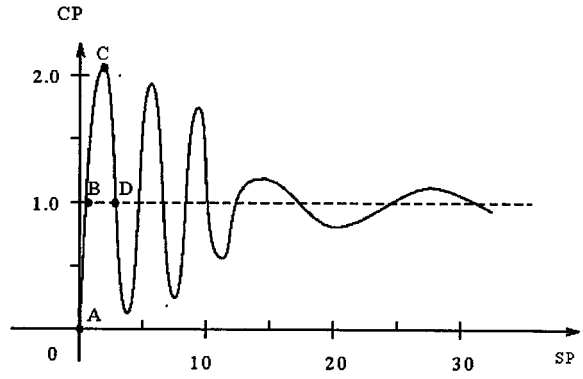


FIGURE 6.5.11 The dynamic process corresponding to Figure 6.5.10.

If the initial rule set does not satisfy a good design for a controller, then It can be modified by intuitive reasoning. A rule set is often symmetrically positioned about the central point, which is the desired stable operating point, where the tracking error (*e*) and the derivative error (*d*) both equal zero and the control action (*u*) is medium. When a positive step increase is imposed to the set point, the tracking error (*e*) has the biggest value and the derivative error (*d*) is zero at the beginning time (point A in the linguistic plane). With the regulating action, the tracking error (*e*) will decrease, the derivative error (*d*) will be negative, and the performance trajectory will enter into the right-bottom block in the linguistic plane. So, the rules in this area have the most important effect on the behavior of the first stage of the dynamic process. The most important area responsible for the behavior of the second stage is the central block.

To avoid strong oscillations, it is apparent that the control actions in the right-bottom block should be decreased. The modified rule set and its simulation of response to a step input are shown in Figure 6.5.12. The performance trajectory expressed in the linguistic plane is a spiral (Figure 6.5.12). We can see that the performance of the control system has been improved, but a small oscillation still exists and there is a little overshoot indicated by point C in Figures 6.5.12 to 6.5.13. Once again, the rule set is modified and the final rule set and its simulation of response to a step input are shown in Figure 6.5.14

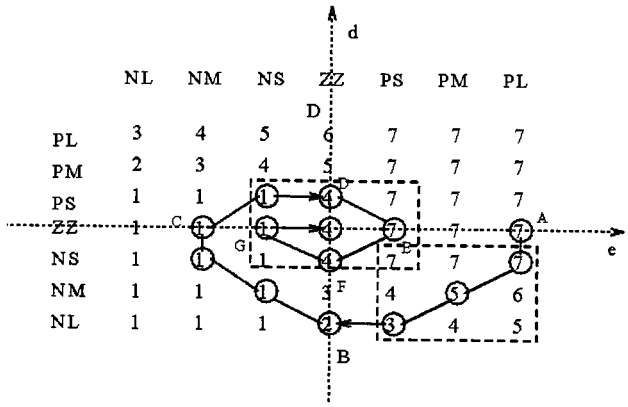


FIGURE 6.5.12 The second rule set on the linguistic plane.

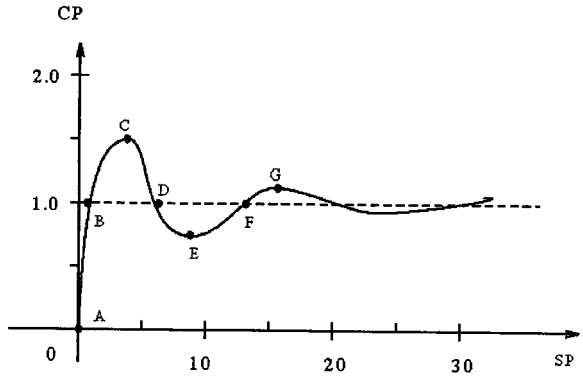


FIGURE 6.5.13 The dynamic process corresponding to Figure 6.5.12.

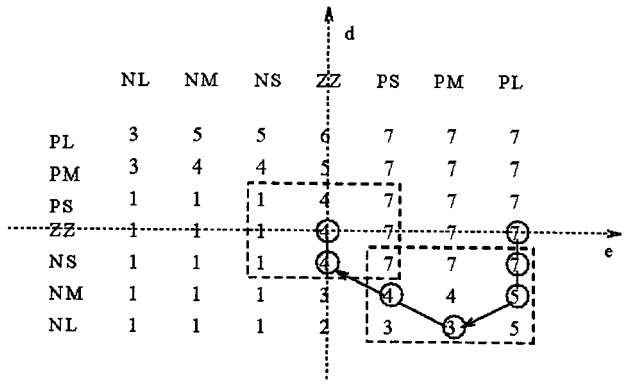


FIGURE 6.5.14 The third rule set on the linguistic plane.

and Figure 6.5.15. The final rule set gives good performance with a short rise time and a very small overshoot and it is considered satisfactory.

By analyzing the performance trajectory on the linguistic plane, a rule set is refined. It relies heavily on intuitive reasoning when comparing the dynamic process of the controlled parameter for the present rule set with the desired one.

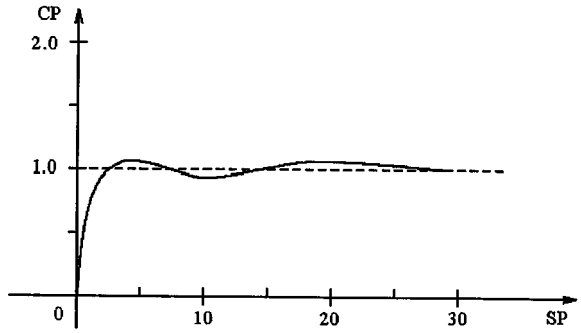


FIGURE 6.5.15 The dynamic process corresponding to Figure 6.5.14.

Completeness and Interaction of Rules and Selection of Membership Functions. The second significant influence on the behavior of an FLC is the membership functions. They should be chosen carefully in the adjustment process. As mentioned in Section 6.3, the fuzzy subsets, language variables NL, NM, NS, ZZ, PS, PM, and PL, are defined on the universe discourse of tracking error (e) or derivative error (d). Some possible membership functions are shown in Figures 6.5.16 to 6.5.18. The membership functions should be chosen to make these language variables have suitable coverage on the universe of discourse. For the case of Figure 6.5.16, the whole range is not covered by these language variables. There are some values of e or d, on which the membership functions of all language variables are zero. In this case, an empty output control action could be created. This means that the control actions are lost for those points which are not covered by any input fuzzy subset. This is referred as the non-completeness of control rules. FLCs should satisfy the condition of completeness for their membership functions. The membership function shown in Figure 6.5.17 cannot be used for an effective FLC. In other words, the union of all fuzzy subsets, X_i, i = [1,7], should be greater than zero for all $e \in E$, i.e.,

$$\forall e \in E \qquad \bigcup_{i=1}^{7} X_i(e) > 0$$

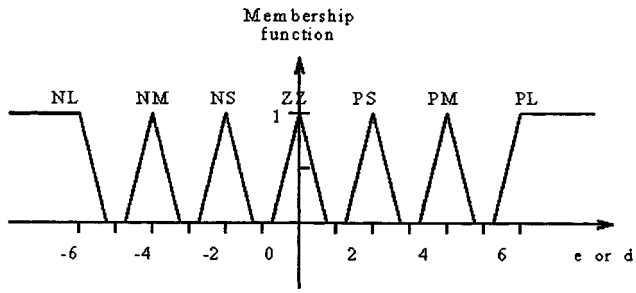


FIGURE 6.5.16 Non-complete membership function.

On the other hand, there can be interaction among the rules if the overlap of fuzzy subsets occurs on the range of the universe of discourse. In this case, the membership functions have the forms shown in Figures 6.5.17 and 6.5.18. The interaction tends to smooth out the set of control rules. Consider the single-input-single-output case for simplicity; the rule set is

IF (e is A_i) THEN (u is C_i) $\qquad i$ = [1,N]

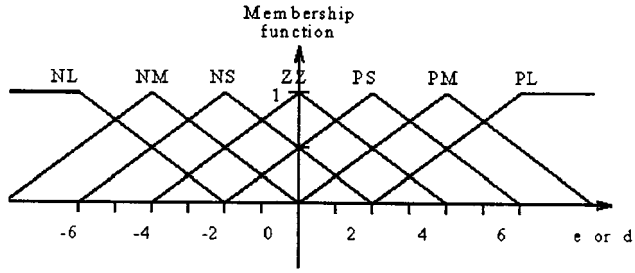


FIGURE 6.5.17 Heavy overlap membership function.

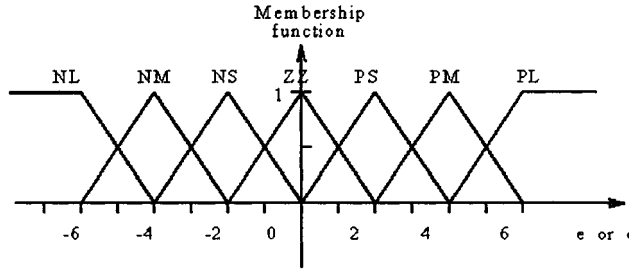


FIGURE 6.5.18 Moderate overlap membership function

where N is the number of rules in the set. These rules are incorporated into a fuzzy relation matrix as follows:

$$R = \bigcup_{i=1}^{N} R_i = \bigcup_{i=1}^{N} \left(A_i \otimes C_i\right)$$

If the fuzzy value of input e is known as $\tilde{e}$, the fuzzy output $\tilde{u}$ then can be calculated as follows:

$$\tilde{u} = \tilde{e} \circ R$$

If $\tilde{e}$ is A_i, $\tilde{u}$ is expected to be C_i. But now the interaction of rules due to overlap results in:

$$C_i \subseteq A_i \circ R$$

The equality is established only when no overlap occurs. This analysis is based on the fuzzy logic scheme including max-min composition operator. A more-detailed example of the numeric calculation is given in the appendices to this section.

If the overlap is heavy as shown in Figure 6.5.17, there will be large deformation and the control rules will lose their original shape. In the limit, as the membership functions become unity for all values, the output of the FLC will always be the same fuzzy quantity. This means that the fuzzy reasoning system conveys no valuable information and the FLC has lost its efficacy.

A moderate overlap, shown in Figure 6.5.18, is desirable to allow for reasoning with uncertainty and the need for completeness of the control rules. How does one determine the "size" of overlap? At present, we use intuitive judgment to choose membership functions when adjusting an FLC. There appears to be some latitude in choosing the amount of overlap, on which the performance of an FLC does not change significantly. The quantitative analysis will be given after further research.

When we modify the control rules in the linguistic plane, the overlapping membership functions let the rules near the performance trajectory have an effect on the output control actions. This is because interactions occur among the neighboring rules.

Scale Factors and Output Gain

The scale factors, K_p and K_d, and the output gain, K_o, shown in Figure 6.5.19, also have significant influence on the behavior of an FLC. Their influence is not as complicated as those of rules and membership functions. The adjustment for the scale factors and output gain is comparatively simple. The scale factor K_p relates the actual range of tracking error (e) to the universe of discourse (E) defined in the fuzzy logic system. In this work, E consists of 13 degrees as indicated in earlier sections. Then K_p is determined as the ratio of the range of E to the range of the real variable:

$$K_p = \frac{E_{\max} - E_{\min}}{e_{\max} - e_{\min}}$$

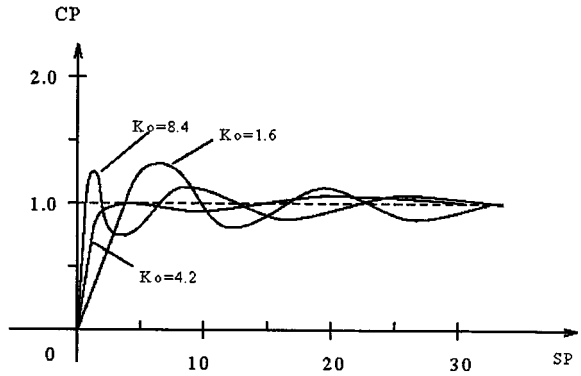


FIGURE 6.5.19 The influence of K_o on the behavior of FLCs.

For scale factor K_d, there is the similar analysis leading to

$$K_d = \frac{D_{\max} - D_{\min}}{d_{\max} - d_{\min}}$$

where D is the universe of discourse for derivative error (d) defined in the fuzzy logic system. Small K_p or K_d will narrow the control band, while large K_p or K_d will lead to loss of control for large inputs.

The output gain K_o is defined as follows:

$$K_o = \frac{u_{\max} - u_{\min}}{U_{\max} - U_{\min}}$$

It is the ratio of range of real output control action (u) to the range of its universe of discourse (U) defined in the fuzzy logic system. K_o acts as an amplification factor of the whole FLC. Figure 6.5.19 shows the influence of K_o on the step response simulation of an FLC with the final rule set used in Figure 6.5.14. Increasing K_o results in a shorter rise time. The performance trajectory in the linguistic plane will become steeper for the first stage and oscillation occurs. Decreasing K_o results in a longer rise time, and the performance trajectory in the linguistic plane will become moderate during the first stage. But, in our simulation, oscillation still occurred. This is because different K_o, larger or smaller, results in a new route of the performance trajectory which will activate the different rules which might

cause oscillation. So the influence of output gain, K_o, should be considered together with the change of the activated rules.

Conclusion

An FLC can perform much better than a conventional controller, such as a PID controller, if the FLC has been well constructed. The main disadvantage of using FLCs today seems to be the lack of a systematic procedure for the design of FLCs. The general method for designing an FLC is to use trial and observation. No useful mathematical tool has yet been developed for the design of an FLC because of its fuzziness, complexity, and nonparameterization.

There are three significant elements that have notable influence on the behavior of an FLC:

1. The control rules expressed in linguistic language,
2. The membership functions defined for fuzzy subsets, and
3. The scale factors attached to the input and the output gains.

The control rules play the main role in forming the dynamics of FLCs. The rule set can be analyzed and modified using the performance trajectory technique and evaluated using the dynamic process curve of the controlled parameter. The membership functions define the "shape" of fuzzy subsets. They should have appropriate width to avoid noncompleteness and suitable interaction among the fuzzy control rules. The scale factors (K_p and K_d) and output gain (K_o) serve as amplification factors.

At present, each application must be individually designed. The initial sets of rules are specifically set up for different applications. Work is now underway to develop a self-adaptive FLC which will choose the initial set of rules automatically according to the process dynamics and refine it on the basis of the global performance evaluation.

References

Anderson, C. 1989. Learning to control an inverted pendulum using neural networks, *IEEE Control Systems Magazine*, April, 31–36.

Askey, S. Y. 1995. Design and Evaluation of Decision Aids for Control of High Speed Trains: Experiments and a Model, Ph.D. Thesis, Massachusetts Institute of Technology, Cambridge, MA, June.

Billings, C.E. 1991. *Human-Centered Aircraft Automation: A Concept and Guidelines*, NASA Ames Research Center, Moffet Field, CA.

Curtiss, P.S., Kreider, J.F., and Brandemuehl, M.J. 1993a. Artificial neural networks proof of concept for local and global control of commercial building HVAC systems, in *Proceedings from the ASME International Solar Energy Conference*, Washington, D.C.

Curtiss, P.S., Kreider, J.F., and Brandemuehl, M.J. 1993b. Energy management in central HVAC plants using neural networks, in *Proceedings from the ASHRAE Annual Winter Meeting*, Chicago, IL.

Curtiss, P.S., Brandemuehl, M.J., and Kreider, J.F. 1993c. Adaptive control of HVAC processes using predictive neural networks, *ASHRAE Trans.*, 99(1).

Daley, S. and Gill, K.F. 1985. The fuzzy logic controller: an alternative design scheme?'' *Comput. Ind.*, 6, 3–14.

Gupta, M.M. and Tsukamoto, Y. 1980. Fuzzy logic controllers — a perspective.'' *Proc. Joint Automatic Control Conf.*, pp. FA10-C, August, San Francisco.

Huang, S.-H. and Nelson, R.M. 1991. A PID-law-combining fuzzy controller for HVAC applications, *ASHRAE Trans.*, 97(2), 768–774.

Huang, S.-H. and Nelson, R.M. 1994. Rule development and adjustment strategies of a fuzzy logic controller for an HVAC system: Part Two — experiment, *ASHRAE Trans.*, 99(2), 851–856.

Leigh, R. and Wetton, M. 1983. Thinking clearly with fuzzy logic, *Process Eng.*, 64, 36–37.

MacArthur, J.W., Grald, E.W., and Konar, A.F. 1989. An effective approach for dynamically compensated adaptive control, *ASHRAE Trans.*, 95(2), 415-423.

Meijer, G. 1992. Fuzzy logic-controlled A/Cs heat pumps, *IEA Heat Pump Cent. Newslett.*, 10(1).

Nguyen, D.H. and Widrow, B. 1989. The truck backer-upper: an example of self learning in neural networks, *Proceedings of the International Joint Conference on Neural Networks*, 2, 357–363.

Ollero, A. and Williams, J. 1989. Direct digital control, auto-tuning, and supervision using fuzzy logic, *Fuzzy Sets Syst.*, 30, 135–153.

Ono, H., Ohnish, T., and Terada, Y. 1989. Combustion control of refuse incineration plant by fuzzy logic, *Fuzzy Sets Syst.*, 32, 193–206.

Radke, F. and Isermann, R. 1987. A parameter-adaptive PID-controller with stepwise parameter optimization, *Automatica*, 23, 449–457.

Ralston, P.A. and Ward, T.L. 1985. Fuzzy control of industrial process, in *Applications of Fuzzy Set Methodologies in Industrial Engineering*, Elsevier Science Publishers North-Holland; Amsterdam, 29–45.

Sakai, Y. and Ohkusa, K. 1985. A fuzzy controller in turning process automation, in *Industrial Application of Fuzzy Control*, Elsevier Science Publishers North-Holland; Amsterdam, 139–151.

Scharf, E.M. and Mandic, N.J. 1985. The application of a fuzzy controller to the control of a multi-degree-of-freedom robot arm, *Industrial Application of Fuzzy Control*, Elsevier Science Publishers North-Holland; Amsterdam, 1–18.

Sheridan, S.E. 1984. Automatic kiln control at Oregon Portland Cement Company's Durkee plant utilizing fuzzy logic, *IEEE Trans. Ind. Appl.*, 20, 562–568.

Sheridan, T.B. 1987. Supervisory control. In G. Salvendy, Ed., *Handbook of Human Factors/Ergonomics*, Wiley, New York.

Sheridan, T.B. 1992. *Telerobotics, Automation and Human Supervisory Control*, MIT Press, Cambridge, MA.

Shinners, S.M. 1978. *Modern Control System Theory and Application*, Addison-Wesley, Reading, MA.

Togai and Maski. 1991. An example of fuzzy logic control, *Comput. Des.*, 30, 93–103.

Wakileh, B.A. and Gill, K.F. 1988. Use of fuzzy logic in robotics, *Comput. Ind.*, 10, 35–46.

Xu, C.W. 1989. Fuzzy system identification, *IEEE Proc.*, 136, Pt. D, No. 4, pp 146–150.

Yasunobu, S. and Miyamoto, S. 1985. Automatic train operation system by predictive fuzzy control, in *Industrial Application of Fuzzy Control*, Elsevier Science Publishers North-Holland; Amsterdam, 1–18.

Further Information

The following are suggested reading for those interested in learning more about neural networks and their use in control systems:

Helferty, J.J., Collins, J.B., Wong, L.C., and Kam, M. 1989. A learning strategy for the control of a one-legged hopping robot, *Proceedings of the 1989 American Control Conference*, 896–901.

Kuperstein, M. and Rubinstein, J. 1989. Implementation of an adaptive neural network controller for sensory-motor coordination, *IEEE Control Syst. Mag.*, April, 25–30.

Lan, M. 1989. Adaptive control of unknown dynamical systems via neural network approach, *Proceedings of the 1989 American Control Conference*, 910–915.

Liu, H., Iderall, T., and Bekey, G. 1989. Neural network architecture for robot hand control, *IEEE Control Syst. Mag.*, April, 38–41.

Miller, R.C. and Seem, J.E. 1991. Comparison of artificial neural networks with traditional methods of predicting return time from night or weekend setback, *ASHRAE Trans.*, 97(2), 500–508.

Psaltis, D., Sideris, A., and Yamamura, A. 1988. A multilayered neural network controller, *IEEE Control Syst. Mag.*, April, 17–21.

Rumelhart, D.E. and McClelland, J.L. 1986. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, MIT Press, Cambridge, MA.

Wasserman, P.D. 1989. *Neural Computing: Theory and Practice*, Van Nostrand Reinhold, New York.

Appendices

Table of Transforms

The following table lists some of the more common transforms used in the analysis of building systems. More-extensive tables can be found in most mathematics and numerical analysis reference books.

List of Some *s*- and *z*-Transforms

Continous-Time Domain	Frequency Domain	Discrete-Time Domain
$1\ t=0 \quad 0\ t\neq 0$	—	1
$1\ t=k \quad 0\ t\neq k$	—	z^{-k}
1	$\frac{1}{s}$	$\frac{z}{z-1}$
t	$\frac{1}{s^2}$	$\frac{Tz}{(z-1)^2}$
e^{-at}	$\frac{1}{s+a}$	$\frac{z}{z-e^{-aT}}$
te^{-at}	$\frac{1}{(s+a)^2}$	$\frac{Tze^{-aT}}{\left(z-e^{-aT}\right)^2}$
$1-e^{-at}$	$\frac{a}{s(s+a)}$	$\frac{z\left(1-e^{-aT}\right)}{(z-1)\left(z-e^{-aT}\right)}$
$e^{-at}-e^{-bt}$	$\frac{b-a}{(s+a)(s+b)}$	$\frac{z\left(e^{-aT}-e^{-bT}\right)}{\left(z-e^{-aT}\right)\left(z-e^{-bT}\right)}$

Special FLC Mathematical Operations

- $\otimes$ aggregation operator
- Θ align-turning operator
- $\circ$ max-min composition operator
- $\bigcup$ union operator
- $\exists$ exists
- $\in$ in
- $\forall$ for all
- $\subseteq$ is the subset of

The aggregation operator ($\otimes$) is used to define a two-dimensional fuzzy variable F from fuzzy subsets A and B (one-dimensional fuzzy variables) as follows:

$$F = A(e) \otimes B(d) = \begin{bmatrix} \mu_F(1,1) & \cdots & \cdots & \cdots & \mu_F(1,N) \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \mu_F(i,j) & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \mu_F(M,1) & \cdots & \cdots & \cdots & \mu_F(M,N) \end{bmatrix}$$

where

$$\mu_F(i,j) = \min\left(\mu_A(i), \mu_B(j)\right)$$

$$A(e) = \left(e_i, \mu_A(i)\right)$$

$$B(d) = \left(d_j, \mu_B(j)\right)$$

e_i is the ith element of the subset $A(e)$ and $\mu_A(i)$ is its membership function; d_j is the jth element of the subset $B(d)$ and $\mu_B(j)$ is its membership function.

The align-turning operator (Θ) is an operator acting on a two-dimensional fuzzy variable to create a one-dimensional fuzzy variable, which has a set of membership functions aligned according to a certain order, as follows:

$$\begin{aligned} S &= \left[A \otimes B\right]^{\Theta} \\ &= \left(\mu_S(1,1), \mu_S(1,2), \ldots, \mu_S(i,j), \ldots, \mu_S(M,N)\right) \end{aligned}$$

where j varies from 1 to N first and i varies from 1 to M.

An Example of Numeric Calculation for Influence of Membership Function

Suppose that there are two rules:

R1: IF (e is PM) THEN (u is VS) and

R2: IF (e is PS) THEN (u is ST).

For the first rule,

$$\tilde{e}_1 = (0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0.1,\ 0.5,\ 0.8,\ 1.0,\ 0.8,\ 0.5)$$

$$\tilde{u}_1 = (0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0.2,\ 0.7,\ 1.0)$$

Then the rule can be interpreted into a fuzzy reasoning matrix as follows:

$$R_1 = \tilde{e}_1 \otimes \tilde{u}_1$$

$$= \begin{bmatrix}
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.1 & 0.1 & 0.1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.2 & 0.5 & 0.5 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.2 & 0.7 & 0.8 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.2 & 0.7 & 1.0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.2 & 0.7 & 0.8 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.2 & 0.5 & 0.5
\end{bmatrix}$$

$$\equiv \mu_{R_1}(i, j)$$

where the membership in R_1 for the element (i,j) of the matrix, $\mu_{R_1}(i, j)$ is

$$\mu_{R_1}(i, j) = \min\left(\mu_{\tilde{e}_1}(i), \mu_{\tilde{u}_1}(j)\right)$$

For the second rule,

$$\tilde{e}_2 = (0,\ 0,\ 0,\ 0,\ 0,\ 0.1,\ 0.5,\ 0.8,\ 1.0,\ 0.8,\ 0.5,\ 0.1,\ 0)$$

$$\tilde{u}_2 = (0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0.2,\ 0.7,\ 1.0,\ 0.7,\ 0.2)$$

Then the second rule can be interpreted into a fuzzy reasoning matrix as follows:

$$R_2 = \tilde{e}_2 \otimes \tilde{u}_2$$

$$= \begin{bmatrix}
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.2 & 0.5 & 0.5 & 0.5 & 0.2 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.2 & 0.7 & 0.8 & 0.7 & 0.2 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.2 & 0.7 & 1.0 & 0.7 & 0.2 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.2 & 0.7 & 0.8 & 0.7 & 0.2 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.2 & 0.5 & 0.5 & 0.5 & 0.2 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0
\end{bmatrix}$$

$$\equiv \mu_{R_2}(i,j)$$

where the membership in R_2 for the element (i,j) of the matrix, $\mu_{R_2}(i,j)$ is

$$\mu_{R_2}(i,j) = \min\left(\mu_{\tilde{e}_2}(i), \mu_{\tilde{u}_2}(j)\right)$$

The general fuzzy relation matrix R is then constructed as the union of these two rules:

$$R_2 = R_1 \cup R_2$$

$$= \begin{bmatrix}
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.1 & 0.1 & 0.1 & 0.1 & 0.1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.2 & 0.5 & 0.5 & 0.5 & 0.2 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.2 & 0.7 & 0.8 & 0.7 & 0.2 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.2 & 0.7 & 1.0 & 0.7 & 0.5 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.2 & 0.7 & 0.8 & 0.7 & 0.8 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.2 & 0.5 & 0.5 & 0.7 & 1.0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.1 & 0.1 & 0.2 & 0.7 & 0.8 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0.2 & 0.5 & 0.5
\end{bmatrix}$$

$$\equiv \mu_R(i,j)$$

where

$$\mu_R(i,j) = \max\left(\mu_{R_1}(i), \mu_{R_2}(j)\right)$$

Assume that there is an input (e) and its fuzzy value $\tilde{e}$ = PM, then the output is expected to be VS according to the first rule. But now the output is calculated through the fuzzy matrix R as follows:

$$\begin{aligned}\tilde{u} &= \tilde{e} \circ R \\ &= \text{PM} \circ R \\ &= (0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0.1,\ 0.5,\ 0.8,\ 1.0,\ 0.8,\ 0.5) \circ R \\ &= (0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0.2,\ 0.7,\ 0.8,\ 0.7,\ 1.0)\end{aligned}$$

where

$$\mu_{\tilde{u}}(j) = \max_i\left(\min\left(\mu_{\tilde{e}}(i), \mu_R(i,j)\right)\right)$$

While

$$\begin{aligned}\tilde{u}_1 &= \text{VS} \\ &= (0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0,\ 0.2,\ 0.7,\ 1.0)\end{aligned}$$

So,

$$\tilde{u}_1 \subseteq \tilde{u}$$

Goswami, D.Y.; et. al. "Energy Resouces"
Mechanical Engineering Handbook
Ed. Frank Kreith
Boca Raton: CRC Press LLC, 1999

Energy Resouces

D. Yogi Goswami
University of Florida

Robert Reuther
U.S. Department of Energy

Richard Bajura
University of West Virginia

Larry Grayson
University of West Virginia

Philip C. Crouse
Philip C. Crouse and Associates, Inc.

Michael C. Reed
U.S. Department of Energy

Lynn L. Wright
Oak Ridge National Laboratory

Ralph P. Overend
National Renewable Energy Laboratory

Carlton Wiles
National Renewable Energy Laboratory

James S. Tulenko
University of Florida

Dale E. Berg
Sandia National Laboratories

Joel L. Renner
Idaho National Engineering Laboratory

Marshall J. Reed
U.S. Department of Energy

This chapter describes the primary as well as derived energy sources. The objective is to provide information on the extent, availability, measurements and estimation, properties, and limitations of each type of resource. These considerations are important for an engineer to know and understand before attempting selection and design of an energy conversion system. The chapter also includes environmental impacts of energy resources since the environmental aspects are expected to play a major role in the selection of energy resources. In addition, there is a brief discussion of the costs associated with each resource to help in the economic analysis and comparison the resources.

The chapter starts with an introduction and background of a historical perspective on energy use and projections of the future energy needs in the U.S., the industrialized countries, and the world. The primary energy sources described in this chapter include fossil fuels such as coal, natural gas, petroleum (including their synthetic derivatives), biomass (including refuse-derived biomass fuels), nuclear, solar radiation, wind, geothermal, and ocean. In addition there is a brief section on derived energy sources including electricity. So, the terminology and units used for each energy resource and their equivalence are provided.

7.1 Introduction

D. Yogi Goswami

Global energy consumption in the last 50 years has increased at a very rapid rate. Present trends in global population growth, rapid industrialization, and urbanization in major population centers of the world suggest that the world energy demand will continue to increase in the next 50 years (U.S. DOE, 1991). Figure 7.1.1 shows the historical and projected world energy consumption compiled by the Energy Information Agency.

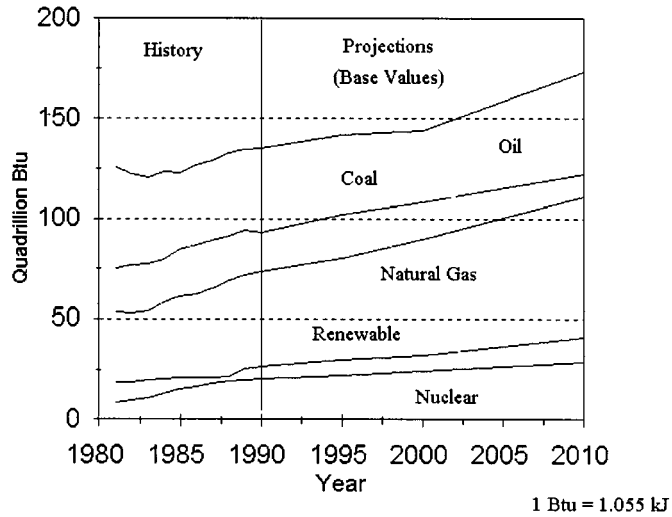


FIGURE 7.1.1 Historical and projected energy consumption. (From EIA. *Annual Energy Outlook 1992*, U.S. DOE, DOE/EIA-0484(92), Washington, D.C.)

The energy resources available to furfill the world demand include

Fossil fuels (oil, coal, natural gas)
Nuclear fuels
Geothermal
Solar radiation
Hydropower

Biomass (crops, wood, municipal solid waste)
Wind
Ocean

Out of all the energy resources, fossil fuels have been used the most (88% of total consumption) because of their extremely high energy densities and simplicity of conversion and use. Figure 7.1.2 shows the world energy production by resource.

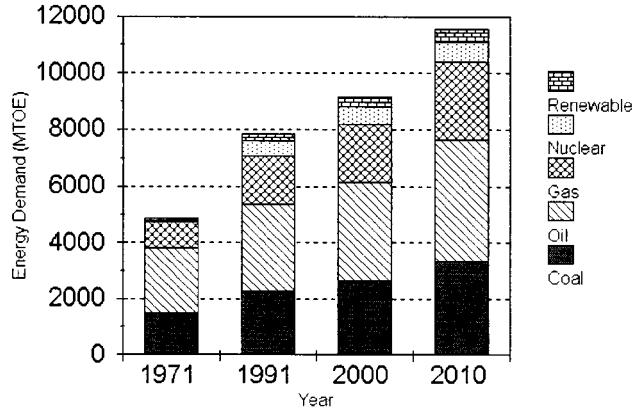


FIGURE 7.1.2 World energy production by resource. (From EIA. *World Energy Outlook*, International Energy Agency, Paris, 1994. With permission.)

Recent concerns about the environment are expected to increase the use of natural gas for power production. Renewable energy resources, such as solar energy, wind, and biomass, are also expected to increase their share of the energy use. There is a strong sentiment in the world in favor of exploiting renewable energy resources, especially because of environmental concerns. How far that sentiment translates into practical use will depend on the development of the renewable energy technologies and prices of the fossil fuels.

Defining Terms

MTOE: Mega tons of oil equivalent; 1 MTOE = 42.63×10^{12} Btu.
Quadrillion Btu: 10^{15} British thermal units (Btu), also known as Quad; 1 Btu = 1055 joules).

References

EIA. 1992. *Annual Energy Outlook* 1992, International Energy Outlook, Energy Information Administration, Office of Integrated Analysis and Forecasting, U.S. DOE, DOE/EIA-0484(92), Washington, D.C.

EIA. 1993. *Annual Energy Outlook 1993*, International Energy Outlook, Energy Information Administration, Office of Integrated Analysis and Forecasting, U.S. DOE, DOE/EIA-0383(93), Washington, D.C.

IEA. 1994. *World Energy Outlook*, Economic Analysis Division, International Energy Agency, Paris.

U.S. DOE. 1991. *National Energy Strategy — Powerful Ideas for America, 1991*. National Technical Information Service, U.S. Department of Commerce, Springfield, VA.

Further Information

Historial and projected energy consumption are published annually by the Energy Information Agency, U.S. Department of Energy, Washington, D.C., and International Energy Agency, Paris.

7.2 Types of Derived Energy

D. Yogi Goswami

Energy from renewable and nonrenewable fuels can be converted to the derived energy forms — thermal, mechanical, and electrical, which are useful for various end uses such as transportation, buildings (heating, cooling, lighting), agricultural, and industrial end uses. The derived energy forms are easily transformed from one type to the other. Figure 7.2.1 shows the projected U.S. energy use by end-use sector.

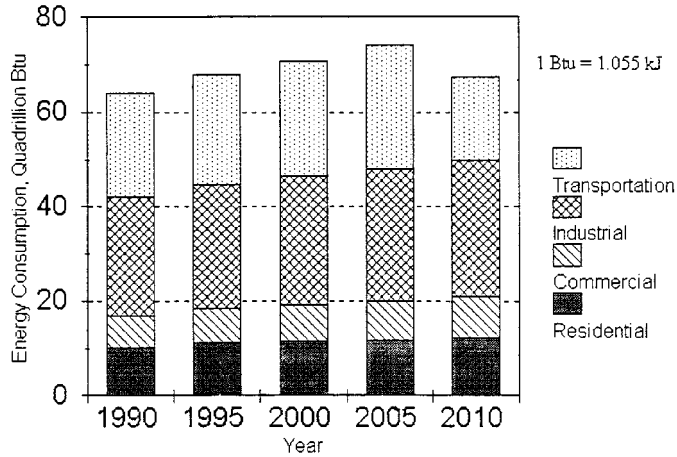


FIGURE 7.2.1 Projected U.S. energy use by end-use sector. (From EIA. *Annual Energy Outlook 1992*, U.S. DOE, DOE/EIA-0484(92), Washington, D.C., 1992.)

Transportation is mainly dependent on oil resources. Efforts to reduce urban air pollution are expected to increase the use of electricity as the preferred energy form for urban transportation. For most of the other end uses electricity will continue to be the preferred energy form. Therefore, it is important to understand the activity in the area of electricity production. Figure 7.2.2 shows the world installed electricity generation capacity by primary energy sources. The United States produces 700 GW (gigawatts or 10^9 W), representing more than 25% of the world electricity capacity. Other major electricity producers are Russia, Europe, Japan, and China. It is expected that China, India, and Southeast Asian countries will add major electricity capacity in the next 20 years.

Integrated resource planning (IRP), or least-cost planning, is the process used to optimize the resource options and minimize the total consumer costs including environmental and health costs that may be attributed to the resource. IRP examines all of the options, including the demand-side options, to minimize the total costs. There is considerable emphasis on IRP in a number of states in United States for future electric capacity and on **demand-side management** (DSM) for the current capacity (Kreith and Burmeister, 1993). The IRP process generally includes some combination of the following steps (Kreith and Burmeister, 1993): development of a load forecast; inventory of existing resources; identification of additional electrical capacity needs; demand-side management programs; screening and identification of options that are feasible; uncertainty analysis in view of uncertainty of future load, fuel prices, capital costs, etc; and selection of a resource or a mix of resources.

Demand Side Management DSM refers to a mix of electrical utility-sponsored custom incentives and disincentives that influence the amount and timing of customer demand in order to better utilize the available resources. Kreith and Burmeister (1993) and SERI (1991) list a number of DSM strategies.

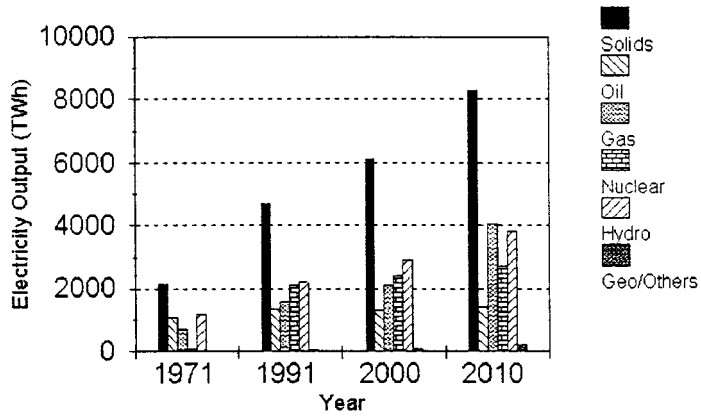


FIGURE 7.2.2 World electricity output. (From EIA, *World Energy Outlook*, International Energy Agency, Paris, 1994. With permission.)

Defining Terms

Demand-side management (DSM): Refers to a mix of incentives and disincentives that influence the amount and timing of energy use in order to better utilize the available resources.

Integrated resource planning (IRP): The process to optimize the resource options and minimize the total consumer costs including environmental and health costs that may be all attributed to the resource.

References

EIA. 1992. *Animal Energy Review.* DOE/EIA-0383(92). Office of Integrated Analysis and Forecasting, U.S. Department of Energy, Washington, D.C.

IEA. 1994. *World Energy Outlook,* Economic Analysis Division, International Energy Agency, Paris.

Kreith, F. and Burmeister, G. 1993. *Energy Management and Conservation.* National Conference of State Legislatures, Denver, CO.

SERI. 1991. Demand Side Management Pocket Guide Book, Volume 1: Residential Technologies; and Volume 2: Commercial Technologies. SERI (Now National Renewable Energy Laboratory), Golden, CO.

Further Information

Annual reviews published by the EIA, U.S. Department of Energy, and the International Energy Agency (see References) provide a wealth of information on electricity capacity and energy consumption by end-use sectors.

7.3 Fossil Fuels

Coal

Robert Reuther

Coal Composition and Classification

Coal is a sedimentary rock formed by the accumulation and decay of organic substances derived from plant tissues and exudates that have been buried over periods of geological time along with various mineral inclusions. Coal is classified by **type and rank.** Coal type classifies coal by the plant sources from which it was derived. Coal rank classifies coal by its degree of metamorphosis from the original plant sources and is therefore a measure of the age of the coal. The process of metamorphosis or aging is termed **coalification.**

The study of coal by type is known as coal petrography. Coal type is determined from the examination of polished sections of a coal sample using a reflected-light microscope. The degree of reflectance and color of a sample are identified with specific residues of the original plant tissues. These various residues are referred to as **macerals.** Macerals are collected into three main groups: vitrinite, inertinite, and exinite (sometimes referred to as liptinite). The maceral groups and their associated macerals are listed in Table 7.3.1 along with a description of the plant tissue from which each distinct maceral type is derived.

TABLE 7.3.1 Coal Maceral Groups and Macerals

Maceral Group	Maceral	Derivation
Vitrinite	Collinite	Humic gels
	Telinite	Wood, bark, and cortical tissue
	Pseudovitrinite	? (Some observors place in the inertinite group)
Exinite	Sporinite	Fungal and other spores
	Cutinite	Leaf cuticles
	Resinite	Resin bodies and waxes
	Alginite	Algal remains
Inertinite	Micrinite	Unspecified detrital matter, <10 μm
	Macrinite	Unspecified detrital matter, 10–100 μm
	Semifusinite	“Burned” woody tissue, low reflectance
	Fusinite	“Burned” woody tissue, high reflectance
	Sclerotinite	Fungal sclerotia and mycelia

Modified from Berkowitz, N., *An Introduction to Coal Technology.* Academic Press, New York, 1979.

Coal rank is the most important property of coal, since it is rank which initiates the classification of coal for use. Rank is a measure of the age or degree of coalification of coal. Coalification describes the process which the buried organic matter goes through to become coal. When first buried, the organic matter has a certain elemental composition and organic structure. However, as the material becomes subjected to heat and pressure, the composition and structure slowly change. Certain structures are broken down, and others are formed. Some elements are lost through volatilization while others are concentrated through a number of processes, including being exposed to underground flows which carry away some elements and deposit others. Coalification changes the values of various properties of coal. Thus, coal can be classified by rank through the measurement of one or more of these changing properties.

In the United States and Canada, the rank classification scheme defined by the American Society of Testing and Materials (ASTM) has become the standard. In this scheme, the properties of **gross calorific value** and **fixed carbon** or **volatile matter content** are used to classify a coal by rank. Gross calorific value is a measure of the energy content of the coal and is usually expressed m units of energy per unit mass. Calorific value increases as the coal proceeds through coalification. Fixed carbon content is a measure of the mass remaining after heating a dry coal sample under conditions specified by the ASTM.

Fixed carbon content increases with coalification. The conditions specified for the measurement of fixed carbon content result in being able alternatively to use the volatile matter content of the coal measured under dry, ash-free conditions as a rank parameter. The rank of a coal proceeds from lignite, the "youngest" coal, through subbituminous, bituminous, and semibituminous, to anthracite, the "oldest" coal. There exist subdivisions within these rank categories which are defined in Table 7.3.2. (Some rank schemes include meta-anthracite as a rank above, or "older" than, anthracite. Others prefer to classify such deposits as graphite. Graphite is a minimal resource and is valuable primarily for uses other than as a fuel.) According to the ASTM scheme, coals are ranked by calorific value up to the high volatile A bituminous rank, which includes coals with calorific values (measured on a moist, mineral matter-free basis) greater than 14,000 Btu/lb (32,564 kJ/kg). At this point, fixed carbon content (measured on a dry, mineral matter-free basis) takes over as the rank parameter. Thus, a high volatile A bituminous coal is defined as having a calorific value greater than 14,000 Btu/lb, but a fixed carbon content less than 69 wt%. The requirement for having two different properties with which to define rank arises because calorific value increases significantly through the lower-rank coals, but very little (in a relative sense) in the higher-ranks, whereas fixed carbon content has a wider range in higher-rank coals, but little (relative) change in the lower-ranks. The most widely used classification scheme outside of North America is that developed under the jurisdiction of the International Standards Organization, Technical Committee 27, Solid Mineral Fuels.

Coal Analysis

The composition of a coal is typically reported in terms of its **proximate analysis** and its **ultimate analysis.** The proximate analysis of a coal is made up of four constituents: volatile matter content, fixed carbon content, moisture content, and ash content, all of which are reported on a weight percent basis. The measurement of these four properties of a coal must be carried out according to strict specifications codified by the ASTM.

Volatile matter in coal includes carbon dioxide, inorganic sulfur- and nitrogen-containing species, and organic compounds. The percentage of various species present depends on rank. Volatile matter content can typically be reported on a number of bases, such as moist; dry, mineral matter-free (dmmf); moist, mineral matter-free; moist, ash-free; and dry, ash-free (daf); depending on the condition of the coal on which measurements were made.

Mineral matter and ash are two distinct entities. Coal does not contain ash, even though the ash content of a coal is reported as part of its proximate analysis. Instead, coal contains mineral matter, which can be present both as distinct mineral entities or inclusions and intimately bound with the organic matrix of the coal. Ash, on the other hand, refers to the solid inorganic material remaining *after combusting* a coal sample. Proximate ash content is the ash remaining after the coal has been exposed to air under specific conditions (ASTM Standard Test Method D 3174). It is reported as the mass percent remaining upon combustion of the original sample on either a dry or moist basis.

Moisture content refers to the mass of water which is released from the solid coal sample when it is heated under specific conditions of temperature and residence time as codified in ASTM Standard Test Method D 3173.

The fixed carbon content refers to the mass of organic matter remaining in the sample after the moisture and volatile matter are released. It is primarily made up of carbon with lesser amounts of hydrogen, sulfur, and nitrogen also present. It is typically reported by difference from the total of the volatile matter, ash, and moisture contents on a mass percent of the original coal sample basis. Alternatively, it can be reported on a dry basis; a dmmf basis; or a moist, mineral matter-free basis.

The values associated with the proximate analysis vary with rank. In general, volatile matter content decreases with increasing rank, while fixed carbon content correspondingly increases. Moisture and ash also decrease, in general, with rank. Typical values for proximate analysis as a function of the rank of a coal are provided in Table 7.3.3.

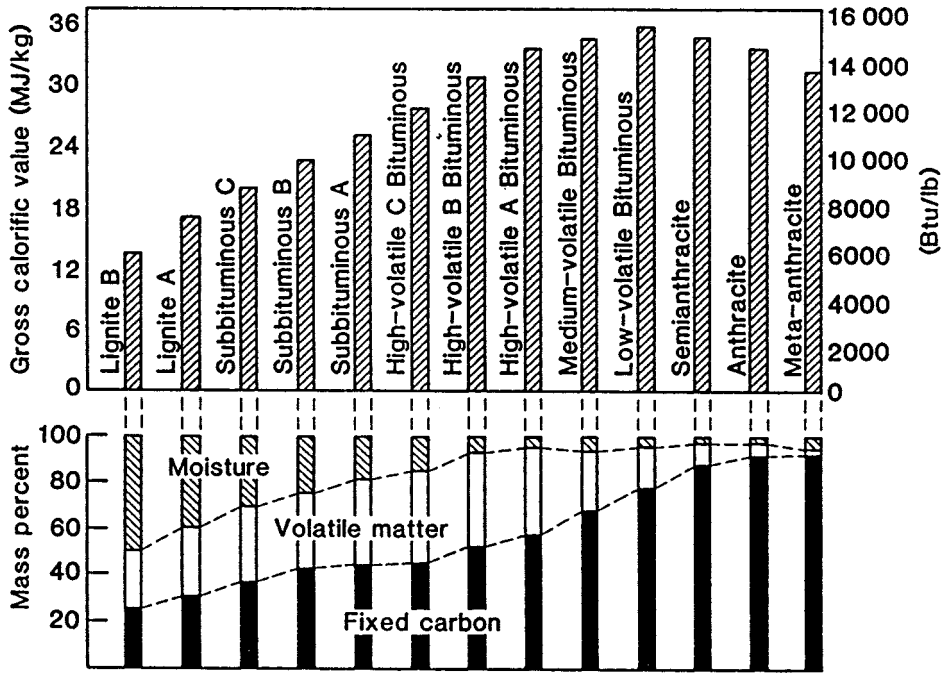
The ultimate analysis of a coal reports the composition of the organic fraction of coal on an elemental basis. Like the proximate analysis, the ultimate analysis can be reported on a moist or dry basis and on

TABLE 7.3.2 Classification of Coals by Rank

Class	Group	Fixed Carbon Limits, % (dmmf)		Volatile Matter limits, % (dmmf)		Gross Calorific Value Limits, Btu/lb (moist, mmf)		Agglomerating Character
		Equal or Greater Than	Less Than	Greater Than	Equal or Less Than	Equal or Greater Than	Less Than	
Anthracitic	Meta-anthracite	98	—	—	2	—	—	Nonagglomerating
	Anthracite	92	98	2	8	—	—	Nonagglomerating
	Semianthracite	86	92	8	14	—	—	Nonagglomerating
Bituminous	Low volatile bituminous	78	86	14	22	—	—	Commonly agglomerating
	Medium volatile bituminous	69	78	22	31	—	—	Commonly agglomerating
	High volatile A bituminous	—	69	31	—	14,000	—	Commonly agglomerating
	High volatile B bituminous	—	—	—	—	13,000	14,000	Commonly agglomerating
	High volatile C bituminous	—	—	—	—	11,500	13,000	Commonly agglomerating
	High volatile C bituminous	—	—	—	—	10,500	11,500	Agglomerating
Subbituminous	Subbituminous A	—	—	—	—	10,500	11,500	Nonagglomerating
	Subbituminous B	—	—	—	—	9,500	10,500	Nonagglomerating
	Subbituminous C	—	—	—	—	8,300	9,500	Nonagglomerating
Lignitic	Lignite A	—	—	—	—	6,300	8,300	Nonagglomerating
	Lignite B	—	—	—	—	—	6,300	Nonagglomerating

From the American Society for Testing and Materials' Annual Book of ASTM Standards. With permission.

TABLE 7.3.3 Calorific Values and Proximate Analyses of Ash-Free Coals of Different Rank



From Averitt, P., Coal Resources of the United States, January 1, 1974. U.S. Geological Survey Bulletin 1412, Government Printing Office, Washington, D.C., 1975.

TABLE 7.3.4 Ultimate Analyses in Mass Percent of Representative Coals of the U.S.

Component	Fort Union Lignite	Power River Subbbituminous	Four Corners Subbbituminous	Illinois C Bituminous	Appalachia Bituminous
Moisture	36.2	30.4	12.4	16.1	2.3
Carbon	39.9	45.8	47 5	60.1	73.6
Hydrogen	2.8	3.4	3.6	4.1	4.9
Nitrogen	0.6	0.6	0.9	1.1	1.4
Sulfur	0.9	0.7	0.7	2.9	2.8
Oxygen	11.0	11.3	9.3	8.3	5.3
Ash	8.6	7.8	25.6	7.4	9.7
Gross calorific value, Btu/lb	6,700	7,900	8,400	10,700	13,400

Modified from Probstein, R. and Hicks, R., *Synthetic Fuels*. McGraw-Hill, New York, 1982. With permission.

an ash-containing or ash-free basis. The moisture and ash reported in the ultimate analyis are found from the corresponding proximate analysis. Nearly every element on Earth can be found in coal. However, the important elements which occur in the organic fraction are limited to a few. The most important of these include carbon, hydrogen, oxygen, sulfur, nitrogen, and, sometimes, chlorine. The scope, definition of the ultimate analysis, designation of applicable standards, and calculations for reporting results on different moisture bases can be found in ASTM Standard Test Method D 3176M. Typical values for the ultimate analysis for various ranks of coal found in the United States are provided in Table 7.3.4.

Coal Properties

Other important properties of coal include swelling, caking, and coking behavior; ash fusibility; reactivity; and calorific value.

Calorific value measures the energy available in a unit mass of coal sample. It is measured by ASTM Standard Test Method D 201 5M, Gross Calorific Value of Solid Fuel by the Adiabatic Bomb Calorimeter, or by ASTM Standard Test Method D 3286, Gross Calorific Value of Solid Fuel by the Isothermal-Jacket Bomb Calorimeter. In the absence of a directly measured value, the gross calorific value, Q, of a coal (in Btu/lb) can be estimated using the Dulong formula (Elliott and Yohe, 1981):

$$Q = 14,544\text{C} + 62,028\left[\text{H} - (\text{O}/8)\right] + 4,050\text{S}$$

where C, H, O, and S are the mass fractions of carbon, hydrogen, oxygen, and sulfur, respectively, obtained from the ultimate analysis.

Swelling, caking, and coking all refer to the property of certain bituminous coals, when slowly heated in an inert atmosphere to between 450 and 550 or 600°F, to change in size, composition, and, notably, strength. Under such conditions, the coal sample initially becomes soft and partially devolatilizes. With further heating, the sample takes on a fluid characteristic. During this fluid phase, further devolatilization causes the sample to swell. Still further heating results in the formation of a stable, porous, solid material with high strength. There are several tests which have been developed based on this property to measure the degree and suitability of a coal for various processes. Some of the more popular are the free swelling index (ASTM Test Method D 720), the Gray-King assay test (initially developed and extensively used in Great Britain), and the Gieseler plastometer test (ASTM Test Method D 2639), as well as a whole host of dilatometric methods (Habermehl et al., 1981). The results of these tests are often correlated with the ability of a coal to form a coke suitable for iron making. In the iron-making process, the high carbon content and high surface area of the coke are utilized to reduce iron oxide to elemental iron. The solid coke must also be strong enough to provide the structural matrix upon which the reactions take place. Bituminous coals which have good coking properties are often referred to as metallurgical coals. (Bituminous coals which do not have this property are, alternatively, referred to as steam coals because of their historically important use in raising steam for motive power or electricity generation.)

Ash fusibility is another important property of coals. This is a measure of the temperature range over which the mineral matter in the coal begins to soften and eventually to melt into a slag and to fuse together. This phenomenon is important in combustion processes; it determines if and at what point the resultant ash becomes soft enough to stick to heat exchanger tubes and other boiler surfaces or at what temperature it becomes molten so that it flows (as slag), making removal as a liquid from the bottom of a combustor possible.

Reactivity of a coal is a very important property fundamental to all coal conversion processes (the most important of which are combustion, gasification, and liquefaction). In general, lower-rank coals are more reactive than higher-rank coals. This is due to several different characteristics of coals, which vary with rank as well as with type. The most important characteristics are the surface area of the coal, its chemical composition, and the presence of certain minerals which can act as catalysts in the conversion reactions. The larger surface area present in lower-rank coals translates into a greater degree of penetration of gaseous reactant molecules into the interior of a coal particle. Lower-rank coals have a less aromatic structure than higher-rank coals, which, along with contributing to larger surface area, also corresponds to a higher proportion of lower-energy, more-reactive chemical bonds. Lower-rank coals also tend to have higher proximate ash contents, and the associated mineral matter is more distributed — down to the atomic level. Any catalytically active mineral matter is thus more highly dispersed, also. However, the reactivity of a coal also varies depending upon what conversion is being attempted. That is, the reactivity of a coal toward combustion (oxidation) is not the same as its reactivity toward liquefaction, and the order of reactivity established in a series of coals for one conversion process will not necessarily be the same as for another process.

Coal Reserves

Coal is found throughout the United States and the world. It is the most abundant fossil energy resource in the United States and in the world. It comprises 95% of United States fossil energy resources and

70% of world fossil energy resources on an energy content basis. All coal ranks can be found in the continental United States. The largest resources in the United States are made up of lignite and subbituminous coals, which are found primarily in the western United States and Alaska. Bituminous coals are found principally in the Midwest, northern Alaska, and the Appalachian region. Principal deposits of anthracite coal are found in northeastern Pennsylvania. The Alaskan coals have not been extensively mined because of their remoteness and the harsh climate. Of the other indigenous coal resources, the anthracite coals have been extensively mined to where there is little economic resource left; the bituminous coals are being heavily mined in the lower 48 states, especially those with low sulfur contents (less than 2.5 wt%); and the low-rank coals in the West have historically been less heavily mined because of their low calorific value, high moisture and ash contents, and their distance from large markets, although with the enactment of the 1990 Amendments to the Clean Air Act of 1970, these coals are replacing high-sulfur Eastern coals in the marketplace. A map showing the general distribution of coal in the lower 48 states is included as Figure 7.3.1.

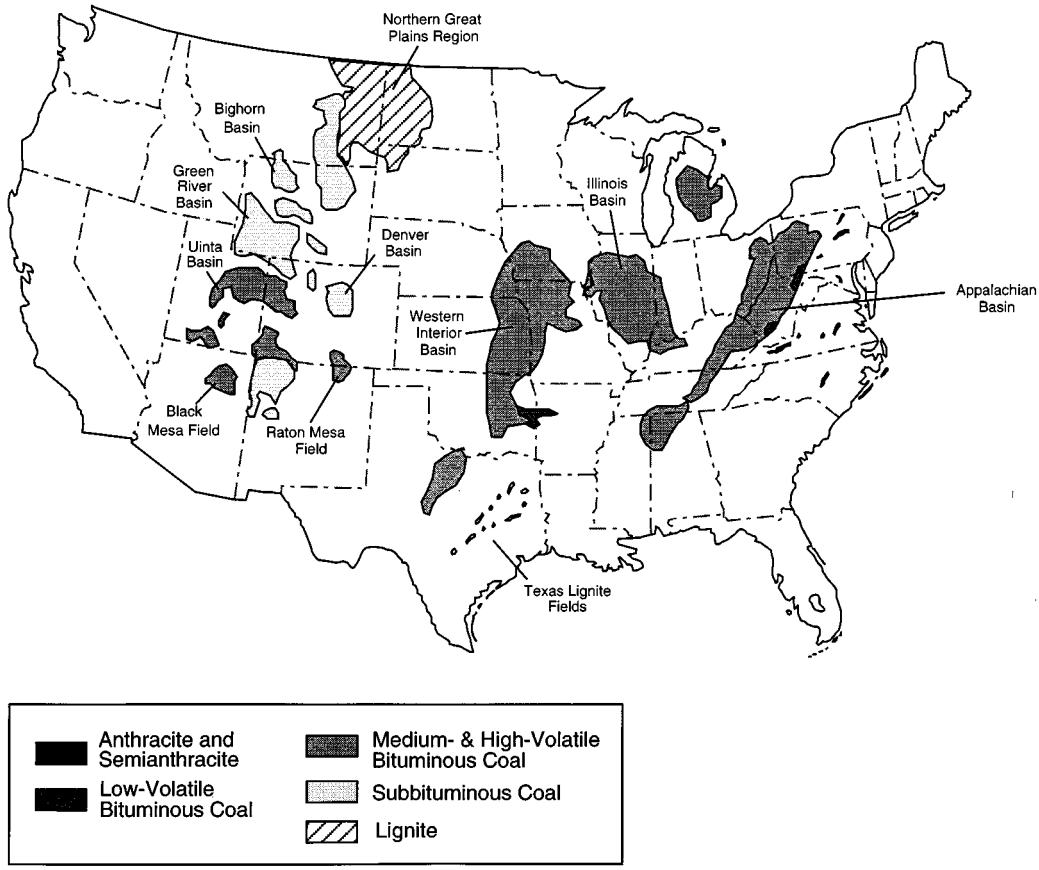


FIGURE 7.3.1 Coal fields of the conterminous (From Averitt, P., Coal Resources of the United States, January 1, 1974. U.S. Geological Survey Bulletin 1412, Government Printing Office, Washington, D.C., 1975.)

The amount of coal (or, for that matter, any minable resource) that exists is not known exactly and is continually changing as old deposits are mined out and new deposits are discovered or reclassified. Estimates are published by many different groups throughout the world. In the United States, the Energy Information Administration, an office within the U.S. Department of Energy, gathers and publishes estimates from various sources, including the World Energy Council and British Petroleum, Ltd. These

latter groups subscribe to a set of definitions for classifying the estimates which are acknowledged both in the United States and the United Kingdom. The most commonly used terms are defined below (as taken from Averitt, (1981):

Resources: The total quantity of coal in the ground within specified limits of thickness of bed overburden. The thickness of overburden defining this term varies from region to region in the world.

Hypothetical Resources: The quantity of coal estimated to be in the ground in unmapped and unexplored parts of known coal basins to an overburden depth of 6000 ft; determined by extrapolation from the nearest areas of identified resources.

Identified Resources: The quantity of coal in the measured, indicated, and inferred resource categories.

Measured Resources: The quantity of coal in the ground based on assured coal bed correlations and on observations spaced about 0.5 mi apart.

Indicated Resources: The quantity of coal in the ground based partly on specific observations and partly on reasonable geological projections; the points of observation and measurement are about 1 mi apart but may be 1.5 mi apart for beds of known continuity.

Inferred Resources: The quantity of coal in the ground based on continuity of coal beds downdip from and adjoining areas containing measured and indicated resources. Generally, inferred resources lie 2 mi or more from outcrops or from points of more precise information.

Demonstrated Reserve Base: The quantity of coal in the ground in the measured and indicated resources categories; restricted primarily to coal in thick and intermediate beds less than 1000 ft below the surface and deemed to be economically and legally available for mining.

Reserves, or Recoverable Reserves: The quantity of coal in the ground that can be recovered from the demonstrated reserve base by application of the recoverability factor.

Recoverability Factor: The percentage of coal in the demonstrated reserve base that can be recovered by established mining practices.

The demonstrated reserve base for coals in the United States as of January 1, 1993, is approximately 474.1 billion (short) tons. It is broken out by rank, state, and mining method (surface or underground) in Table 7.3.5. As of December 31, 1990 (December 30, 1991, for the United States), the world recoverable reserves are estimated to be 1145 billion (short) tons. A breakdown by region and country is provided in Table 7.3.6. The recoverability factor for all coals can vary from approximately 40% to over 90% depending on the individual deposit. The recoverable reserves in the United States represent approximately 56% of the demonstrated reserve base. Thus, the United States contains (by weight) approximately 23% of the recoverable reserves of coal in the world.

Environmental Aspects

Richard Bajura

Along with coal production and use comes a myriad of potential environmental problems, most of which can be ameliorated or effectively addressed during recovery, processing, conversion, or reclamation.

Underground coal reserves are recovered using the two principal methods of room-and-pillar mining (60%) and longwall mining (40%). In room-and-pillar mining, coal is removed from the seam in a checkerboard pattern (the “room”) as viewed from above, leaving pillars of coal in an alternate pattern to support the roof of the mine. When using this technology, generally half of the reserves are left underground. Depending upon the depth of the seam and characteristics of the overburden, subsidence due to the removal of the coal may affect the surface many years after the mining operation is completed. Because of the danger of collapse and movement of the surface, undermined lands are not used as building sites for large, heavy structures.

Longwall mining techniques employ the near-continuous removal of coal in rectangular blocks with a vertical cross section equal to the height of the seam multiplied by the horizontal extent (width) of the panel being mined. As the longwall cutting heads advance into the coal seam, the equipment is

TABLE 7.3.5 Coal Demonstrated Reserve Base, January 1, 1993 (billion short tons)

Region and State	Anthracite	Bituminous Coal[1]		Lignite	Total		
	Underground and Surface[2]	Underground	Surface	Surface[3]	Underground	Surface	Total
Appalachian	7.4	81.8	16.6	1.1	85.7	21.1	106.8
Alabama	0	1.4	2.2	1.1	1.4	3.3	4.7
Kentucky, Eastern	0	7.1	1.5	0	7.1	1.5	8.6
Ohio	0	17.9	6.0	0	17.9	6.0	23.8
Pennsylvania	7.2	20.7	1.1	0	24.6	4.5	29.1
Virginia	0.1	1.6	0.7	0	1.7	0.7	2.5
West Virginia	0	31.8	4.7	0	31.8	4.7	36.5
Other[4]	0	1.2	0.4	0	1.2	0.4	1.6
Interior	0.1	92.7	26.7	13.7	92.8	40.4	133.3
Illinois	0	62.6	15.4	0	62.6	15.4	78.0
Indiana	0	8.9	1.2	0	8.9	1.2	10.1
Iowa	0	1.7	0.5	0	1.7	0.5	2.2
Kentucky, Western	0	16.4	3.8	0	16.4	3.8	20.2
Missouri	0	1.5	4.5	0	1.5	4.5	6.0
Oklahoma	0	1.2	0.3	0	1.2	0.3	1.6
Texas	0	0	0	13.2	0	13.2	13.2
Other[5]	0.1	0.3	1.1	0.5	0.4	1.6	2.0
Western	(6)	140.3	63.7	29.9	140.4	93.6	234.0
Alaska	0	5.4	0.7	(6)	5.4	0.7	6.1
Colorado	(6)	12.0	0.6	4.2	12.0	4.8	16.8
Montana	(6)	71.0	33.2	15.8	71.0	48.9	119.9
New Mexico	0	2.1	2.3	0	2.1	2.3	4.4
North Dakota	0	0	0	9.6	0	9.6	9.6
Utah	0	5.8	0.3	0	5.8	0.3	6.0
Washington	0	1.3	0.1	(6)	1.3	0.1	1.4
Wyoming	0	42.5	26.5	0	42.5	26.5	69.1
Other[7]	0	0.1	0.1	0.4	0.1	0.5	0.6
U.S. Total	7.5	314.8	107.1	44.7	318.9	155.1	474.1
States East of the Mississippi River	7.4	169.9	36.9	1.1	173.8	41.4	215.2
States West of the Mississippi River	0.1	145.0	70.1	43.6	145.1	113.7	258.8

[1] Includes subbituminous coal.
[2] Includes 3,396.4 million short tons of surface mine reserves, of which 3,380.8 million tons are in Pennsylvania and 15.6 million tons are in Arkansas.
[3] There are no underground demonstrated reserves of lignite.
[4] Georgia, Maryland, North Carolina, and Tennessee.
[5] Arkansas, Kansas, Louisiana, and Michigan.
[6] Less than 0.05 billion short tons.
[7] Arizona, Idaho, Oregon, and South Dakota.

Notes: Data represent 100 percent of known measured and indicated coal, with qualifying seam thicknesses and depths, in place as of January 1, 1993. Recoverability varies from less than 40 percent to more than 90 percent for individual deposits. Fifty-six percent of the demonstrated reserve base of coal in the United States is estimated to be recoverable.

Totals may not equal sum of components due to independent rounding.

Sources: Energy Information Administration (EIA), *Coal Production 1992* (October 1993), Tables A1, A2, A3, and A4, and EIA, *U.S. Coal Reserves: An Update by Heat and Sulfur Content* (February 1994, page 23.

TABLE 7.3.6 World Recoverable Reserves of Coal (million short tons)

Region and Country	Anthracite and Bituminous Coal		Subbituminous Coal and Lignite		Total	
	World Energy Council[1]	British Petroleum[1]	World Energy Council[1]	British Petroleum[1]	World Energy Council[1]	British Petroleum[1]
North America	123,741	123,741	152,544	152,342	276,285	276,083
Canada	4,970	4,970	4,535	4,535	9,505	9,505
Mexico	1,380	1,380	516	516	1,896	1,896
United States[2]	117,391	117,391	147,291	147,291	264,682	264,682
Other	—	—	202	—	202	—
Central and South America	6,226	6,226	4,478	4,476	10,703	10,702
Brazil	—	—	2,600	2,600	2,600	2,600
Chile	34	—	1,268	—	1,302	—
Colombia	4,674	4,674	330	330	5,003	5,003
Peru	1,058	—	110	—	1,168	—
Other	460	1,552	169	1,546	629	3,098
Western Europe	32,411	32,334	92,493	74,506	124,904	106,840
Germany	26,366	26,366	61,895	61,895	88,261	88,261
Greece	—	—	3,307	3,307	3,307	3,307
Spain	937	—	661	—	1,598	—
Turkey	179	179	7,701	7,701	7,879	7,879
United Kingdom	3,638	3,638	551	551	4,189	4,189
Yugoslavia	77	—	18,188	—	18,265	—
Other	1,215	2,152	189	1,052	1,404	3,204
Eastern Europe and U.S.S.R.	150,021	150,098	179,436	197,625	329,457	347,723
Bulgaria	33	—	4,079	—	4,112	—
Czechoslovakia	2,061	—	3,858	—	5,919	—
Hungary	657	—	4,260	—	4,917	—
Poland	32,628	32,628	12,787	12,787	45,415	45,415
U.S.S.R.	114,640	114,640	151,017	151,017	265,657	265,657
Other	1	2,830	3,436	33,821	3,437	36,651
Africa	67,024	67,033	1,397	1,397	68,420	68,429
Botswana	3,858	—	—	—	3,858	—
South Africa	60,994	60,994	—	—	60,994	60,994
Swaziland	—	—	1,101	—	1,101	—
Zimbabwe	809	809	—	—	809	809
Other	1,434	5,229	295	1,397	1,730	6,626
Far East, Oceania, and Middle East	188,450	188,523	146,710	146,668	335,160	335,194
Australia	49,979	49,979	50,265	50,251	100,244	100,230
China	68,564	68,564	57,651	57,635	126,215	126,198
India	66,853	66,853	2,094	2,094	68,947	68,947
Indonesia	1,060	1,060	34,283	34,273	35,343	35,334
Japan	912	912	19	19	930	930
Other	1,082	1,156	2,398	2,397	3,479	3,552
World	567,946	567,743	577,057	577,015	1,145,002	1,144,758

[1] See Note 3 at end of Section.
[2] U.S. data are more current than other data on this table. They represent recoverable reserves as of December 31, 1991; data for the other countries are as of December 31, 1990, the most recent period for which they are available. U.S. reserves represent both measured and indicated tonnage. The U.S. term "measured" approximates the term "proved," which is used by the World Energy Council and British Petroleum. The U.S. "measured and indicated" data have been combined prior to depletion adjustments and cannot be recaptured as "measured alone."

— = Not applicable.

Notes: The EIA does not certify the international reserves data but reproduces the information as a matter of convenience for the reader.

Totals may not equal sum of components due to independent rounding.

Source: EIA, *International Energy Annual 1992* (January 1994), Table 37.

automatically moved forward. The roof of the mine collapses behind the shields, and most of the effects of subsidence are observed on the surface within several days of mining. If the longwall mining operation proceeds in a continuous fashion, subsidence may occur smoothly so that little damage occurs to surface structures. Once subsidence has occurred, the surface remains stable into the future. Longwall mining operations may influence water supplies as a result of fracturing of water-bearing strata far removed from the panel being mined.

When coal occurs in layers containing quartz dispersed in the seam or in the overburden, miners are at risk of being exposed to airborne silica dust, which is inhaled into their lungs. Coal workers pneumonoconiosis, commonly called black lung disease, reduces the ability of a miner to breathe because of the effects of fibrosis in the lungs.

Surface mining of coal seams requires the removal of large amounts of overburden, which must eventually be replaced into the excavated pit after the coal resource is extracted. When the overburden contains large amounts of pyrite, exposure to air and water produces a discharge known as acid mine drainage, which can contaminate streams and waterways. Iron compounds formed as a result of the chemical reactions precipitate in the streams and leave a yellow- or orange-colored coating on rocks and gravel in the streambeds. The acid caused by the sulfur in the pyrite has been responsible for significant destruction of aquatic plants and animals. New technologies have been and continue to be developed to neutralize acid mine drainage through amendments applied to the soil during the reclamation phases of the mining operation. Occasionally, closed underground mines fill with water and sufficient pressure is created to cause “blowouts” where the seams reach the surface. Such discharges have also been responsible for massive fish kills in receiving streams.

The potential for acid rain deposition from sulfur and nitrogen oxides released to the atmosphere during combustion is a significant concern. About 95% of the sulfur oxide compounds can be removed through efficient stack gas cleaning processes such as wet and dry scrubbing. Also, techniques are available for removing much of the sulfur from the coal prior to combustion. Combustion strategies are also being developed which reduce the formation and subsequent release of nitrogen oxides.

The potential for greenhouse warming due to emissions of carbon dioxide during combustion (as well as methane during mining and mine reclamation) has also been raised as a significant concern. Since coal is largely composed of carbon with relatively little hydrogen, its combustion leads to a higher level of carbon dioxide emissions per unit of energy released than for petroleum-based fuels or natural gas.

Transportation

Larry Grayson

According to figures compiled by the United States Department of Energy’s Energy Information Administration (Coleman, 1994), 878 million (short) tons of coal were distributed domestically in 1993, primarily via rail (59.6%), followed by barge (13.9%), truck (13.2%), and conveyor and slurry pipeline (10.8%). Shipments to electric utilities amounted to 773.5 million (short) tons during the same year, with 61.5% transported by rail, followed by 14.0% by barge, 11.7% by truck, and 11.0% by conveyor and slurry pipeline.

Villagran (1989) compared the cost advantage of barge transportation over rail transportation. For distances between 100 and 300 mi, the advantage is between $11 and 38/kilo ton-mi; while the advantage drops to $7 to 11/kilo ton-mi for distances greater than 300 mi. For distances less than 100 mi, rail hauls are very expensive, leading to the predominance of truck haulage in this range when river transportation is not possible.

Defining Terms:

Coalification: The physicochemical transformation which coal undergoes after being buried and subjected to elevated temperature and pressure. The classification of a particular coal by rank is a measure of the extent of its coalification. Thus, coalification is a measure of the “age” of a particular coal.

Fixed carbon content: Fixed carbon content is one of the constituents which make up the proximate analysis of a coal. It is normally measured by difference. That is, one measures the volatile matter content and the moisture and ash contents, if the fixed carbon content is reported on a basis containing one or both of those constituents, and subtracts the result(s) from 100% to find the fixed carbon content. One should not confuse the fixed carbon content of a coal with its (elemental) carbon content found in the ultimate analysis. Although there is certainly carbon in the material making up the fixed carbon content, it is not all of the carbon present in the original coal, and other elements are also present.

Gross calorific value: Calorific value is a measure of the energy content of a material, in this case, a coal sample. Calorific value is measured by ASTM Standard Test Method D 2015M, Gross Calorific Value of Solid Fuel by the Adiabatic Bomb Calorimeter, or by ASTM Standard Test Method D 3286, Gross Calorific Value of Solid Fuel by the Isothermal-Jacket Bomb Calorimeter. The *gross* calorific value takes into account the additional heat gained by condensing any water present in the products of combustion, in contrast to the *net* calorific value, which assumes that all water remains in the vapor state.

Maceral: Macerals are organic substances or optically homogeneous aggregates of organic substances in a coal sample that possess distinctive physical and chemical properties.

Proximate Analysis: Proximate analysis is a method to measure the content of four separately identifiable constituents in a coal: volatile matter content, fixed carbon content, moisture content, and ash content, all of which are reported on a weight percent basis. The standard method for obtaining the proximate analysis of coal or coke is defined by the ASTM in Standard Test Method D 3172.

Rank: Coal rank is a classification scheme for coals which describes the extent of coalification which a particular coal has undergone. The structure, chemical composition, and many other properties of coals vary systematically with rank. The standard method for determining the rank of a coal sample is defined by the ASTM in Standard Test Method D 388.

Type: Coal type is a classification scheme for coals which references the original plant material from which the coal was derived.

Ultimate Analysis: Ultimate analysis is a method to measure the elemental composition of a coal sample. Typical ultimate analyses include carbon, hydrogen, oxygen, sulfur, and nitrogen contents, but other elements can also be reported. These other elements are usually not present to any appreciable extent. However, if they are reported, the sum of all the elements reported (including moisture and ash content) should equal 100%. The standard method for the ultimate analysis of coal or coke is defined by the ASTM in Standard Test Method D 3176.

Volatile matter content: Volatile matter content measures the mass of material released upon heating the coal sample under specific conditions, defined by the ASTM Standard Test Method D 3175.

References

Averitt, P. 1981. Coal resources, in *Chemistry of Coal Utilization, Second Supplementary Volume*, M.A. Elliott, Ed., John Wiley & Sons, New York, 55–90.

Coleman, L.L. 1994. Chapter III, in *Coal Data 1994*, National Coal Association, III-9, III-19, and III-20.

Conaway, D. 1994. Panel discussion on independent contractors, *Proc. 25th Annual Institute on Mining Health, Safety and Research*, G.R. Tinney, A. Bacho, and M. Karmis, Eds., Virginia Polytechnic Institute and State University, Blacksburg, VA, 154.

Elliott, M.A. and Yohe, G.R. 1981. The coal industry and coal research and development in perspective, in *Chemistry of Coal Utilization, Second Supplementary Volume*, M.A. Elliott, Ed., John Wiley & Sons, New York, 26.

Habermehl, D., Orywal, F., and Beyer, H.-D. 1981. Plastic properties of coal, in *Chemistry of Coal Utilization, Second Supplementary Volume*, M.A. Elliott, Ed., John Wiley & Sons, New York, 319–328.

Neavel, R. 1981. Origin, petrography, and classification of coal, in *Chemistry of Coal Utilization, Second Supplementary Volume*, M.A. Elliott, Ed., John Wiley & Sons, New York, 91–158.

Villagran, R.A. 1989. *Acid Rain Legislation: Implications for the Coal Industry*, Shearson, Lehman, Hutton, New York, 37–39.

Further Information

An excellent resource for understanding coal, its sources, its uses, and its limitations and potential problems is the book by Elliott referenced above under Averitt (1981), Habermehl et al. (1981), and Neavel (1981). A reader wishing an understanding of coal topics could find no better resource.

Another comprehensive book which includes more-recent information but is not quite as weighty as Elliott's (664 pages vs. 2374 pages) is *The Chemistry and Technology of Coal*, edited (second edition, revised and expanded) by James G. Speight.

For up-to-date information specific to the environmental problems associated with the use of coal, the reader is referred to Norbert Berkowitz's chapter entitled "Environmental Aspects of Coal Utilization" in *An Introduction to Coal Technology.*

For information on the standards for coal analyses and descriptions of the associated procedures, the reader is referred to any recent edition of the American Society for Testing and Materials's *Annual Book of ASTM Standards.* Section 5 covers petroleum products, lubricants, and fossil fuels, including coal and coke.

Oil

Philip C. Crouse, P.E.

Overview

Fossil fuel energy accounted for 86.3% of all world energy in 1990. The Energy Information Administration (EIA) of the U.S. Department of Energy estimates that in the year 2010, fossil fuels will account for 85.9% of all world energy consumption — only a 0.4% percentage decrease in usage (Figure 7.3.2). According to EIA estimates, coal is expected to decline slightly from about a 27% to about a 25% share of consumption, and consumption of natural gas is expected to increase from 21 to 24% over the 20-year period. Over the same period, oil is forecasted to continue to be world major energy source with only slight declines from the present 39% of consumption.

Recent efforts in the United States have been to foster growth in natural gas usage as an energy source, causing an estimated growth of 2.3% per year. Total energy usage is expected to grow from 345.6 to 476.0 quadrillion Btu — or a 38% growth in energy usage over 20 years.

Crude Oil Classification and World Reserves

Obtaining accurate estimates of world petroleum and natural gas resources and reserves is difficult and uncertain. Terminology used by industry to classify resources and reserves has no broadly accepted standard classification. Such classifications have been a source of controversy in the international oil and gas community. Confusion persists in regard to classification. This section uses information provided by the Department of Energy classification system. The next chart shows the relationship of resources to reserves. **Recoverable resources** include discovered and undiscovered resources. **Discovered resources** are those resources that can be economically recovered (Figure 7.3.3).

Discovered resources include all production already out of the ground and reserves. Reserves are further broken down into proved reserves and other reserves. Again, there are many different groups that classify reserves in different ways, such as *measured*, *indicated*, *internal*, *probable*, and *possible.* Most groups break reserves into producing and nonproducing categories. Each of the definitions is quite voluminous and the techniques for qualifying reserves vary globally.

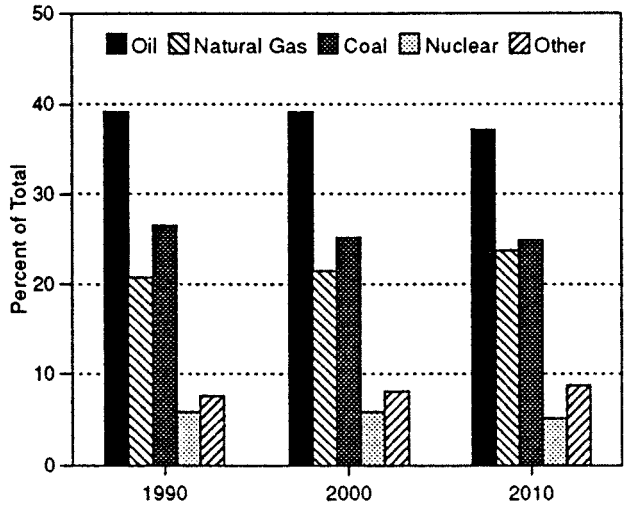
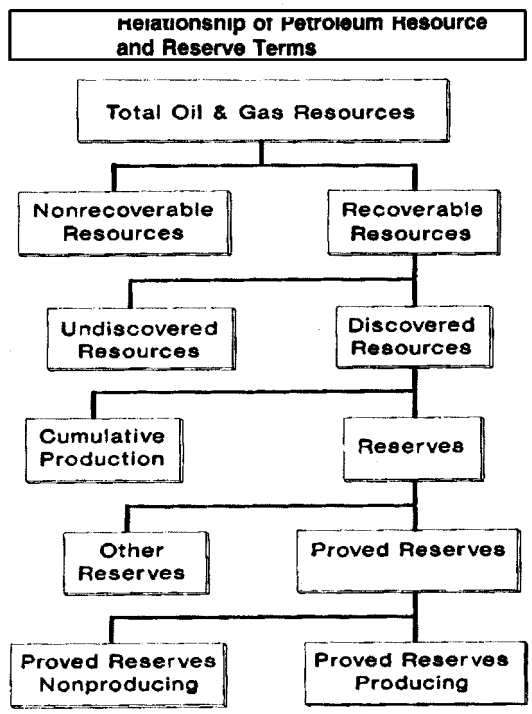


FIGURE 7.3.2 Share of world energy consumption by primary energy source, 1990–2010. (From EIA, *International Energy Outlook 1994*, EIA, Washington, D.C., July 1994.)



Source: Energy Information Administration, Office of Oil and Gas.

FIGURE 7.3.3 Relationship of petroleum resource and reserve terms. (Courtesy of Energy Information Administration, Offiice of Oil and Gas.)

TABLE 7.3.7 Oil Reserves (Year End) — million barrels

	1970	1980	1990
North America	53,160	80,053	84,134
South America	25,557	26,922	69,686
Western Europe	5,698	18,698	21,289
Eastern Europe	59,880	84,140	65,412
Middle East	342,134	357,578	601,987
Africa	46,356	52,650	59,159
Asia-Pacific	21,973	38,517	56,151
Total	554,777	658,557	957,818
OPEC	404,441	428,139	715,502
Non-OPEC	150,336	230,418	242,316

Source: EIA, *International Oil and Gas Exploration and Development 1991*, Washington, D.C., December 1993, 36–39.

TABLE 7.3.8 Annual Oil Production — million barrels

	1970	1980	1990
North America	4,157	4,379	4,136
South America	1,738	1,331	1,574
Western Europe	116	869	1,478
Eastern Europe	2,706	4,550	4,204
Middle East	5,063	6,760	6,120
Africa	2,246	2,265	2,367
Asia-Pacific	651	1774	2371
Total	16,678	21,928	22,049
OPEC	8,545	9,839	8,645
Non-OPEC	8,133	12,089	13,604

Source: EIA, International Oil and Gas Exploration and Development 1991, Washington, D.C., December 1993, 30–33.

Proved reserves are generally defined as: "Those volumes of oil and gas that geological and engineering data demonstrate with reasonable certainty to be recoverable in future years from known reservoirs under existing economic and operating conditions."

OPEC (the Organization of Petroleum-Exporting Countries) has been key in setting global fossil fuel prices over the last two decades. With very large reserves, OPEC can provide much of the world future needs for crude oil and petroleum products. About two-thirds of the world known petroleum reserves are located in the Middle East as shown in Table 7.3.7.

Table 7.3.8 shows that the annual world crude oil production has steadily grown from 16.7 billion barrels in 1970 to 22 billion barrels in 1990.

Both crude oil demand and production are forecast to increase over the next 20 years. OPEC production is relatively level at 8.6 billion barrels in 1990 compared with 8.5 billion barrels in 1970. During the same time, non-OPEC production increased from 8.1 to 13.6 billion barrels. As the "swing producer", OPEC's production in 1980 increased by over 1 billion barrels when non-OPEC production could not meet total demand. They then decreased production by a similar amount in 1990 when production in the rest of the world increased by 1 billion to a non-OPEC total of 13.6 billion barrels. With a low price environment, OPEC is expected to gain market share in global production over the next 20 years.

Standard Fuels

Petroleum is refined into petroleum products that are used to meet individual product demands. The general classifications of products are

1. NATURAL GAS LIQUIDS AND LIQUEFIED REFINERY GASES
   This category includes ethane (C_2H_6), ethylene (C_2H_4), propane (C_3H_8), propylene (C_3H_6), butane and isobutane (C_4H_{10}), and butylene and isobutylene (C_4H_8).
2. FINISHED PETROLEUM PRODUCTS
   This category includes motor gasoline, aviation gasoline, jet fuel, kerosene, distillate, fuel oil, residual fuel oil, petrochemical feed stock, naphthas, lubricants, waxes, petroleum coke, asphalt and road oil, and still gas.
   - *Motor gasoline* includes reformulated gasoline for vehicles and oxygenated gasoline such as gasohol (a mixture of gasoline and alcohol).
   - *Jet fuel* is classified by use such as industrial or military and naphtha and kerosene-type. Naphtha fuels are used in turbo jet and turbo prop aircraft engines and excludes ram-jet and petroleum rocket fuel.
   - *Kerosene* is used for space heaters, cook stoves, wick lamps, and water heaters.
   - *Distillate fuel oil* is broken into subcategories: No. 1 distillate, No. 2 distillate, and No. 4 fuel oil which is used for commercial burners.
   - *Petrochemical feedstock* is used in the manufacture of chemicals, synthetic rubber, and plastics.
   - *Naphthas* are petroleums with an approximate boiling range of 122 to 400°F.
   - *Lubricants* are substances used to reduce friction between bearing surfaces, used as process materials, and as carriers of other materials. They are produced from distillates or residues. Lubricants are paraffinic or naphthenic and separated by viscosity measurement.
   - *Waxes* are solid or semisolid material derived from petroleum distillates or residues. They are typically a slightly greasy, light colored or translucent, crystallizing mass.
   - *Asphalt and road oil*. Asphalt is a cementlike material containing bitumens. Road oil is any heavy petroleum oil used as a dust pallatine and road surface treatment.
   - *Still Gas* is any refinery by-product gas. It consists of light gases of methane, ethane, ethylene, butane, propane, and the other associated gases. Still gas typically used as a refinery fuel.

World Refining Capacity. Refining capacity grew from 48 million barrels per day in 1970 to about 75 million barrels per day in 1990 — a 55% growth in capacity. Table 7.3.9 shows world refining capacity beginning in 1970. The peak year was 1982 in which capacity was 81.4 million barrels per day. Utilization of refinery capacity was about 80% in 1990, pointing to underutilization.

TABLE 7.3.9 World Refining Capacity

	1970	1980	1990
North America	13.2	20.2	17.4
Latin America	4.8	8.6	7.2
Western Europe	14.7	20.3	14.1
Middle East	2.2	3.6	4.2
Africa	0.7	1.7	2.6
Asia-Pacific	5.1	10.4	10.3
Central Planned Economies	7.5	15.4	17.6
Total world	48.2	80.0	73.4

Natural Gas

Philip C. Crouse, P.E.

Natural gas has been called the environmentally friendly fossil fuel since it releases fewer harmful contaminants. World production of dry natural gas was 73.7 trillion ft^3 and accounted for over 20% of world energy production. In 1990 Russia accounted for about one third of world natural gas. The second largest producer was the United States having about one quarter of world 1990 natural gas production.

Natural Gas Production Measurement

Natural gas production is generally measured as "dry" natural gas production. It is determined as the volume of natural gas withdrawn from a reservoir less (1) the volume returned for cycling and repressuring reservoirs; (2) the shrinkage resulting from the removal of lease condensate and plant liquids; (3) the nonhydrocarbon gases. The parameters for measurement are 60°F and 14.73 lb standard per square inch absolute.

World Production and Reserves of Dry Natural Gas

From 1983 to 1992, dry natural gas production rose from 54.4 to 75 trillion ft^3. The breakdown by region of world is shown in Table 7.3.10.

TABLE 7.3.10 World Dry Natural Gas Production — trillion ft^3

	1983	1992
North, Central, and South America	21.20	25.30
Western Europe	6.20	7.85
Eastern Europe and former U.S.S.R.	21.09	28.60
Middle East and Africa	2.95	6.87
Far East and Oceania	2.96	6.38
World total	54.40	75.00

Source: EIA, *Annual Energy Review 1993*, EIA, Washington, D.C., July 1994, 305.

World natural gas reserves estimated by the *Oil and Gas Journal* as of December 31, 1991 are in Table 7.3.11. OPEC accounted for 40% of world reserves yet processes only about 12% of the world production. The former U.S.S.R. accounts for about 40% and Iran another 15% of world reserves.

TABLE 7.3.11 World Natural Gas Reserves — billion ft^3

North America	343,677
South America	166, 850
Western Europe	177,844
Eastern Europe	1,766,358
Middle East	1,319,823
Africa	310,241
Asia-Pacific	299,288
Total	4,384,081
OPEC	1,729,205
Non-OPEC	2,654,876

Compressed Natural Gas

Environmental issues have countries examining and supporting legislation to subsidize the development of cleaner vehicles that use compressed natural gas (CNG). Even with a push toward the use of CNG-burning vehicles, the numbers are quite small when compared with gasoline vehicles. Italy has used

CNG since 1935 and has the largest usage with 300,000 vehicles. The United States ranked fifth with an estimated 30,000 vehicles in 1994. Argentina, which ranked sixth, had 15,000 vehicles.

Liquefied Natural Gas (LNG)

Natural gas can be liquefied by lowering temperature until a liquid state is achieved. It can be transported by refrigerated ships. The process of using ships and providing special-handling facilities adds significantly to the final LNG cost. If oil prices stay low, prospects for LNG development will remain low in the future. However, LNG projects planned by OPEC member countries may become significant over the next 20 years with shipments of LNG exports ultimately accounting for up to 25% of all gas exports.

Physical Properties of Hydrocarbons

The most important physical properties from a crude oil classification standpoint are density or specific gravity and the viscosity of liquid petroleum. Crude oil is generally lighter than water. A Baume-type scale is predominantly used by the petroleum industry and is called the **API** (American Petroleum Institute) **gravity** scale (see Table 7.3.12). It is related directly to specific gravity by the formula:

$$\varphi = (141.5)/(131.5 + {}^{\circ}\text{API})$$

where φ = specific gravity. Temperature and pressure are standardized at 60°F and 1 atm pressure.

TABLE 7.3.12 Relation of API Gravity, Specific Gravity, and Weight per Gallon of Gasoline

Degree API	Specific Gravity	Weight of gallon in lbs.
8	1.014	8.448
9	1.007	8.388
10	1.000	8.328
15	0.966	8.044
20	0.934	7.778
25	0.904	7.529
30	0.876	7.296
35	0.850	7.076
40	0.825	6.870
45	0.802	6.675
50	0.780	6.490
55	0.759	6.316
58	0.747	6.216

Note: The specific gravity of crude oils ranges from about 0.75 to 1.01.

Other key physical properties involve the molecular weight of the hydrocarbon compound and the boiling point and liquid density. Table 7.3.13 shows a summation of these properties.

Defining Terms

API Gravity: A scale used by the petroleum industry for specific gravity.

Discovered resources: Discovered resources include all production already out of the ground and reserves.

Proved resources: Resources that geological and engineering data demonstrate with reasonable certainly to be recoverable in future years from known reservoirs under existing economic and operating conditions.

Recoverable resources: Recoverable resources include discovered resources.

TABLE 7.3.13 Other Key Physical Properties of Hydrocarbons

Compound	Molecular Weight	Boiling Point at 14.7 psia in °F	Liquid Density at 14.7 psia and 60°F-lb/gal
Methane	16.04	–258.7	2.90
Ethane	30.07	–125.7	4.04
Propane	44.09	–43.7	4.233
Isobutane	58.12	10.9	4.695
n-Butane	58.12	31.1	4.872
Isopentane	72.15	82.1	5.209
n-Pentane	72.15	96.9	5.262
n-Hexane	86.17	155.7	5.536
n-Heptane	100.2	209.2	5.738
n-Octane	114.2	258.2	5.892
n-Nonane	128.3	303.4	6.017
n-Decane	142.3	345.4	6.121

Further Information

The Energy Information Agency of the U.S. Department of Energy, Washington, D.C., publishes *International Energy Outlook* periodically.

7.4 Biomass Energy

Michael C. Reed, Lynn L. Wright, Ralph P. Overend, and Carlton Wiles

Biomass energy encompasses a wide variety of renewable energy technologies that use plant matter, plant residues, or plant-derived process wastes as fuel. These biomass resources can be used directly as solid fuels to produce heat, or they can be converted to other energy carriers such as liquid and gaseous fuels and electricity. Because the energy in biomass is less concentrated than the energy in fossil fuels, high-efficiency conversion technologies are necessary to make this energy resource cost competitive.

Photosynthesis

Biomass fuels are derived from green plants which capture solar energy and store it as chemical energy through the **photosynthetic** reduction of atmospheric carbon dioxide. Plant leaves are biological solar collectors while the stems, branches, and roots are the equivalent of batteries storing energy-rich complex carbon compounds. Elemental analysis shows both wood and grasses to be about 50% carbon. The average photosynthetic efficiency of converting solar energy into organic carbon compounds on an annual basis varies from less than 0.5% in temperate and tropical grasslands to about l.5% in moist tropical forests (Cralle and Vietor, 1989). While quite low, it represents stored energy on an annual basis and the diversity and adaptability of plants allows this solar collection to occur on most parts of Earth's surface. The worldwide annual storage of photosynthetic energy in terrestrial biomass is huge, representing approximately ten times world annual use of energy (Hall et al., 1993).

Biomass Production, Yield, and Potential

Forest Land

The amount of harvestable woody biomass produced by natural forests on an annual basis ranges from about 2 to 6 dry t/ha/year (metric ton/hectare/year), or 0.9 to 2.7 dry ton/acre/year with the higher yields usually in tropical regions. Managed forests often accumulate biomass at approximately double the rate of natural unmanaged stands. The productivity rates of natural forests could be increased to 4 to 12 dry t/ha/year (1.8 to 5.4 dry t/acre/year) if brought under active management. Such management might include optimizing harvesting strategies for production, fertilization, and replanting natural stands with faster-growing species. At present 10% of world forests, or 355 million ha (876 million acres), are actively managed. If managed forests were increased to 20%, and if 20% of the harvested material were used for energy, the annual worldwide resource of wood for energy from currently forested land would amount to somewhere between 284 and 852 million t (315 to 946 million dry tons) of available wood or about 5.6 to 17 Exajoules (5 to 14.5 quadrillion Btu) of primary energy based on potential yield ranges of managed forests. The most optimistic estimates of biomass potential would provide a total primary energy resource of about 30 Exajoules from managed forests (Sampson et al., 1993).

Cropland and Grassland

The production of perennial plants (either woody crops grown for 3 to 10 years between harvests or perennial grasses), using genetically superior materials, established on agricultural land and managed as an agricultural crop is believed to be the method of choice for producing biomass for energy on a large scale. Temperate shelterbelts and tropical agroforestry plantings could also contribute greatly to biomass energy resources. Energy crop yields are highest in locations where genetically superior material is planted on land with plenty of access to sunshine and water. Operational yields in the range of 20 to 30 dry t/ha/year (9 to 13 dry ton/acre/year) are achievable now. Such yields have been observed over large areas with eucalyptus and tropical grasses in the moist tropics and subtropics, and with hybrid poplars in irrigated portions of the dry Pacific Northwest of the United States. In many areas, however, suitable genetically superior materials are not yet available, water is a limiting factor, and irrigation may not be an economically or environmentally desirable alternative. Temperate wood energy yields are more

commonly in the range of 9 to 13 dry t/ha/year (4 to 6 dry ton/acre/year) today with the expectation that they could increase to the 12 to 20 dry t/ha/year (5.5 to 9 dry ton/acre/year) relatively rapidly with continued genetic improvement and clonal selection (Wright, 1994). Yields of properly managed perennial grasses in field test plots are currently in the range of 11 to 22 dry t/ha/year (5 to 10 dry ton/acre/year) with the expectation that yields of 16 to 27 dry t/ha/year (7 to 12 dry ton/ac/year) could be attained with further breeding and improvement of management regimes. The best current perennial grass yields on nonirrigated test plots have averaged around 25 dry t/ha/year (11 dry ton/acre/year) for 6 years in the southeastern U.S. where a long growing season favors warm season grasses.

The total amount of biomass potentially available from cropland conversion can only be speculated on since it requires making a number of assumptions involving such issues as land availability, food crop requirements, development of biomass energy markets, and crop yields. Such speculations have been made to attempt to evaluate the potential effect of biomass energy use on carbon emission reduction. These speculations have assumed that 10 to 15% of cropland worldwide could be available for the production of biomass resources as energy crops, shelterbelts, or for agroforestry use with energy as one product (Hall et al., 1993; Sampson et al., 1993). The primary energy resource potential of cropland conversion has been estimated to range from a low of 18 Exajoules to a high of 49 Exajoules (Sampson et al., 1993). Another 25 to 110 Exajoules has been estimated to be available from the conversion of grasslands and degraded areas to the production of biomass energy resources worldwide (Sampson et al., 1993).

Biomass Residues

Biomass residues are the organic by-products of food, fiber, and forest production and processing. Major sources of residues include the residues of grain crops such as corn, wheat, rice; animal dung; and forest roundwood harvest and processing. A signifiant portion of residues is not economically collectible for energy because of wide dispersal and low bulk density, which makes recovery, transport, and storage too costly. In many cases, residues have greater economic value being left on the land to restore nutrients, reduce erosion, and stabilize soil structure, or they may be recovered for other domestic, industrial, or agricultural uses. Recoverable crop residues in the United States are estimated to range between 70 and 190 million dry tons (Day, 1989) and recoverable forest residues are about double the level of crop residues (Hall et al., 1993). Recoverable crop and forest residues worldwide are estimated to have an energy value of 12.5 and 13.5 Exajoules respectively (Hall et al., 1993).

Terrestrial Limitations

The biomass productivity of forests, grasslands, and energy crop plantings is considerably less than theoretically possible based on calculations of maximum photosynthetic potential. Under artificially optimal conditions of temperature, water, and nutrients, average solar radiation near Des Moines, Iowa, would theoretically be aufficient to produce maximum total tree dry weight yields (above and below ground) as high as 102 t/ha/year (45 ton/acre/year) (Sampson et al., 1993). Actual seasonal temperature variations alone, however, reduce the maximum yield potential to 60 t/ha/year (27 ton/acre/year). Of that, the usable biomass portion of trees (the stems and branches only) amounts to only about 35 t/ha/year (15 ton/acre/year). Yet this is many times the harvestable biomass dry weight yield of trees from natural forest stands and more than double the best tree yields observed in experimental trials in that region. Water stress is a major reason for reduced yields. Perennial grasses may have a slightly higher harvestable yield potential in the region because of their higher water use efficiency and the more-efficient carbon metabolism of warm season grasses in temperate climates (Jones, 1985). While water stress, pests outbreaks, and inadequate nutrients all contribute to reduced yields, it is also clear that the genetic potential of perennials is yet to be fully explored. Genetic improvement of perennials for energy is only beginning in most regions of the United States; thus, the potential for finding substantially higher-yielding varieties and clones is excellent.

Environmental Impact

An emerging consensus of environmentalists, forestry organizations, researchers, and others is that if energy crops are treated as an agricultural crop and established and tended in a considered and informed way on appropriate land, environmental damage can be avoided. In fact, preliminary data suggests that there can be significant environmental and ecological benefits achieved in association with the development of a fully sustainable energy resource. Global benefits include the reduced emission of carbon into the atmosphere, while local benefits include soil conservation, reduction of chemical use, and wildlife biodiversity in an agricultural setting.

Biomass Conversion Technologies

As noted previously, biomass can be used to produce heat, electricity, and prepared fuels such as charcoal and liquid fuels. To do this, the biomass conversion technologies have to treat a wide range of biomass resources efficiently and convert them into the desired products. For example, biomass resources may include fuelwood, agricultural residues, farm animal wastes, forest industry processing residues (e.g., sawdust, black liquor from pulping processes), agricultural processing by-products (e.g., bagasse from sugarcane processing, food-processing by-products such as olive pits), and crops (e.g., grains, oil seeds, energy crops).

Each of these resources has different handling requirements and different process characteristics. Worldwide, the most common biomass resources are fuelwood and agricultural residues such as cereal straws. However, many other biomass resources offer opporturuties notjust for heat, but also in improving the environment (as when animal wastes are converted into methane and fertilizer through anaerobic fermentation processes). Each conversion process must take into account the differences in biomass composition which can vary widely in terms of moisture and ash content.

Biomass Composition

As received, biomass can range from very clean wood chips at 50% moisture to urban wood residues that are dry but contaminated with ferrous and other materials, animal residues, sludges, and the organic component of murucipal solid waste (MSW). Most biomass is a **lignocellulosic** material, such as wood or straw, which is composed of **cellulose**, **hemicellulose**, and **lignin**. Although different plant species have differing proportions of these polymers, it turns out that on a moisture and ash-free basis, the majority have essentially the same calorific values in the range of 16 to 19.6 GJ/t (8000 to 8500 Btu/lb). Biomass ash content varies considerably, being lowest in the clean wood fraction and very high in cereal straws such as rice straw, as can be seen in Table 7.4.1.

TABLE 7.4.1 Some Fuel Properties of Four Different Biomass Types

Property	Pine Shavings	Switchgrass	Rice Hull	Rice Straw
Ash %	1.43	10.10	18.34	15.90
Carbon	48.54	47.79	40.96	41.78
Hydrogen	5.85	5.76	4.30	4.63
Nitrogen:	0.47	1.17	0.40	0.70
Sulfur	0.01	0.10	0.02	0.08
Oxygen	43.69	35.07	35.86	36.57
Btu/lb	8337	7741	6944	7004
GJ/t	19.38	17.99	16.14	16.28

Direct Combustion, Combined Heat, and Power and Steam Electric Generation

By far the most frequent use of biomass is in direct-combustion applications, with combustion devices ranging from 5 kW cook stoves (Prasad, 1985) to 125 MW boilers burning wood chips and producing 150 thm/hr of steam at 11 MPa and 540°C superheat temperatures (Anonymous, 1994). The available types of combustors range from pile burners or Dutch-oven-type burners, which can accept biomass in

a large size range, to units that are suspension fired and require not only extensive fuel preparation but also fuel drying. The most popular types of boilers seem to be stoker-fired units with moving grates. More recently, the trend has been to bubbling and circulating fluidized-bed units, which offer a high degree of fuel flexibility by accepting fuel mixes ranging from coal to paper mill sludges that have very high moisture and ash contents. Boiler efficiency for typically moist biomass feedstocks ranges from 65 to 75%.

Biomass-fired steam turbine power generation technology is very similar to that of conventional fossil-fueled cycles. Biomass is combusted in a boiler, and the heat which is liberated is used to produce steam with conditions ranging from 28 bar and 398°C (400 psig and 750°F) for typical 25 MW units to 86 bar and 510°C (1250 psig and 950°F) for larger systems. The superheated steam generated in the boiler is used to drive a steam turbine generator. Net plant efficiency ranges from 20 to 30% (17,060 to 11,373 Btu/kWhr). Larger-scale boilers would permit several stages of reheat and the use of supercritical steam cycles. Developers of the Whole Tree Energy™ System have proposed the combustion of whole trees that are dried in a flue gas-assisted dome in order to provide supercritical steam to a turbine. System efficiencies are projected to be in the range of 40% (heat rate 8530 Btu/kWhr) and would be comparable to gas turbine-based combined-cycle units. The ability to utilize whole trees, as opposed to having to process wood feedstocks to "chip" size is a key advantage of the Whole Tree Energy System, as it can significantly reduce feedstock-handling costs.

Gasification

Biomass gasification processes can convert biomass materials into gaseous fuels (carbon- and hydrogen-rich fuel gases) that can be more easily utilized, often with a gain in efficiency and environmental performance. Gasification is a two-step, endothermic process in which a solid fuel (biomass or coal) is thermochemically converted into a low- or medium-Btu gas (see Figure 7.4.1). In the first reaction, pyrolysis, the volatile components of the fuel are vaporized at temperatures below 600°C (1100°F) by a set of complex reactions. Included in the volatile vapors are hydrocarbon gases, hydrogen, carbon monoxide, carbon dioxide tar, and water vapor. Because biomass fuels tend to have more volatile components (70 to 86% on a dry basis) than coal (30%), pyrolysis plays a larger role in biomass gasification than in coal gasification. Char (fixed carbon) and ash are the by-products of pyrolysis which are not vaporized. The second reaction, char conversion, involves the gasification and/or combustion of the carbon that remains after pyrolysis. In this reaction a portion of the char is burned to provide the heat required for pyrolysis and for gasification of any remaining char.

Thermal gasification is typically 80 to 85% efficient in converting the organic content of the feed into a fuel gas mixture mainly composed of hydrogen, carbon monoxide, and methane along with inert constituents. As a comparison, natural gas energy content, measured on a dry gas basis, is typically 37.8 $MJ \cdot m^{-3}$, or 1000 Btu/ft^3. Biomass gases produced through thermal gasification with air as the oxidant typically have gas energy content of less than 5.6 $MJ \cdot m^{-3}$ or 150 Btu/ft^3, while gasification with oxygen and steam or indirect gasification (which elimmintes the nitrogen dilution of the product) results in a heating value of 13 to 17 $MJ \cdot m^{-3}$ or about 350 to 450 Btu/ft^3. Typically, an air-blown gasifier would have over 50% nitrogen, a 25 to 30% by volume mix of hydrogen and carbon monoxide, and a small amount of methane, and the balance would be carbon dioxide.

Gasifier systems usually comprise the biomass fuel-handling and -feeding system which is coupled by means of air locks to the gasifier. In the thermal processes, the gasifier is usually a refractory-lined vessel and the gasification is carried out at high temperatures of approximately 850°C, at either atmospheric or elevated pressures. In a fluidized-bed gasifier, a continuous feed of biomass and inert heat-distributing material (e.g., sand) is "fluidized" by an oxidant and/or steam, and heat is supplied to the gasifier either "directly" or "indirectly" (U.S. Department of Energy, 1992).

In a directly heated fluidized-bed gasifier, heat required for gasification comes from char combustion in the gasifier reactor. The Institute of Gas Technology RENUGAS™ Pressurized Gasifier is an example of a direct-gasification technology. In the RENUGAS process, biomass is fed into a single, fluidized-bed gasifier vessel that operates at 300 to 500 psig. Inert solids in the vessel form the stable fluidized

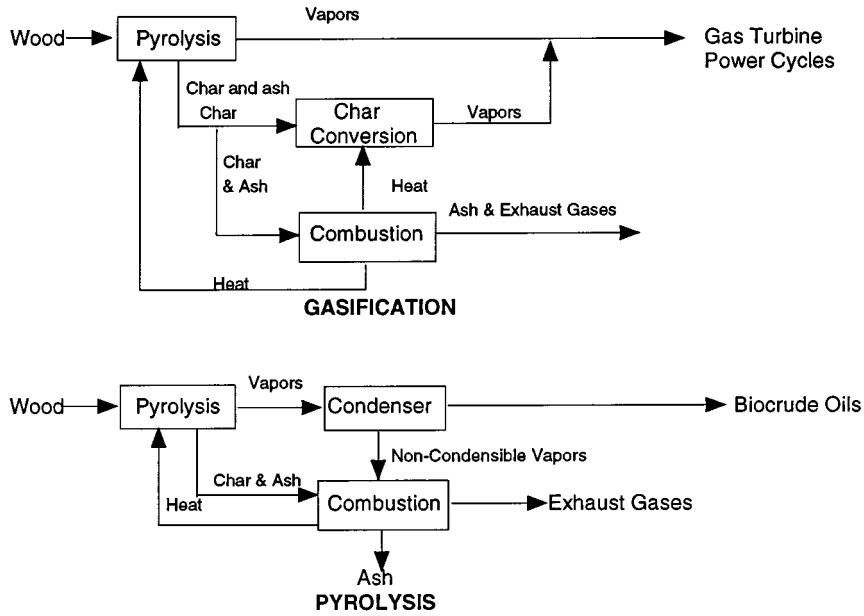


FIGURE 7.4.1. A diagram illustrating biomass gasification and pyrolysis processes. (From U.S. Department of Energy, *Electricity from Biomass: A Development Strategy.* DOE/CH10093-l52, DE920l0590. Washington, D.C., 1992.)

bed into which the biomass is fed. All of the biomass ash is carried overhead with the product gases. By using a deep, single-stage bed of inert solids, the RENUGAS process is able to achieve high carbon conversion with low oils and tars production. The gasifier is capable of producing either an industrial fuel gas or a chemical synthesis gas, depending on air- or oxygen-blown operation.

In an indirectly heated fluidized-bed gasifier, char is removed from the gasifier and burned in a separate vessel. The resulting heat is transferred to the gasifier by either in-bed heat exchangers or by recirculating the inert bed material heated in the char combustor. The advantage of indirect heating of the gasifier is that the gasification product is not diluted with the char combustion by-products (U.S. Department of Energy, 1992). The Battelle Gasification System is an example of an indirect gasification technology (see Figure 7.4.2). The Battelle system is a two-zone circulating-bed gasifier that produces a medium-energy content gas with a heating value of 500 Btu/ft^3 (18.63 MJ/m^3). The product gas can be used in existing natural gas-fired equipment including boilers, kilns, or gas turbines. The Battelle process produces this medium-Btu gas without the need for an oxygen plant, and uses two physically separate reactors (a gasification reactor in which biomass is converted into the product gas and residual char, which is then burned in a separate combustion reactor to provide the heat required for gasification). Heat transfer between the reactors is accomplished by circulating sand between the gasifier reactor and the combustor reactor. The separation of the gasifier from the combustor provides a means to maintain a constant heating value of the product gas regardless of changes in the feed moisture or ash content.

Biomass fuels can provide significant environmental advantages over competing fossil fuels, especially with regard to coal products. These advantages include little or no sulfur content and zero net carbon dioxide production. Additionally, chemicals produced during biomass combustion are absorbed in the photosynthesis process of new biomass growth. However, the fuel conversion process (gasification/combustion) generates emissions such as particulates, tar, and alkali that can cause erosion, corrosion, and deposition problems within the components of advanced conversion systems, especially combustion turbines. As such, some form of gas cleanup must be employed to ensure dependable system performance.

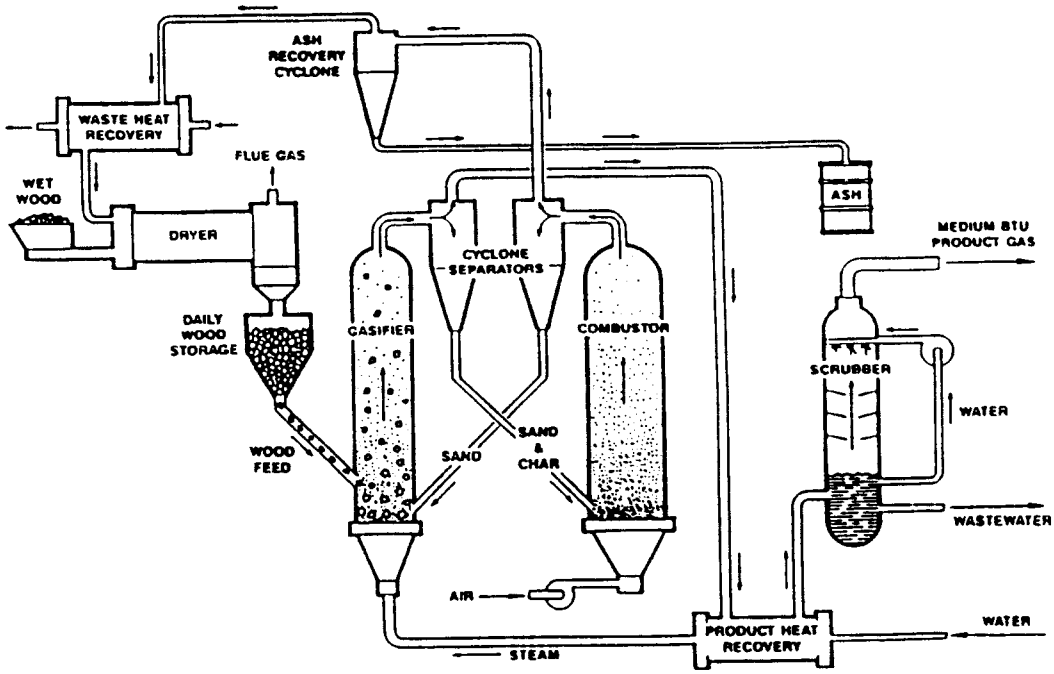


FIGURE 7.4.2. A diagram illustrating the Battelle Biomass Indirect Gasification System.

Biological Gasification or Anaerobic Digestion

The biological gasification process is carried out on feedstock which is either dissolved or slurried in water and typically produces a medium-energy-content gas composed of methane and carbon dioxide. A variety of designs are available, but all of them put anaerobic and methane-forming bacteria in contact with the biomass, the products are the gas (a mixture of methane and carbon dioxide) and a slurry of nonreacted feedstock which can usually be converted into a compost/fertilizer. In the case of landfill gas, the landfill "naturally" produces methane, which is currently being managed for greenhouse gas emissions. The contaminants, which can be highly odoriferous and include hydrogen sulfide, are removed with available technology prior to combustion in either a medium-speed diesel or turbine engine.

Liquefaction to Produce a Biofuel Oil

Biomass feedstocks can also be utilized to create biofuel oils, also known as biocrude. Through a process known as rapid pyrolysis, biofuel oils can be produced and used as fuel for gas turbines, diesel engines, or by co-firing the oil in an existing pulverized coal- or oil-fired boiler. Rapid pyrolysis occurs when heat is transferred to prepared biomass feedstocks (typical thickness <2 mm, moisture <8%) and the solid particles are thermochemically converted to a mixture of noncondensible gases, water vapor, char particles, and pyrolysis oil vapors. The pyrolysis oil vapors typically have a yield of 60 to 80% and are condensed to form a black, viscous, medium-Btu (9000 Btullb or 18.6 MJ/kg) mixture of organic compounds. Because the fuel oils can be stored and transported, the pyrolysis process can be decoupled from the power generation cycle, increasing the flexibility with respect to the proximity of the pyrolysis process to the end user.

Biofuel oil can be produced by pyrolyzing dried wood chips in a **vortex reactor**, followed by removal of solids and condensation of the vapors to recover the higher-molecular-weight hydrocarbons (see Figure 7.4.3). Noncondensible and light hydrocarbons are recycled as carrier gas, or burned with char to generate steam and/or electricity for internal use and export. Table 7.4.2 contains a comparison of biocrude oil characteristics and those of No. 6 fuel oil.

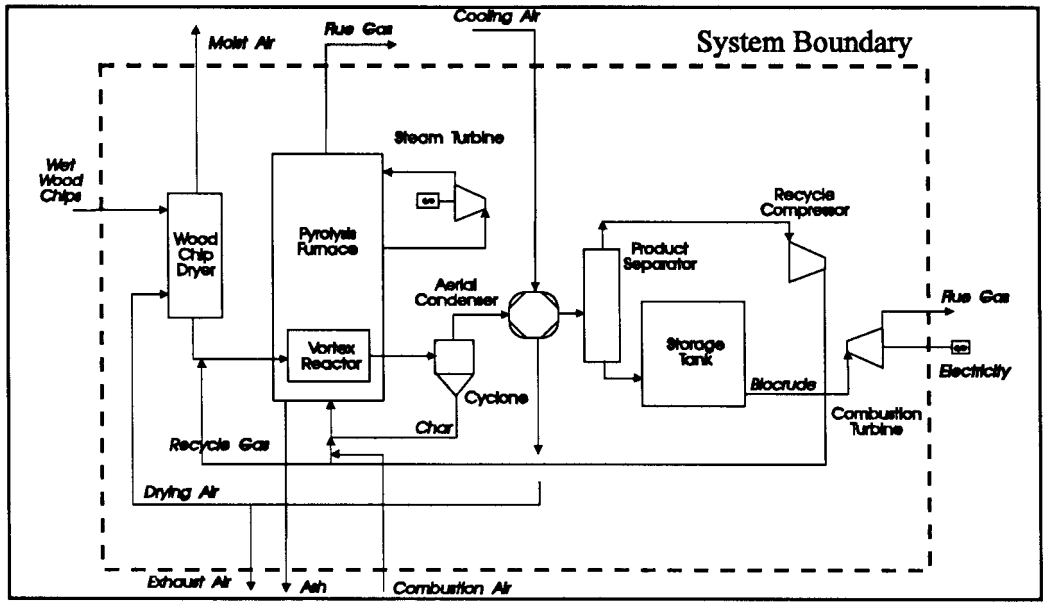


FIGURE 7.4.3. A diagram illustrating the biomass fast pyrolysis process for biocrude oil production.

TABLE 7.4.2 Fuel Oil Characteristics

Type	Biocrude Oil	No. 6 Fuel Oil
Source	Biomass pyrolysis	Petroleum distillation
Sulfur, ppm	5	3,000–30,000
Ash %	0.1–1.0	0.2–1.5
HHV, MJ/kg (BTU/lb)	21–24 (9,000–10,300)	43 (18,400)
HHV, MJ/l (Btu/gal)	25.2 (90,000)	42.1 (150,500)
Density, kg/m^3 @ 16°C (lb/gal @ 60°F)	1,200 (10.14)	950 (8.0)

Source: U.S. Department of Energy, *Biomass Power: A Development Strategy,* DOE/CH10093-152, DE92010590, Washington, D.C., 1992, 31.

Biomass Liquefaction

Ethanol

Ethanol (CH_3CH_2OH) is a renewable liquid fuel that can be produced by fermenting a wide range of biomass materials, including sugars, starches, and lignocellulosics. It is an excellent fuel for spark-ignition internal combustion engines with a high octane number. The automobile industry has developed specifications for E85, an 85% by volume blend with gasoline. Lower-level blends such as El0 are in broad commercial use. Heavy-duty compression ignition engines are usually modified to provide ignition assistance such as glow or spark plugs because of the low cetane and poor autoignition qualities. Although the volumetric energy density of ethanol is two thirds that of gasoline, it has properties, such as the increased octane number, that allow the use of more-efficient high-compression engines. The heat of vaporization is twice that of gasoline, thus increasing the cooled fuel/air mixture that reaches each cylinder. This same property, however, increases cold-starting difficulties.

Currently, corn is used to provide the starch substrate for ethanol production in the United States which is about 1.4 billion gal. Lignocellulosic materials, such as short-rotation woody crops (SRWC) and herbaceous energy crops (HEC), offer an alternative feedstock, one that is dependent on the development of new technology that will allow the cellulose and hemicelluloses that make up

approximately 70% of the lignocellulosic feedstocks to produce ethanol. Technology developments in the last decade in lignocellulosic-to-ethanol conversion have shown major potential for cost reductions that have positioned ethanol from lignocellulosics as a near commercial opportunity. These improvements have been in pretreatment, cellulase enzyme production, xylose fermentation, and cellulose conversion.

A typical flow sheet for ethanol production from lignocellulosics such as SRWC or HEC comprises the following operations: feedstock handling; pretreatment to expose the polymers such as hemicellulose and cellulose; production of the biocatalysts such as **cellulase** and hemicellulase that liberate the sugars glucose and xylose; fermentation to convert sugars into ethanol; ethanol recovery process such as distillation; and waste treatment with combined heat and power production from the lignin and other unconverted process residues. A combination of revolutionary advances and incremental improvements in this process chain has reduced costs of ethanol from $3.60 in 1980 to an estimated $1.27/gal in 1996; this latter is already below the cost of production of ethanol from corn. Further development promises to bring the cost into the range of $0.7 to 0.8/gal. This projected cost is competitive with equivalent fuels produced from oil costing $25/barrel.

Pretreatments to expose the carbohydrate polymers have been systematically optimized and developed to be low cost and to have low environmental impact. Dilute acid **hydrolysis** has been found to be effective at exposing cellulose for further hydrolysis. This pretreatment step is followed by a separation of the soluble hemicellulose sugars from the lignin and cellulose fiber. The cellulose material is used both to produce the cellulase enzymes for the hydrolysis and as the major source of sugar that produces ethanol in a combined saccharification and fermentation step. This combines the release of glucose with its immediate fermentation to ethanol in a single process step that results in a large capital cost saving, decreased fermentation times, and increased enzyme productivity, factors that have made a major reduction in overall cost. Work continues on improving the cellulase productivity, the tolerance of the fermentation yeasts to ethanol, and increases in the solids concentration in the fermenters. The current yield of xylose from the hemicellulose polymer, xylan, is also being improved from 80 to 90% (Zhang et al., 1995).

Methanol from Biomass

Methanol (CH_3OH) can easily be produced by gasifying biomass and then catalytically combining the carbon monoxide and the hydrogen. It is the one-carbon analog of ethanol and shares many of the same fuel properties such as a very high octane number. For spark-ignition engine use, M85, an 85% by volume blend with gasoline, is usually specified. The 15% gasoline provides added fuel volatility, which assists cold starting and adds luminosity to the flame for added safety since neat methanol (M100) burns with a nearly invisible flame.

Methanol is very toxic to humans, but is not a great threat to either soil or water environments, and like ethanol, it is essentially nonthreatening to the environment. In any case, methanol is an important industrial chemical and represents one third of the mass of the MTBE (methyl tertiarybutylether) oxygenate used in auto-fuel oxygenate programs. Currently, it is produced from natural gas using technology that has existed for over 50 years.

Biodiesel

Biodiesel is a diesel fuel that in the United States is derived from oil seed crops such as sunflower, soybean, and rape seed (canola). The oil is esterified in a simple process by reacting it with an alcohol such as methanol or ethanol to produce an ester that has been shown to be an attractive diesel fuel with an excellent cetane number and with much-reduced particulate emissions. The oil is isolated by pressing and solvent extraction from the seed cake, and the meal that remains is a nutritious animal feed. The esterification process produces a glycerine by-product. Both the meal and the glycerine have to be marketed in parallel with the oil production and currently the basic price of the oil seed — a food or feedstuff — is still too high, even though in subsidized markets in Europe production is increasing.

Municipal Solid Waste

Quantities, Characteristics, and Fuel Value

MSW contains energy, typically ranging from about 4000 to 6500 Btu/lb. MSW includes waste discarded from residential, commercial, institutional, and industrial sources. In 1993, over 200 million tons were generated with approximately 16% combusted for energy recovery. The quantities and heating values available will vary depending on the materials collected and processed into fuel. For instance, tires may be defined as MSW, but most will not be collected and processed into refuse-derived fuel (RDF). Yard waste contains some fuel value but may be composted. Of the approximate 207 million tons of MSW generated in the United States in 1993, paper and paperboard contributed about 77.8 million tons; plastics, 19.3 million tons; wood, 13.7 million tons. Paper has a typical heating value of 7200 Btu/lb; cardboard, 7000 Btu/lb; plastics, 14,000 Btu/lb; and wood, 8500 Btu/lb. Other components, such as leather, textiles, yard trimmings, rubber, etc., also have significant heating values. Processed RDF will have higher heating values depending on the degree of processing to remove noncombustibles and the components used to prepare the fuel. The average passenger tire has a heating value of 13,000 to 15,000 BtuAb. However, there are only two dedicated tires-to-energy facilities in the United States combusting tires for energy recovery; a third was scheduled to begin operation some time in 1995. There are a number of cement kilns, pulp and paper mills, electric-generating utility boilers, and industrial boilers that use tire-derived fuel as a portion of their fuel.

Refuse-Derived Fuels

Fuels derived from MSW have been classified according to the degree of processing. They range from waste used as a fuel in its discarded form with only bulky waste removed (RDF-1) to the combustible fraction of the waste processed into gaseous fuel (RDF-7). Some RDF may be densified (d-RDF) into pellets, cubes, or briquets for use in industrial and utility boilers as a supplement to coal.

MSW-to-Energy Conversion Technologies

Although there are several potential methods to recover energy from MSW (e.g., conversion to ethanol, anaerobic digestion, gasification, pyrolysis, etc.), combustion is the most-advanced and the most-applied conversion technology.

Combustion. Direct combustion is the most prevalent technology used to reclaim the energy value in MSW. In 1993, there were 125 municipal waste-to-energy facilities in the United States with a rated MSW capacity of approximately 2500 MW of energy. The most common options are mass burn and RDF facilities, with mass burn being the most prevalent.

In a typical mass burn facility, waste is received into a pit where an overhead crane mixes the waste and removes oversize materials. The waste is fed onto a grate of the furnace, which agitates and moves the waste across the combustion chamber. Air for combustion is introduced from under the grate (underfire air) and from nozzles located above the grate (overfire air). The formation and emission of pollutants (e.g., CO, NO_x, SO_x, hydrocarbons, dioxins, etc.) are affected significantly by the design and operation of the combustor. Energy is recovered from the hot flue gases in a waterwall boiler and is recovered as hot water and steam.

The cooled flue gases then pass through air pollution control (APC) equipment which normally includes combinations of scrubbers to remove acid gases (e.g., HCl) electrostatic precipitators (particulate removal), and/or fabric filters (fine particulates).

Anaerobic Digestion. Anaerobic digestion is the biological degradation of biodegradable materials, in the absence of oxygen. Anaerobic microorganisms convert the biodegradable fraction of the waste to carbon dioxide and methane, which can be collected and used as an energy source. Commercial-scale anaerobic digestion of MSW has not been successful in the U.S. There are approximately 24 plants operating outside of the U.S.

Defining Terms

Cellulase: A group of enzymes, found in many fungi and bacterial organisms, that hydrolyze the cellulose molecule into the component glucose molecules.

Cellulose: A complex carbohydrate found in stems, bark, and fibrous parts of plants and in products such as paper made from plant material. The insoluble molecule is a long chain of six carbon, glucose molecules linked together.

Hemicellulose: A complex carbohydrate molecule also found in plant material. This molecule is more soluble than cellulose and made up of five carbon sugars such as xylose.

Hydrolysis: Chemical decomposition involving the formation of water. Enzymes are capable of catalyzing the breakdown of cellulose resulting in the release of a water molecule and glucose molecules.

Lignin: A hard material embedded in the cellulose matrix of vascular plant cell walls. The molecule is a branched chain of organic rings, and gives strength to the fibers. This material is insoluble and must be removed in order to hydrolyze the cellulose.

Lignocellulosics: Substances composed of lignin and cellulose and/or hemicellulose, for example, wood. Lignocellulosic materials are the initial form of biomass feedstock before pretreatment removes or loosens the lignin and liberates the cellulose for further treatment

Photosynthesis: A process by which plants combine water and carbon dioxide in the presence of light and chlorophyll to make carbohydrates for food.

Vortex reactor: A pyrolysis reaction chamber design whereby a carrier gas and a dry feed stream are injected into a cylindrical reaction chamber, entering the reactor in a helical pattern and flowing over the inside surface of the cylinder, where heat is transferred to the biomass.

References

Anonymous. 1994. *Sweden's Largest Biofuel-Fired CFB Up and Running*, Tampere, Finland, 16–17.

Cralle, H.T. and Vietor, D.M. 1989. Productivity: solar energy and biomass, in *Biomass Handbook*, O. Kitani and C.W. Hal, Eds., Gordon and Breach Science Publishers, New York, 11–20.

Day, D.L. 1989. Biomass waste: agricultural waste, crop residues, in *Biomass Handbook*, O. Kitani and C.W. Hal, Eds., Gordon and Breach Science Publishers, New York, 11–20.

Hall, D.O., Rosillo-Calle, F., Williams, R.H., and Woods, J. 1993. Biomass for energy: supply prospects, in *Renewable Energy: Sources for Fuels and Electricity*, T. B. Johansson, H. Kelly, A.K.N. Reddy and R.H. Williams, Eds., Island Press, Washington, D.C., 593–651.

Jones, C.A. 1985. *C-4 Cereals and Grasses. Growth, Development and Stress Response*, John Wiley & Sons, New York.

Prasad, K. 1985. Stove design for improved dissemination, in *Wood-Stove Dissemination*, Robin Clarke, Ed., Intermediate Technology Publications, London, 59–74.

Sampson, R.N. et al. 1993. Biomass management and energy, *Water, Air, Soil Pollut.*, 70, 139–159.

U.S. Department of Energy. 1992. *Biomass Power: A Development Strategy*, DOE/CH10093-152, DE92010590, Washington, D.C.

Wright, L.L. 1994. Production technology status of woody and herbaceous crops, *Biomass and Bioenerg.*, 6(3), 191–209.

Wyman, C.E. 1994. Ethanol from lignocellulosic biomass: technology, economics, and opportunities, *BioResource Technol.*, 50(1), 3–16.

Zhang, M. et al. 1995. Metabolic engineering of a pentose metabolism pathway in ethanologenic *Zymomonas mobilis*, *Science*, 267, 240.

Further Information

Additional information on energy efficiency and renewable energy technologies can be obtained from the Energy Efficiency and Renewable Energy Clearinghouse (EREC), which can provide publications, tailored technical and business responses, and referrals to energy organizations. Contact EREC at P.O. Box 3048, Merrifield, VA, 22116.

7.5 Nuclear Resources

James S. Tulenko

The Nuclear Fuel Cycle

Sources of Nuclear Fuels and World Reserves

Nuclear power can use two naturally occurring elements, uranium and thorium, as the sources of its fissioning energy. Uranium can be a fissionable source (fuel) as mined (Candu Reactors in Canada), while thorium must be converted in a nuclear reactor into a fissionable fuel. Uranium and thorium are relatively plentiful elements ranking about 60th out of 80 naturally occurring elements. All isotopes of uranium and thorium are radioactive. Today, natural uranium contains, in atomic abundance, 99.2175% Uranium-238 (U^{238}); 0.72% Uranium-235 (U^{235}); and 0.0055% Uranium-234 (U^{234}). Uranium has atomic number 92, meaning all uranium atoms contain 92 protons, with the rest of the mass number being composed of neutrons. Uranium-238 has a half-life of 4.5×10^9 years (4.5 billion years), U-235 has a half-life of 7.1×10^8 years (710 million years), and U-234 has a half-life of 2.5×10^5 years (250 thousand years). Since the age of the earth is estimated at 3 billion years, roughly half of the U-238 present at creation has decayed away, while the U-235 has changed by a factor of sixteen. Thus, when the earth was created, the uranium-235 enrichment was on the order of 8%, enough to sustain a natural reactor of (there is evidence of such an occurrence in Africa). The U-234 originally created has long disappeared, and the U-234 currently present occurs as a product of the decay of U-238.

Uranium was isolated and identified in 1789 by a German scientist, Martin Heinrich Klaproth, who was working with pitchblend ores. No one could identify this new material he isolated, so in honor of the planet Uranus which had just been discovered, he called his new material Uranium. It wasn't until 1896, when the French scientist Henri Becquerel accidentally placed some uranium salts near some paper-wrapped photographic plates, that radioactivity was discovered.

Until 1938, when the German scientists Otto Hahn and Fritz Shassroen succeeded in uranium fission by exposure to neutrons, uranium had no economic significance except in coloring ceramics, where it proved valuable in creating various shades of orange, yellow, brown, and dark green. When a uranium atom is fissioned it releases 200 million electron volts of energy; the burning of a carbon (core) atom releases 4 electron volts. This diffenence of 50 million times in energy release shows the tremendous difference in magnitude between chemical and nuclear energy.

Uranium is present in the earth's crust to the extent of four parts per million. This concentration makes uranium about as plentiful as beryllium, hafnium, and arsenic; and greater in abundance than tungsten, molybdenum, and tantalum. Uranium is an order of magnitude more plentiful than silver and a hundred times more plentiful than gold. It has been estimated that the amount of uranium in the earth's crust to a depth of 12 miles is of the order of 100 trillion tons.

Thorium, which is composed of only one isotope, Thorium-232, has a half-life of 14 billion years (1.4×10^{10} yr), is more than three times more abundant than uranium, and is in the range of lead and gallium in abundance. Thorium was discovered by Berjelius in 1828 and named after Thor, the Scandinavian god of war. For reference, copper is approximately five times more abundant than thorium and twenty times more abundant than uranium.

Uranium is chemically a reactive element; therefore, while it is relatively abundant, it is found chemically combined as an oxide (U_3O_8 or UO_2) and never as a pure metal. Uranium is obtained in three ways, either by underground mining, open pit mining, or in situ leaching. An economic average ore grade is normally viewed as .2% (4 pounds per short ton), though recently ore grades as low as .1% have been exploited. A large quantity of uranium exists in sea-water which has an average concentration of 3×10^{-3} ppm, yielding an estimated uranium quantity available in sea-water of 4000 million tons. A pilot operation was successfully developed by Japan to recover uranium from sea-water, but the cost was about $900/lb, and the effort was shut down as uneconomical.

The major countries with reserves of urnaium in order of importance are Australia, United States, Russia, Canada, South Africa, and Nigeria. The countries with major thorium deposits are India, Brazil, and the United States. It is estimated that for a recovery value of $130/kg ($60/lb), the total uranium reserves in these countries are approximately 1.5 million tonnes of uranium in the U.S., 1 million tonnes of uranium in Australia, .7 million tonnes of uranium in Canada, and 1.3 million tonnes of uranium in the former Soviet Union. As mentioned earlier, thorium reserves are approximately four times greater. With the utilization of breeder reactors, there is enough uranium and thorium to provide electrical power for the next thousand years at current rates of usage.

Processing of Nuclear Fuel

Once the uranium ore is mined it is sent to a concentrator (mill) where it is ground, treated, and purified. Since the ore is of a grade of .1 to .2% uranium, a ton of ore contains only between 1 to 2 kilograms of uranium per 1000 kilograms of ore. Thus, thousands to tonnes of ore have to be extracted and sent to a mill to produce a relatively small quantity of uranium. In the concentration process approximately 95% of the ore is recovered as U_3O_8 (yellowcake) to a purity grade of about 80%. Thus, assuming 0.15% uranium ore, the milling and processing of a metric ton (1000 kg) of ore yields a concentrate of 1.781 kg (1.425 kg of uranium and 0.356 kg of impurities). For this reason the mills must be located relatively close to the mine site. The ore tailings (waste) amounts to 998.219 kg and contains quantities of radon and other urnaium decay products and must be disposed of as a radioactive waste.

The U_3O_8 concentrate is then taken to a conversion plant where the concentrate is further purified (the 20% impurities are removed) and the uranium yellowcake is converted to uranium hexafluoried UF_6). The uranium hexafluoride is a gas at fairly low temperature and is an ideal material for the U-235 isotope enriching processes of either gaseous diffusion or gaseous centrifuge. The UF_6 is shipped in steel cylinders in a solid state, and UF_6 is vaporized by putting the cylinder in a steam bath.

If the uranium is to be enriched to 4% U^{235}, then 1 kilogram of 4% U^{235} product will require 7.4 kilograms of natural uranium feed and will produce 6.4 kilograms of waste uranium (tails or depleted uranium) with a U^{235} isotope content of 0.2%. This material is treated as a radioactive waste. Large quantities of tails (depleted uranium) exist as UF_6 in their original shipping containers at the enriching plants. Depleted uranium (a dense material) has been used as shields for radioactive sources, armor piercing shells, balancing of helicopter rotor tips, yacht hold ballast, and balancing of passenger aircraft.

The enriched UF_6 is then sent to a fabrication plant where it is converted to a uranium dioxide (UO_2) powder. The powder is pressed and sintered into cylindrical pellets which are placed in zircaloy tubes (an alloy of zirconium), pressurized with helium, and sealed. The rods are collected in an array (~17 × 17) bound together by spacer grids, with top and bottom end fittings connected by tie rods or guide tubes. Pressurized water reactor fuel assemblies, each containing approximately 500 kilograms of uranium, are placed in a reactor for 3 to 4 years. A single fuel assembly produces 160,000,000 kilowatt hours of electricity and gives 8,000 people their yearly electric needs for its three years of operation. When the fuel assembly is removed from the reactor it must be placed in a storage pond to allow for removal of the decay heat. After approximately five years of wet storage, the fuel assembly can be removed to dry storage in concrete or steel containers. In the United States the current plan is to permanently store the nuclear fuel, with the Department of Energy assuming responsiblility for the "spent" fuel. The money for the government to handle the storage comes from a fee of 1 mill per kilowatt hour paid by consumers of nuclear-generated electricity. A mill is a thousandth of a dollar or a tenth of a penny. Thus, the fuel assembly described above would have collected $160,000 in the waste fund for the Department of Energy to permanently store the fuel. In Europe, when the fuel is taken out of wet storage it is sent to a reprocessing plant where the metal components are collected for waste disposal; and the fuel is chemically recovered as 96% uranium, which is converted to uranium dioxide for recycling to the enrichment plant, 1% plutonium, which is converted to fuel or placed in storage, and 3% fission products which are encased in glass and permanently stored.

The important thing to remenber about the fuel cycle is the small quantitiy of radioactive fission products (1.5 kilograms) which are created as radioactive waste in producing power which can serve the yearly eletricity needs of 8,000 people for the three years that it operates. The schematic of the entire fuel cycle showing both the United States system (once-through) and the European (recycle) system is given in Figure 7.5.1.

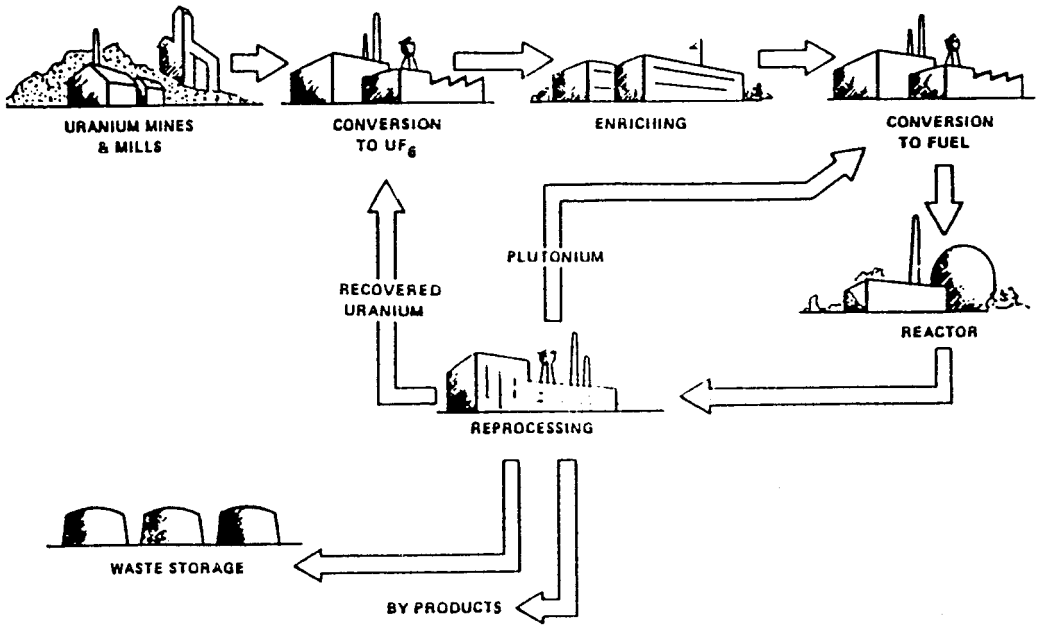


FIGURE 7.5.1 The nuclear fuel cycle.

7.6 Solar Energy Resources

D. Yogi Goswami

The sun is a vast nuclear power plant of the fusion variety which generates power in the form of radiant energy at a rate of 3.8×10^{23} kW. An extremely small fraction of this is intercepted by Earth, but even this small fraction amounts to the huge quantity of 1.8×10^{14} kW. On the average, about 60% of this energy, incident at the outer edge of the atmosphere, reaches the surface. To compare these numbers with our energy needs, consider the present electrical-generating capacity in the United States, which is approximately of 7×10^8 kW. This is equivalent to an average solar radiation falling on only 1000 square miles in a cloudless desert area. It must, however, be remembered that solar energy is distributed over the entire surface of Earth facing the sun, and it seldom exceeds 1.0 kW/m^2. Compared to other sources, such as fossil fuels or nuclear power plants, solar energy has a very low energy density. However, solar radiation can be concentrated to achieve very high energy densities. Indeed, temperatures as high as 3000 K have been achieved in solar furnaces.

Solar energy technology has been developed to a point where it can replace most of the fossil fuels or fossil fuel-derived energy. In many applications it is already economical, and it is a matter of time before it becomes economical for other applications as well.

This section deals in the availability of solar radiation, including methods of measurement, calculation, and available data.

Solar Energy Availability

Detailed information about solar radiation availability at any location is essential for the design and economic evaluation of a solar energy system. Long-term measured data of solar radiation are available for a large number of locations in the United States and other parts of the world. Where long-term measured data are not available, various models based on available climatic data can be used to estimate the solar energy availability. The solar energy is in the form of electromagnetic radiation with the wavelengths ranging from about 0.3 μm (10^{-6} m) to over 3 μm, which correspond to ultraviolet (less than 0.4 μm), visible (0.4 and 0.7 μm), and infrared (over 0.7 μm). Most of this energy is concentrated in the visible and the near-infrared wavelength range (see Figure 7.6.1). The incident solar radiation, sometimes called **insolation**, is measured as irradiance, or the energy per unit time per unit area (or power per unit area). The units most often used are watts per square meter (W/m^2), British thermal units per hour per square foot (Btu/hr-ft^2), and Langleys (calories per square centimeter per minute, cal/cm^2-min).

The amount of solar radiation falling on a surface normal to the rays of the sun outside the atmosphere of the earth (extraterrestrial) at mean Earth-sun distance (D) is called the **solar constant**, I_o. Measurements by NASA indicated the value of solar constant to be 1353 W/m^2 (±1.6%). This value was revised upward and the present accepted value of the solar constant is 1377 W/m^2 (Quinlan, 1979) or 437.1 Btu/hr-ft^2 or 1.974 langleys. The variation in seasonal solar radiation availability at the surface of Earth can be understood from the geometry of the relative movement of Earth around the sun.

Earth-Sun Relationships

Figure 7.6.2 shows the annual motion of Earth around the sun. The **extraterrestrial solar radiation** varies throughout the year because of the variation in the Earth-sun distance (D) as:

$$I = I_o\left(D/D_o\right)^2 \tag{7.6.1}$$

which may be approximated as (Spencer, 1971)

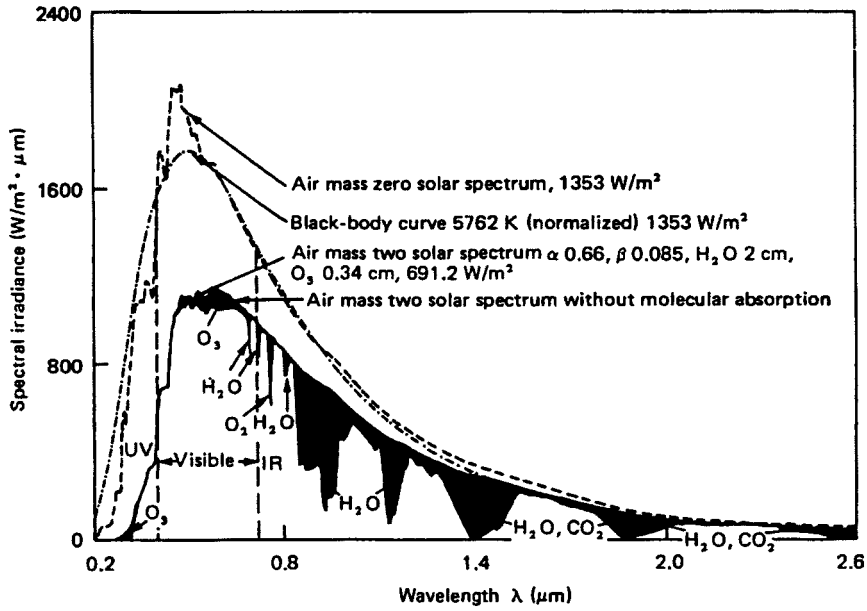


FIGURE 7.6.1 Spectral distribution of solar energy at sea level. (Reprinted by permission from Kreith, F. and Kreider, J.F., *Principles of Solar Engineering*, Hemisphere, Washington, D.C., 1978.)

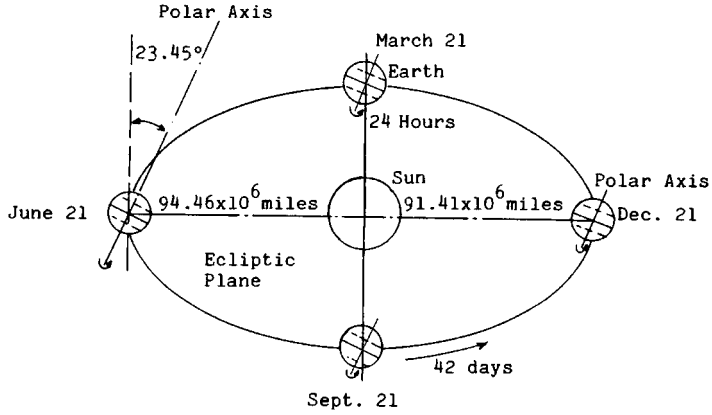


FIGURE 7.6.2 Annual motion of the Earth around the sun.

$$\left(D/D_o\right)^2 = 1.00011 + 0.034221\cos(x) + 0.00128\sin(x) + 0.000719\cos(2x) + 0.000077\sin(2x) \quad (7.6.2)$$

where

$$x = 360(N-1)/365° \quad (7.6.3)$$

and N = Day number (starting from January 1 as 1). The axis of the Earth is tilted at an angle of 23.45° to the plane of its elliptic path around the sun. This tilt is the major cause of the seasonal variation of solar radiation available at any location on Earth. The angle between the Earth-sun line and a plane through the equator is called **solar declination**, δ. The declination varies between –23.45° to +23.45° in 1 year. It may be estimated by the relation:

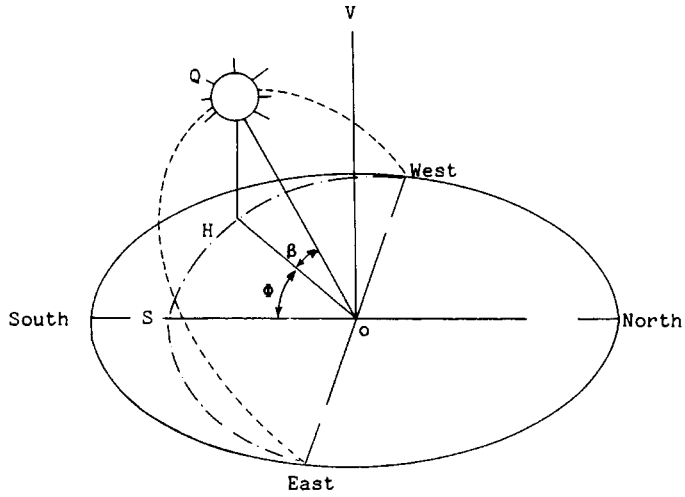


FIGURE 7.6.3 Apparent daily path of the sun across the sky from sunrise to sunset, showing the solar altitude and azimuth angles.

$$\delta = 23.45^\circ \sin\left[360(284 + N)/365^\circ\right] \tag{7.6.4}$$

The apparent motion of the sun around the earth is shown in Figure 7.6.3. The **solar altitude angle**, β, and the **solar azimuth angle**, Φ, describe the position of the sun at any time.

Solar Time

The sun angles are found from the knowledge of solar time, which differs from the local time. The relationship between solar time and local standard time (LST) is given by

$$\text{Solar Time} = \text{LST} + \text{ET} + 4\left(L_{st} - L_{loc}\right) \tag{7.6.5}$$

where ET is the **equation of time**, which is a correction factor in minutes that accounts for the irregularity of the motion of the Earth around the sun. L_{st} is the standard time meridian and L_{loc} is the local longitude. ET can be calculated from the following empirical equation:

$$\text{ET(in minutes)} = 9.87\sin 2B - 7.53\cos B - 1.5\sin B \tag{7.6.6}$$

where $B = 360(N - 81)/365^\circ$.
The sun angles β (altitude) and Φ (azimuth) can be found from the equations:

$$\sin\beta = \cos\ell\cos\delta\cos H + \sin\ell\sin\delta \tag{7.6.7}$$

where ℓ = latitude angle,

$$\sin\Phi = \cos\delta\sin H/\cos\beta \tag{7.6.8}$$

and

$$H = \text{Hour angle} = \frac{\text{Number of minutes from local solar noon}}{4 \text{ min/degree}} \tag{7.6.9}$$

(At solar noon, $H = 0$, so $\beta = 90 - |\ell - \delta|$ and $\Phi = 0$.)

Solar Radiation on a Surface

As solar radiation, I, passes through the atmosphere, some of it is absorbed by air and water vapor, while some gets scattered by molecules of air, water vapor, aerosols, and dust particles. The part of solar radiation that reaches the surface of the Earth with essentially no change in direction is called **direct or beam normal radiation**, I_{bN}. The scattered radiation reaching the surface from the atmosphere is called **diffuse radiation**, I_d.

I_{bN} can be calculated from the extraterrestrial solar irradiance, I, and the atmospheric optical depth τ as (Goswami et al., 1981; ASHRAE, 1995)

$$I_{\text{bN}} = Ie^{-\tau \sec\theta_z} \tag{7.6.10}$$

where θ_z is the solar zenith angle (angle between the sun rays and the vertical). The atmospheric optical depth determines the attenuation of the solar radiation as it passes through the atmosphere. Threlkeld and Jordan (1958) calculated values of τ for average atmospheric conditions at sea level with a moderately dusty atmosphere and amounts of precipitable water vapor equal to the average value for the United States for each month. These values are given in Table 7.6.1. To account for the differences in local conditions from the average sea level conditions Equation (7.6.10) is modified by a parameter called Clearness Number, Cn, introduced by Threlkeld and Jordan (1958):

$$I_{\text{bN}} = \text{Cn}Ie^{-\tau \sec\theta_z} \tag{7.6.11}$$

values of Cn vary between 0.85 and 1.15.

TABLE 7.6.1 Average Values of Atmospheric Optical Depth (τ) and Sky Diffuse Factor (C) for 21st Day of Each Month

Month	1	2	3	4	5	6	7	8	9	10	11	12
τ	0.142	0.144	0.156	0.180	0.196	0.205	0.207	0.201	0.177	0.160	0.149	0.142
C	0.058	0.060	0.071	0.097	0.121	0.134	0.136	0.122	0.092	0.073	0.063	0.057

Source: Threlkeld, J.L. and Jordan, R.C., *ASHRAE Trans.*, 64, 45, 1958.

Solar Radiation on a Horizontal Surface

Total incident solar radiation on a horizontal surface is given by

$$I_{t,\text{Horizontal}} = I_{\text{bN}} \cos\theta_z + CI_{\text{bN}} \tag{7.6.12}$$

$$= I_{\text{bN}} \sin\beta + CI_{\text{bN}} \tag{7.6.13}$$

where θ_z is called the solar zenith angle and C is called the sky diffuse factor, as given in Table 7.6.1.

Solar Radiation on a Tilted Surface

For a surface of any orientation and tilt as shown in Figure 7.6.4, the angle of incidence, θ, of the direct solar radiation is given by

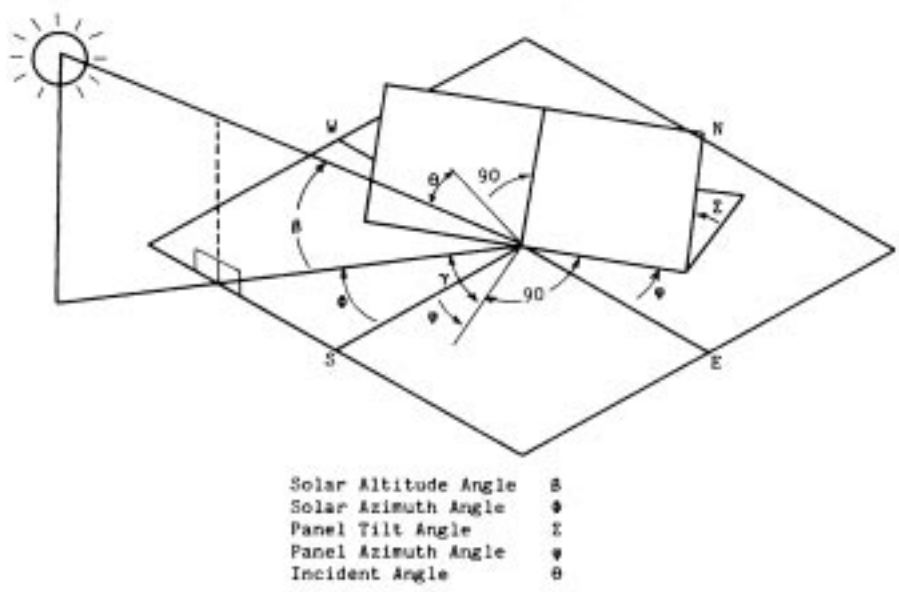


FIGURE 7.6.4 Definitions of solar angles for a tilted surface.

$$\cos\theta = \cos\beta\cos\gamma\sin\Sigma + \sin\beta\cos\Sigma \tag{7.6.14}$$

where γ is the angle between horizontal projections of the rays of the sun and the normal to the surface. Σ is the tilt angle of the surface from the horizontal.
For a tilted surface with angle of incidence θ, the total incident solar radiation is given by

$$I_b = I_{bN}\cos\theta + I_{diffuse} + I_{reflected} \tag{7.6.15}$$

where

$$I_{diffuse} = CI_{bN}(1+\cos\Sigma)/2 \tag{7.6.16}$$

and

$$I_{reflected} = \rho I_{bN}(C+\sin\beta)(1-\cos\Sigma)/2 \tag{7.6.17}$$

where ρ is the reflectivity of the surroundings. For ordinary ground or grass, ρ is approximately 0.2 while for ground covered with snow it is approximately 0.8.

Solar Radiation Measurements

Two basic types of instruments are used in measurements of solar radiation. These are (see Figure 7.6.5):

1. *Pyranometer:* An instrument used to measure global (direct and diffuse) solar radiation on a surface. This instrument can also be used to measure the diffuse radiation by blocking out the direct radiation with a shadow band.
2. *Pyrheliometer:* This instrument is used to measure only the direct solar radiation on a surface normal to the incident beam. It is generally used with a tracking mount to keep it aligned with the sun.

More-detailed discussions about these and other solar radiation measuring instruments can be found in Zerlaut (1989).

FIGURE 7.6.5 Two basic instruments for solar radiation: (a) pyranometer; (b) pyrheliometer.

Solar Radiation Data

Measured values of solar radiation data for locations in the United States are available from the National Climatic Center in Asheville, NC. A number of states have further presented solar radiation data for locations in those states in readily usable form. Weather services and energy offices in almost all the countries have available some form of solar radiation data or climatic data that can be used to derive solar radiation data for locations in those countries. Tables 7.6.2 to 7.6.8 give solar radiation data for clear days for south-facing surfaces in the Northern Hemisphere (and northern-facing surfaces in the Southern Hemisphere) tilted at 0°, 15°, 30°, 45°, 60°, 75°, and vertical, for latitudes 0°, 10°, 20°, 30°, 40°, 50°, and 60°. The actual average solar radiation data at a location is less than the values given in these tables because of the cloudy and partly cloudy days in addition to the clear days. The actual data can be obtained either from long-term measurements or from modeling based on some climatic parameters, such as percent sunshine. Tables 7.6.9 to 7.6.12 give hourly solar angles for northern latitudes 0°, 20°, 40°, and 60°.

TABLE 7.6.2 Average Daily Total Solar Radiation on South-Facing Surfaces in Northern Hemisphere; Latitude = 0°N

Month	Horiz.	15°	30°	45°	60°	75°	90°
1	31.11	34.13	35.13	34.02	30.90	25.96	19.55
2	32.34	33.90	33.45	31.03	26.80	21.05	14.18
3	32.75	32.21	29.79	25.67	20.12	13.53	6.77
4	31.69	29.13	24.93	19.39	12.97	6.59	4.97
5	29.97	26.08	20.81	14.64	8.34	4.92	5.14
6	28.82	24.43	18.81	12.54	6.66	5.07	5.21
7	29.22	25.08	19.66	13.48	7.45	5.17	5.31
8	30.59	27.48	22.87	17.13	10.82	5.58	5.32
9	31.96	30.51	27.34	22.65	16.78	10.18	5.33
10	32.18	32.82	31.54	28.44	23.73	17.72	10.84
11	31.33	33.80	34.28	32.72	29.24	24.08	17.58
12	30.51	33.90	35.27	34.53	31.73	27.05	20.83

TABLE 7.6.3 Average Daily Total Solar Radiation on South-Facing Surfaces in Northern Hemisphere; Latitude = 10°N

Month	Horiz.	15°	30°	45°	60°	75°	90°
1	27.19	31.27	33.48	33.67	31.83	28.08	22.69
2	29.64	32.43	33.31	32.20	29.17	24.45	18.35
3	31.84	32.63	31.51	28.56	23.98	18.08	11.27
4	32.71	31.37	28.23	23.55	17.65	11.09	5.63
5	32.48	29.58	25.07	19.35	12.96	6.95	5.50
6	32.01	28.46	23.42	17.36	10.97	5.78	5.68
7	32.13	28.88	24.10	18.23	11.87	6.33	5.75
8	32.31	30.28	26.57	21.48	15.42	9.10	5.58
9	31.93	31.74	29.73	26.05	20.95	14.79	8.17
10	30.25	32.14	32.16	30.29	26.67	21.54	15.25
11	27.85	31.43	33.14	32.87	30.64	26.60	21.02
12	26.30	30.66	33.19	33.71	32.20	28.75	23.59

TABLE 7.6.4 Average Daily Total Solar Radiation on South-Facing Surfaces in Northern Hemisphere; Latitude = 20°N

Month	Horiz.	15°	30°	45°	60°	75°	90°
1	22.47	27.33	30.55	31.91	31.32	28.83	24.59
2	25.96	29.83	31.92	32.10	30.34	26.77	21.63
3	29.83	31.91	32.12	30.43	26.97	21.97	15.78
4	32.65	32.59	30.67	27.02	21.91	15.74	9.16
5	34.01	32.26	28.73	23.73	17.68	11.23	6.09
6	34.31	31.79	27.57	22.05	15.73	9.44	6.05
7	34.11	31.93	28.03	22.76	16.59	10.26	6.09
8	33.00	32.16	29.53	25.30	19.83	13.57	7.51
9	30.80	31.87	31.11	28.58	24.44	18.98	12.61
10	27.28	30.32	31.58	30.97	28.54	24.44	18.97
11	23.50	27.95	30.73	31.67	30.68	27.85	23.36
12	21.34	26.38	29.83	31.47	31.18	28.99	25.03

Note: Values are in megajoules per square meter. Clearness number = 1.0; ground reflection = 0.2.

TABLE 7.6.5 Average Daily Total Solar Radiation on South-Facing Surfaces in Northern Hemisphere; Latitude = 30°N

Month	Horiz.	15°	30°	45°	60°	75°	90°
1	17.19	22.44	26.34	28.63	29.15	27.86	24.85
2	21.47	26.14	29.25	30.59	30.06	27.70	23.68
3	26.81	30.04	31.50	31.09	28.84	24.90	19.54
4	31.48	32.71	32.06	29.57	25.44	19.96	13.60
5	34.49	33.96	31.56	27.49	22.08	15.82	9.49
6	35.61	34.24	31.03	26.28	20.40	13.97	8.02
7	35.07	34.06	31.21	26.76	21.11	14.77	8.68
8	32.60	33.00	31.54	28.35	23.68	17.89	11.57
9	28.60	30.87	31.35	30.02	26.97	22.42	16.67
10	23.41	27.38	29.74	30.33	29.10	26.14	21.66
11	18.50	23.48	27.05	28.98	29.14	27.51	24.20
12	15.90	21.19	25.21	27.68	28.44	27.43	24.71

TABLE 7.6.6 Average Daily Total Solar Radiation on South-Facing Surfaces in Northern Hemisphere; Latitude = 40°N

Month	Horiz.	15°	30°	45°	60°	75°	90°
1	11.62	16.72	20.82	23.63	24.96	24.72	22.93
2	16.36	21.45	25.25	27.51	28.07	26.90	24.07
3	22.86	27.03	29.61	30.41	29.39	26.60	22.25
4	29.26	31.69	32.29	31.04	28.01	23.44	17.67
5	33.92	34.63	33.42	30.43	25.88	20.15	13.77
6	35.91	35.73	33.64	29.84	24.65	18.52	12.10
7	35.02	35.20	33.48	30.03	25.13	19.19	12.82
8	31.15	32.73	32.48	30.43	26.73	21.68	15.68
9	25.41	28.72	30.36	30.23	28.35	24.83	19.93
10	18.78	23.37	26.59	28.21	28.12	26.33	22.96
11	13.09	18.11	22.05	24.64	25.70	25.16	23.06
12	10.29	15.25	19.29	22.14	23.60	23.57	22.06

TABLE 7.6.7 Average Daily Total Solar Radiation on South-Facing Surfaces in Northern Hemisphere; Latitude = 50°N

Month	Horiz.	15°	30°	45°	60°	75°	90°
1	6.16	10.32	13.85	16.51	18.13	18.60	17.88
2	10.89	15.84	19.83	22.59	23.95	23.81	22.17
3	18.13	22.92	26.36	28.20	28.33	26.73	23.51
4	26.04	29.52	31.29	31.26	29.41	25.90	20.96
5	32.36	34.22	34.21	32.37	28.83	23.87	17.91
6	35.29	36.23	35.28	32.55	28.25	22.71	16.42
7	34.01	35.32	34.75	32.39	28.42	23.14	17.01
8	28.70	31.35	32.26	31.37	28.76	24.62	19.27
9	21.33	25.43	28.06	29.04	28.32	25.92	22.03
10	13.59	18.36	22.05	24.40	25.26	24.56	22.36
11	7.62	11.98	15.62	18.30	19.82	20.10	19.11
12	4.95	8.70	11.93	14.41	15.98	16.53	16.02

Note: Values are in megajoules per square meter. Clearness number = 1.0; ground reflection = 0.2.

TABLE 7.6.8 Average Daily Total Solar Radiation on South-Facing Surfaces in Northern Hemisphere; Latitude = 60°N

Month	Horiz.	15°	30°	45°	60°	75°	90°
1	1.60	3.54	5.26	6.65	7.61	8.08	8.03
2	5.49	9.38	12.71	15.25	16.82	17.32	16.72
3	12.82	17.74	21.60	24.16	25.22	24.73	22.71
4	21.96	26.22	28.97	30.05	29.38	27.00	23.09
5	30.00	32.79	33.86	33.17	30.73	26.72	21.45
6	33.99	35.82	35.93	34.29	31.00	26.26	20.46
7	32.26	34.47	34.97	33.71	30.78	26.36	20.80
8	25.37	28.87	30.80	31.02	29.53	26.42	21.94
9	16.49	21.02	24.34	26.22	26.54	25.27	22.51
10	8.15	12.39	15.90	18.45	19.85	20.01	18.92
11	2.70	5.27	7.53	9.31	10.51	11.03	10.84
12	0.82	2.06	3.16	4.07	4.71	5.05	5.07

Note: Values are in megajoules per square meter. Clearness number = 1.0; ground reflection = 0.2.

Defining Terms

Diffuse radiation: Scattered solar radiation coming from the sky.

Direct or beam normal radiation: Part of solar radiation coming from the direction of the sun on a surface normal to the sun's rays.

Equation of time: Correction factor in minutes, to account for the irregularity of the Earth's motion around the sun.

Extraterrestrial solar radiation: Solar radiation outside Earth's atmosphere.

Insolation: Incident solar radiation measured as W/m^2 or $Btu/hr\text{-}ft^2$.

Solar altitude angle: Angle between the solar rays and the horizontal plane.

Solar azimuth angle: Angle between the true south horizontal line and the horizontal projection of the sun's rays.

Solar constant: Extraterrestrial solar radiation at the mean Earth-sun distance.

Solar declination: Angle between the Earth-sun line and a plane through the equator.

References

ASHRAE. 1995. 1995 *HVAC Applications*, ASHRAE, Atlanta, GA.

Goswami, D.Y. 1986. *Alternative Energy in Agriculture*, Vol. 1, CRC Press, Boca Raton, FL.

Goswami, D.Y., Klett, D.E., Stefanakos, E.K., and Goswami, T.K. 1981. Seasonal variation of atmospheric clearness numbers for use in solar radiation modelling, *AIAA J. Energ.*, 5(3) 185.

Kreith, F. and Kreider, J.F. 1978. *Principles of Solar Engineering*, Hemisphere Publishing, Washington, D.C.

Quinlan, F.T., Ed. 1979. *SOLMET Volume 2: Hourly Solar Radiation — Surface Meteorological Observations*, National Oceanic and Atmospheric Administration, Asheville, NC.

Spencer, J.W. 1971. Fourier series representation of the position of the sun, *Search*, 2, 172.

Threlkeld, J.L. and Jordan, R.C. 1958. Direct radiation available on clear days, *ASHRAE Trans.*, 64, 45.

Zerlaut, G. 1989. Solar Radiation Instrumentation, Chapter 5, *Solar Resources*, R.L. Hulstrom, Ed., The MIT Press, Cambridge, MA.

TABLE 7.6.9 Hourly Sun Angles for Latitude = 0° N

Solar time		Jan. 21		Feb. 21		March 21	
AM	PM	Altitude	Azimuth	Altitude	Azimuth	Altitude	Azimuth
12	0	69.9	0.0	79.1	0.0	89.9	0.0
11	1	65.1	35.3	71.5	53.4	75.0	89.5
10	2	54.4	53.8	58.3	69.0	60.0	89.7
9	3	41.6	62.6	44.0	74.8	45.0	89.8
8	4	28.0	67.1	29.4	77.5	30.0	89.8
7	5	14.1	69.2	14.7	78.7	15.0	89.9
6	6	0.0	69.9	0.0	79.1	0.0	89.9

		April 21		May 21		June 21	
		Altitude	Azimuth	Altitude	Azimuth	Altitude	Azimuth
12	0	78.5	180.0	70.0	180.0	66.6	180.0
11	1	71.2	128.3	65.2	144.6	62.4	149.2
10	2	58.1	112.2	54.5	126.0	52.6	130.9
9	3	43.9	106.1	41.6	117.2	40.4	121.5
8	4	29.3	103.3	28.0	112.8	27.3	116.6
7	5	14.7	101.9	14.1	110.6	13.7	114.2
6	6	0.0	101.5	0.0	110.0	0.0	113.4

		July 21		Aug. 21		Sept. 21	
		Altitude	Azimuth	Altitude	Azimuth	Altitude	Azimuth
12	0	69.3	180.0	77.6	180.0	88.9	180.0
11	1	64.7	145.5	70.6	130.4	75.0	94.1
10	2	54.1	127.0	57.7	113.8	60.0	92.1
9	3	41.4	118.1	43.7	107.3	45.0	91.5
8	4	27.9	113.5	29.2	104.3	30.0	91.2
7	5	14.0	111.3	14.6	102.9	15.0	91.1
6	6	0.0	110.7	0.0	102.4	0.0	91.1

		Oct. 21		Nov. 21		Dec. 21	
		Altitude	Azimuth	Altitude	Azimuth	Altitude	Azimuth
12	0	79.6	0.0	70.3	0.0	66.6	0.0
11	1	71.8	54.8	65.4	35.8	62.4	30.8
10	2	58.4	69.9	54.6	54.4	52.6	49.1
9	3	44.1	75.5	41.7	63.1	40.5	58.5
8	4	29.5	78.1	28.1	67.5	27.3	63.4
7	5	14.7	79.3	14.1	69.7	13.7	65.8
6	6	0.0	79.6	0.0	70.3	0.0	66.6

Note: To convert from solar time to local time, apply corrections for longitude and equation of time.

Further Information

Cinquenami, V., Owensby, J.R., and Baldwin, R.G. 1978. Input Data for Solar Systems, Report prepared for the United States Department of Energy. National Climatic Center, Asheville, NC.

Solar Resources, edited by R.H. Hulstrom, MIT Press, Cambridge, MA, 1989.

TABLE 7.6.10 Hourly Sun Angles for Latitude = 20° N

Solar time		Jan. 21		Feb. 21		March 21	
AM	PM	Altitude	Azimuth	Altitude	Azimuth	Altitude	Azimuth
12	0	49.9	0.0	59.1	0.0	69.9	0.0
11	1	47.3	21.0	55.8	26.9	65.1	37.9
10	2	40.3	38.0	47.3	46.3	54.4	59.2
9	3	30.4	50.4	36.0	59.1	41.6	71.0
8	4	18.9	59.3	23.4	67.9	28.0	78.7
7	5	6.4	65.9	10.0	74.4	14.0	84.6
6	6	−6.8	71.0	−3.7	79.8	0.0	89.9
		April 21		**May 21**		**June 21**	
		Altitude	Azimuth	Altitude	Azimuth	Altitude	Azimuth
12	0	81.5	0.0	90.0	0.0	86.6	180.0
11	1	73.3	61.8	75.9	92.5	75.7	106.6
10	2	60.0	78.2	61.8	95.2	62.0	102.5
9	3	46.0	86.0	47.8	98.1	48.2	103.2
8	4	31.9	91.4	33.9	101.2	34.5	105.3
7	5	17.9	96.1	20.2	104.7	21.1	108.3
6	6	3.9	100.9	6.7	108.9	7.8	112.2
		July 21		**Aug. 21**		**Sept. 21**	
		Altitude	Azimuth	Altitude	Azimuth	Altitude	Azimuth
12	0	89.3	0.0	82.4	0.0	71.1	0.0
11	1	75.9	95.3	73.7	64.5	66.1	39.6
10	2	61.9	96.6	60.3	79.9	55.1	60.9
9	3	47.9	99.0	46.3	87.2	42.1	72.4
8	4	34.1	101.9	32.2	92.4	28.4	80.0
7	5	20.4	105.4	18.1	97.0	14.5	85.8
6	6	6.9	109.5	4.2	101.7	0.4	91.0
		Oct. 21		**Nov. 21**		**Dec. 21**	
		Altitude	Azimuth	Altitude	Azimuth	Altitude	Azimuth
12	0	59.6	0.0	50.3	0.0	46.6	0.0
11	1	56.2	27.3	47.7	21.2	44.2	19.3
10	2	47.6	46.9	40.6	38.3	37.6	35.4
9	3	36.3	59.7	30.7	50.7	28.3	47.4
8	4	23.6	68.4	19.1	59.6	17.2	56.3
7	5	10.2	74.9	6.5	66.2	5.0	62.8
6	6	−3.5	80.3	−6.6	71.4	−7.8	67.8

Note: To convert from solar time to local time, apply corrections for longitude and equation of time.

TABLE 7.6.11 Hourly Sun Angles for Latitude = 40° N

Solar time		Jan. 21		Feb. 21		March 21	
AM	PM	Altitude	Azimuth	Altitude	Azimuth	Altitude	Azimuth
12	0	29.9	0.0	39.1	0.0	49.9	0.0
11	1	28.3	16.0	37.2	18.6	47.6	22.6
10	2	23.7	30.9	32.0	35.4	41.4	41.8
9	3	16.7	43.9	24.2	49.6	32.7	57.2
8	4	8.0	55.2	14.8	61.6	22.4	69.5
7	5	−2.0	65.2	4.2	72.0	11.3	80.1
6	6	****	74.3	−7.0	81.6	−0.1	89.9

		April 21		May 21		June 21	
		Altitude	Azimuth	Altitude	Azimuth	Altitude	Azimuth
12	0	61.5	0.0	70.0	0.0	73.4	0.0
11	1	58.6	29.1	66.2	37.1	69.2	41.9
10	2	51.1	51.3	57.5	60.9	59.8	65.8
9	3	41.2	67.1	46.8	76.0	48.8	80.2
8	4	30.3	79.2	35.4	87.2	37.4	90.7
7	5	18.8	89.4	24.0	96.6	25.9	99.7
6	6	7.4	98.9	12.7	105.6	14.8	108.4
5	7	−3.8	108.5	1.9	114.7	4.2	117.3

		July 21		Aug. 21		Sept. 21	
		Altitude	Azimuth	Altitude	Azimuth	Altitude	Azimuth
12	0	70.7	0.0	62.4	0.0	51.1	0.0
11	1	66.8	37.9	59.4	29.8	48.8	23.1
10	2	57.9	61.8	51.8	52.2	42.5	42.7
9	3	47.2	76.8	41.9	68.0	33.6	58.1
8	4	35.8	87.8	30.8	80.0	23.3	70.5
7	5	24.3	97.2	19.4	90.1	12.1	81.1
6	6	13.1	106.1	8.0	99.6	0.7	90.8
5	7	2.4	115.2	−3.2	109.1	****	100.6

		Oct. 21		Nov. 21		Dec. 21	
		Altitude	Azimuth	Altitude	Azimuth	Altitude	Azimuth
12	0	39.6	0.0	30.3	0.0	26.6	0.0
11	1	37.8	18.8	28.7	16.1	25.0	15.2
10	2	32.5	35.7	24.1	31.0	20.7	29.4
9	3	24.7	49.9	17.0	44.1	14.0	42.0
8	4	15.1	61.9	8.3	55.5	5.5	53.0
7	5	4.6	72.4	−1.7	65.5	−4.2	62.7

Note: To convert from solar time to local time, apply corrections for longitude and equation of time.

TABLE 7.6.12 Hourly Sun Angles for Latitude = 60° N

Solar time		Jan. 21		Feb. 21		March 21	
AM	PM	Altitude	Azimuth	Altitude	Azimuth	Altitude	Azimuth
12	0	9.9	0.0	19.1	0.0	29.9	0.0
11	1	9.0	14.2	18.1	15.5	28.7	17.2
10	2	6.3	28.2	15.2	30.6	25.5	33.6
9	3	2.0	41.6	10.6	44.9	20.6	49.1
8	4	−3.6	54.6	4.7	58.6	14.4	63.4
7	5	****	67.1	−2.1	71.7	7.3	76.9
		April 21		**May 21**		**June 21**	
		Altitude	Azimuth	Altitude	Azimuth	Altitude	Azimuth
12	0	41.5	0.0	50.0	0.0	53.4	0.0
11	1	40.3	19.4	48.6	21.6	52.0	22.7
10	2	36.7	37.7	44.7	41.4	47.9	43.2
9	3	31.3	54.2	38.9	58.7	42.0	60.8
8	4	24.7	69.1	32.1	73.8	35.0	76.0
7	5	17.5	82.8	24.7	87.5	27.6	89.6
6	6	10.0	95.8	17.2	100.3	20.1	102.2
5	7	2.7	108.7	10.0	112.8	13.0	114.5
		July 21		**Aug. 21**		**Sept. 21**	
		Altitude	Azimuth	Altitude	Azimuth	Altitude	Azimuth
12	0	50.7	0.0	42.4	0.0	31.1	0.0
11	1	49.2	21.8	41.2	19.6	29.9	17.4
10	2	45.3	41.7	37.5	38.0	26.7	34.0
9	3	39.5	59.1	32.1	54.6	21.7	49.5
8	4	32.6	74.2	25.5	69.6	15.4	63.9
7	5	25.3	87.9	18.2	83.3	8.4	77.5
6	6	17.8	100.7	10.7	96.3	0.9	90.5
5	7	10.6	113.1	3.4	109.1	−6.5	103.6
		Oct. 21		**Nov. 21**		**Dec. 21**	
		Altitude	Azimuth	Altitude	Azimuth	Altitude	Azimuth
12	0	19.6	0.0	10.3	0.0	6.6	0.0
11	1	18.6	15.6	9.4	14.3	5.7	13.8
10	2	15.7	30.7	6.6	28.3	3.0	27.3
9	3	11.1	45.1	2.3	41.8	−1.1	40.5
8	4	5.2	58.8	−3.2	54.7	−6.6	53.1
7	5	−1.6	71.9	−9.8	67.3	****	65.5

Note: To convert from solar time to local time, apply corrections for longitude and equation of time.

7.7 Wind Energy Resources*

Dale E. Berg

The mechanical devices that are used to convert kinetic energy in the wind into useful shaft power are known as windmills (the earliest machines were used to mill grain), wind machines, or wind turbines. The earliest use of wind machines appears to have been in ancient Persia, where they were used for grinding grain and pumping water. By the 14th century, completely different types of mills known as post and cap mills had become a major source of energy for milling, water pumping, and other tasks throughout northern Europe, and they remained so well into the 19th century, when the steam engine displaced them in many applications.

A new form of windmill appeared in United States in the second half of the 19th century — the multivane or annular windmill, also sometimes known as the American windmill (see Figure 7.7.1). These small, lightweight machines were designed to survive high winds with no human intervention by automatically shedding power, and they played a large role in the settlement of the American West — an arid country where little surface water is available. Many windmills of this basic type are still in use for water pumping around the world today.

FIGURE 7.7.1 The multiblade American windmill Photograph by Paul Gipe. (Adapted from *Wind Power for Home & Business*, Chelsea Green Publishing. 1993).

By the end of the 19th century, efforts to adapt wind power to electricity generation were underway in several countries. In the early 20th century, small wind turbine generators utilizing only two or three aerodynamic blades and operating at a higher rotational speed than the multibladed windmills were developed. Many thousands of generators of this type have been used to provide electricity in the remote areas of the world over the past 85 years.

Large-scale wind turbines designed to generate electrical power were built and tested in several European countries and the U.S. between 1935 and 1970. However, economic studies showed that the electricity generated by the machines would be every expensive, and no effort was made to develop the machines as a serious alternative energy source.

* This work was supported by the United States Department of Energy under Contract DE-AC04-94AL85000.

As a result of research and development since the mid 1970s, the cost of energy or wind-generated electricity has decreased from around 30¢ per kilowatt-hour (kWhr) in the early 1980s to less than 5¢/kWhr for a modern wind farm at a good site in 1995, and wind turbine availability (the fraction of time the machine is operational; i.e., not disabled for repairs or maintenance) has increased from 50 to 60% to better than 95% over the same period. At the end of 1995, there were over 26,000 wind turbines operating worldwide with an installed capacity of over 5000 MW. Twenty-five percent of that capacity was installed in 1995, and plans for 1996 call for the installation of an additional 25%. About 2500 MW of the 5000 MW was in Europe, 1700 MW was in the U.S., and over 400 MW was in India.

Wind Characteristics

Wind Speed and Shear

The primary cause of atmospheric air motion, or wind, is uneven heating of Earth by solar radiation. For example, land and water along a coastline absorb radiation differently, and this creates the light winds or breezes normally found along a coast. Earth's rotation is also an important factor in creating winds.

Wind moving across Earth's surface is slowed by trees, buildings, grass, rocks, and other obstructions in its path. The effect of these obstructions decreases with increasing height above the surface, typically resulting in a wind speed that varies with height above the Earth's surface — phenomenon known as **wind shear**. For most situations, wind shear is positive and wind speed increases with height, but situations in which the wind shear is negative or inverse are not unusual. In the absence of actual data for a specific site, a commonly used approximation for wind shear is

$$U/U_o = \left(h/h_o\right)^{\alpha} \tag{7.7.1}$$

where U is the velocity at a height h, U_o is the measured velocity at height h_o, and α is the non-dimensional wind shear exponent.

The wind shear, α, varies with terrain characteristics, but usually is between 0.10 and 0.25. Over a smooth, level, grass-covered terrain such as the United States Great Plains, α is normally about 0.14. For wind over row crops or low bushes with a few scattered trees, a value of 0.20 is more common while a value of 0.25 is normally a good value for wind over a heavy stand of trees, several buildings, or hilly or mountainous terrain.

A specific site may display much different wind shear behavior than that given in Equation 7.7.1, and that will dramatically affect site energy capture, making it important to measure the wind resource at the specific site and height where the wind turbine will be located, if at all possible.

Wind Energy Resource

The available power in the wind passing through a given area at any given velocity is due to the kinetic energy of the wind and is given by

$$\text{Power} = \frac{1}{2}\rho A U^3 \tag{7.7.2}$$

where the power is in watts, ρ is the air density in kg/m^3, A is the area of interest perpendicular to the wind in m^2, and U is the wind velocity in m/sec.

Air density decreases with increasing temperature and increasing altitude. The effect of temperature on density is relatively weak and is normally ignored, as these variations tend to average out over the period of a year. The density difference due to altitude, however, is significant and does not average out. For example, the air density at sea level is approximately 14% higher than that at Denver, CO (elevation 1600 m or 5300 ft above sea level), so wind of any velocity at sea level contains 14% more power than wind of the same velocity at Denver.

From Equation 7.7.2, it is obvious that the most important factor in the available power is the velocity of the wind — available wind power is proportional to the cube of the wind speed, so doubling the wind speed increases the available power by a factor of 8. An increase in wind velocity of only 20%, say, from 5 to 6 m/sec (11.2 to 13.4 mph), yields a 73% increase in available power.

The available wind power at a site can vary dramatically with height due to wind shear. For example, for $\alpha = 0.20$, Equations (7.7.1) and (7.7.2), reveal that the available power at a height of 30 m is approximately $\{(30/10)^{0.2}\}^3 = 1.93$ times the available power at 10 m. The amount of energy available from the wind (the **wind energy resource**) is the average amount of power available from the wind over a specified period of time, frequently a year. If the wind speed is 20 m/sec, the available power is very large, but if it only blows at that speed for 10 h/year and the rest of the time the wind speed is 0, the resource for the year is very small. Therefore, the site **wind speed distribution**, or frequency at which each wind speed occurs, is very important. The distribution is often presented in histogram form as the probability that the wind occurs at any particular wind speed between 0 and the maximum wind speed (see Figure 7.7.9). If the actual wind speed distribution is not available, it is frequently approximated with a Rayleigh distribution based on the yearly average wind speed. The Rayleigh distribution is given by

$$f(U) = \frac{\pi}{4} \cdot \frac{U}{\overline{U}} \exp\left[\frac{-\pi U^2}{4\overline{U}^2}\right] \tag{7.7.3}$$

where $f(U)$ is the frequency of occurrence of wind speed U and $\overline{U}$ is the yearly average wind speed.

How big is the wind energy resource? It varies dramatically from site to site. In 1981, scientists at Battelle Pacific Northwest Laboratory (PNL) in the U.S. analyzed national weather data, ship data, and terrain and made an estimate of the world-wide wind energy resource. That estimate is summarized in Figure 7.7.2. Very little data is available over much of the world, but even where excellent data is available, only very crude estimates can be portrayed on a map of this scale. More detailed analysis of available data for a specific country can yield a much better estimate of the wind resource for that country.

For example, the PNL scientists have carefully analyzed and interpreted the available long-term wind data for the U.S. and have summarized their estimate of the wind energy resource in a series of maps in the *Wind Energy Resource Atlas of the United States* (Elliott et al., 1987). A summary for the entire U.S. is reproduced in Figure 7.7.3 which presents the resource in terms of the power class rating. Sites with power class 4 (at least 250 W/m^2 at 10 m height or 500 W/m^2 at 50 m height) and above are considered economic for utility-scale wind power development with available wind technology, while sites with power class 3 are not considered economic today but are likely to become economic with near-term wind technology advances. Sites in power classes 2 or lower are not considered economic for utility-scale wind power development, but may well be economic for remote or hybrid wind power systems.

Much of the land in the United States is not available for wind energy development because it is in environmental exclusion areas such as parks, monuments, wilderness areas, wildlife refuges, and other protected areas, or only a portion of it may be developed because of land-use restrictions for forest, agriculture, range, and urban lands. Schwartz et al. (1992) estimated the wind energy that could be generated (the **wind energy potential**) for the United States land available for wind energy development. For the purposes of this study, they assumed that turbines with a 50-m hub height, arranged in arrays with 10 diameter by 5 diameter spacing, operating at 25% efficiency, and experiencing 25% power losses, were used to capture the available wind. Their results for power class 4 and above are summarized in Figure 7.7.4 and Table 7.7.1. They estimate that the average power that could be produced from this resource is 512,000 MW, significantly above the 1994 U.S. average electricity generation of 350,000 MW. They also estimate that the available windy lands with power class 3 or better resources could yield an average power of 1,241,000 MW, more than three times the current United States electricity generation. However, since wind energy is intermittent, is typically concentrated away from the major

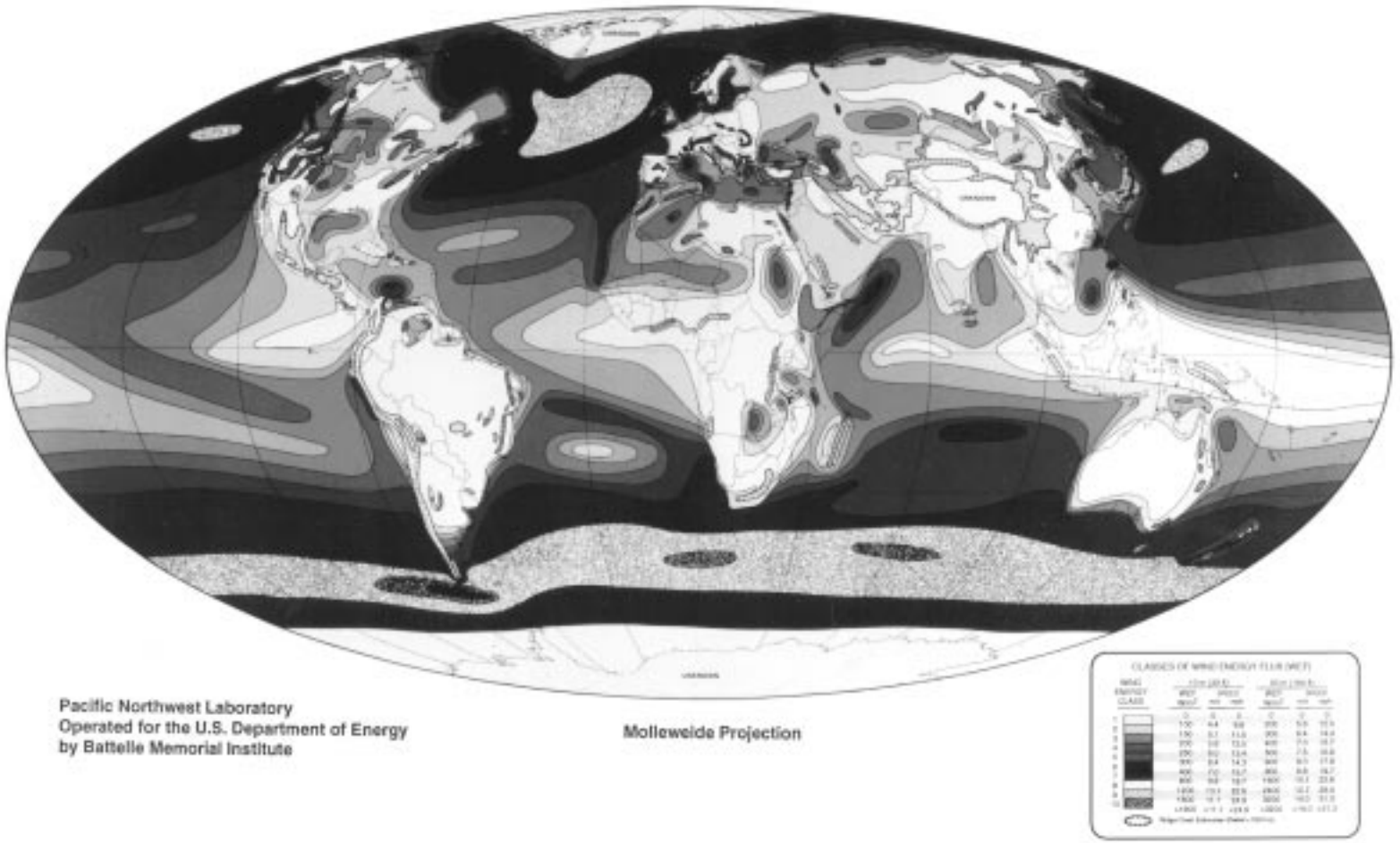


FIGURE 7.7.2 Map of world-wide wind energy resources. (Courtesy of Pacific Northwest Laboratory, Richmond, Washington.)

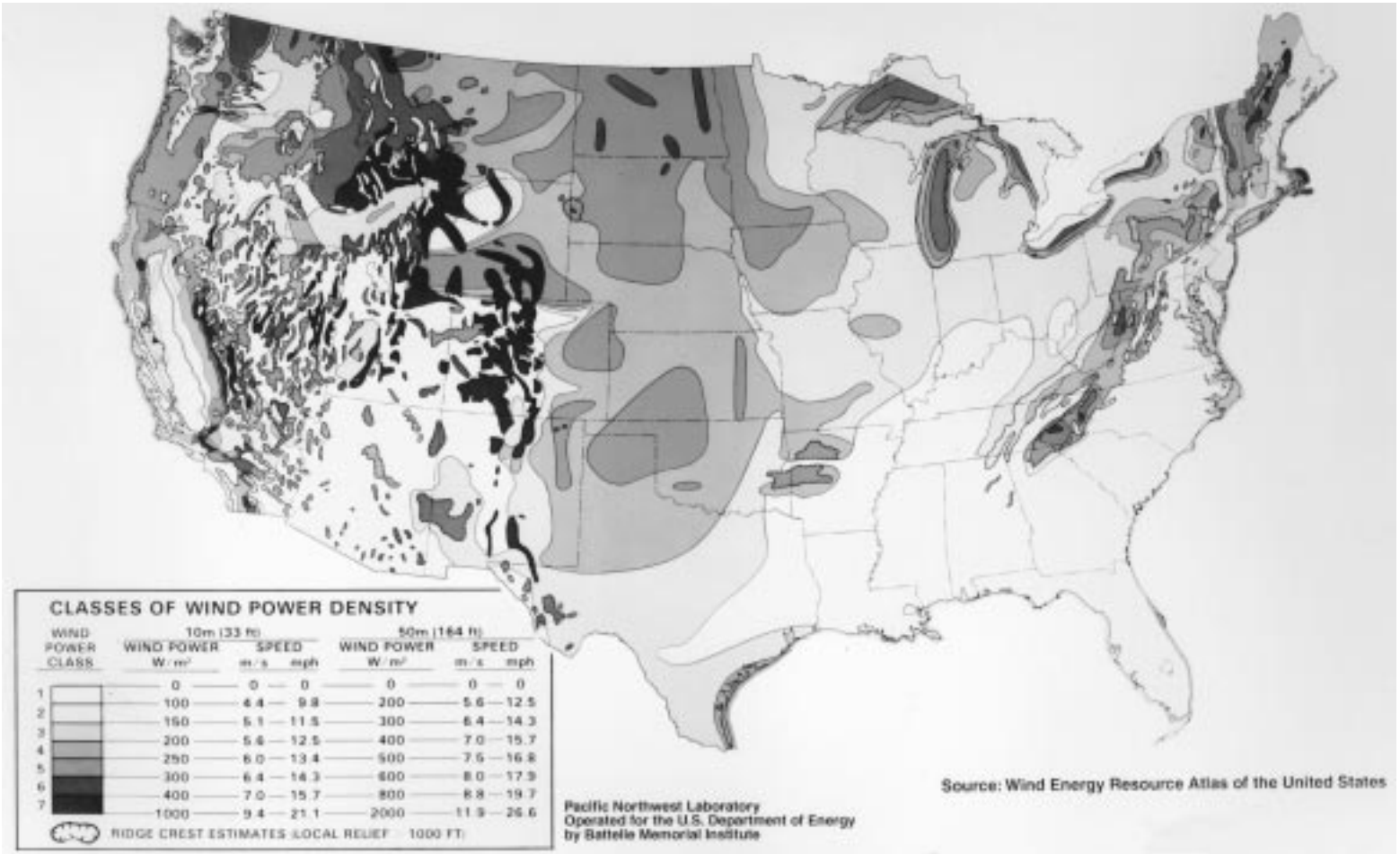


FIGURE 7.7.3 Map of United States wind energy resources. Reproduced from Elliott, Barchet, Foote, and Sandusky, 1987. *Wind Energy Resource Atlas of the United States*. (Courtesy of Pacific Northwest Laboratory, Richmond, Washington.

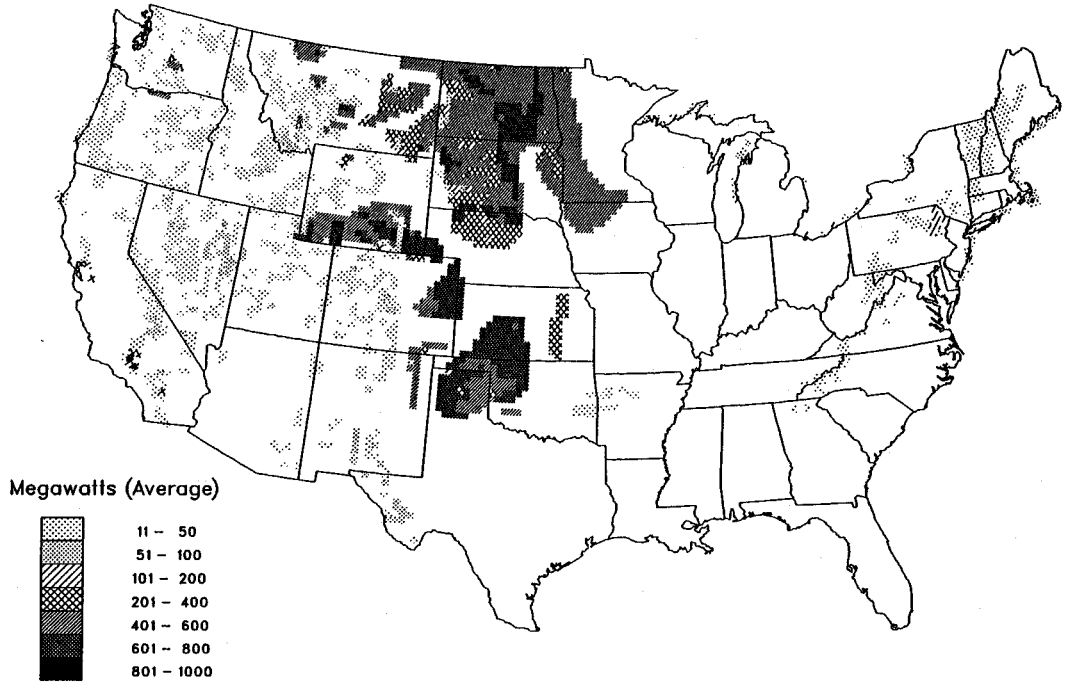


FIGURE 7.7.4 Map of United States wind electric potential. Reproduced from Schwartz, Elliott, and Gower, 1992. *Gridded State Maps of Wind Electric Potential*. (Courtesy of Pacific Northwest Laboratory, Richmond, Washington.)

TABLE 7.7.1 Estimates of Class 4 Wind Land Area and Wind Energy Potential

State	Average Power, MW	Windy Land, km^2
North Dakota	108,000	98,000
South Dakota	70,900	66,500
Wyoming	49,800	39,600
Montana	48,400	42,800
Minnesota	44,800	42,100
Kansas	35,700	33,600
Texas	28,900	27,100
Oklahoma	27,200	25,500
Colorado	26,400	24,300
Nebraska	22,200	20,900
Iowa	15,800	14,900
New Mexico	5,200	4,600
National	512,300	464,200

Note: The top 12 states and total for 48 contiguous states. Wind resource class 4 and above, for a 30-m hub height.

population centers where the electricity demand is highest, and is not always available when the demand is great, it will probably never supply more than 15 to 20% of the U.S. electricity needs.

Similar assessments may be performed for any country of interest. The PNL scientists have recently performed a similar assessment of the available long-term wind data for Mexico and have estimated the wind resource there (Schwartz and Elliott, 1995), as shown in Figure 7.7.5. While wind resources in class 2 are too low to be economic for grid-connected wind power, they are useful for rural power applications. Scientists at Denmark's Risø National Laboratory have done similar work in Europe,

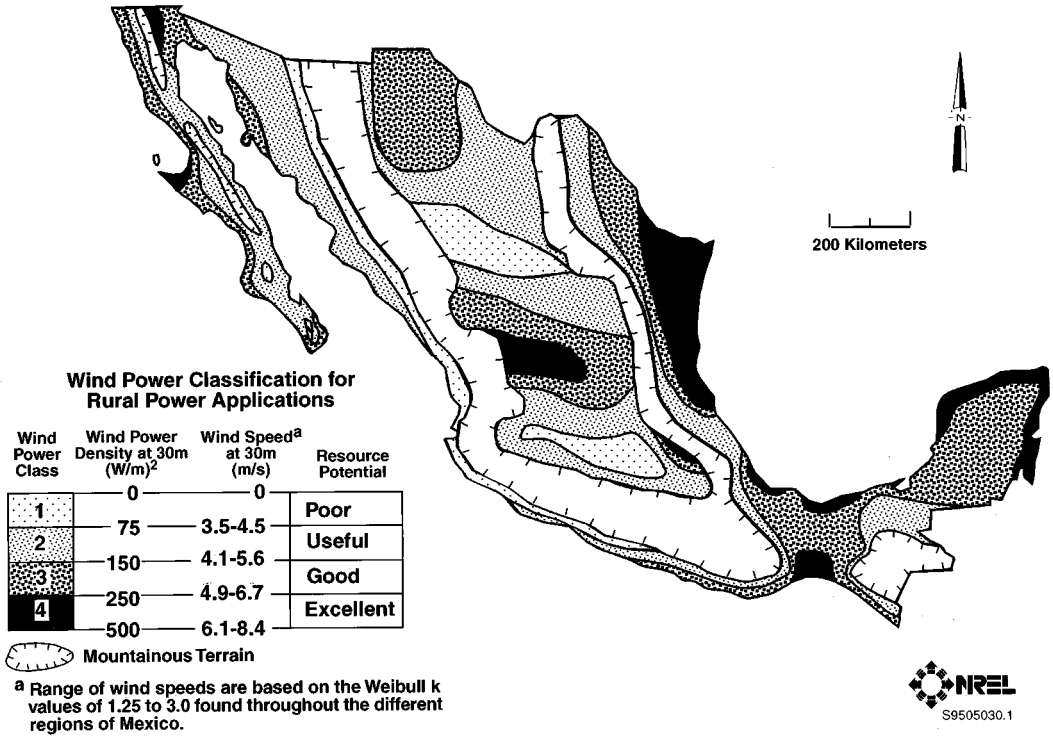


FIGURE 7.7.5 Map of Mexican remote application wind energy resources. Reproduced from Schwartz and Elliott, 1995. *Mexico Wind Resource Assessment Project*. (Courtesy of National Renewable Energy Laboratory, Golden, Colorado.)

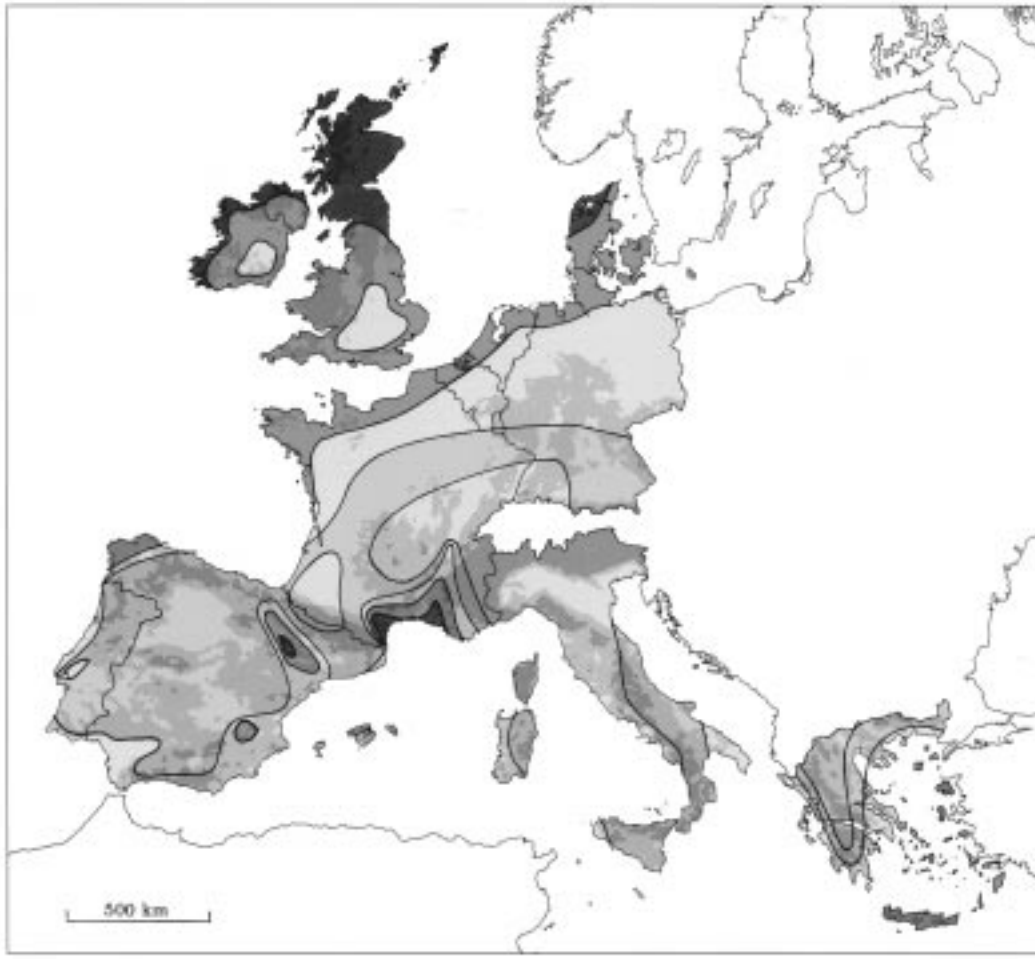
producing a *European Wind Atlas* (Troen and Petersen, 1989) that examines the wind resources of the European Community countries and summarizes the resource available at a 50 m height above smooth, flat terrain, with no obstacles. These results are presented in Figure 7.7.6. Finland, Sweden, Algeria, and Jordan have produced similar reports, and atlases for Poland, Canada, Egypt, Syria, and Turkey are currently in preparation. Resource maps for several other countries may be found in Rohatgi and Nelson (1994).

Existing meteorological station data, in general, significantly underestimate the wind energy resources because of lack of maintenance of wind sensors, poor locations at which they are placed, and cursory methods used in analyzing these data. Assessments such as these only identify the wind energy resource or potential on a very coarse grid. The actual wind resources in any specific area can vary dramatically from those estimates and must be determined with site-specific measurements. Even areas with a very low large-scale resource may contain small areas with a very high resources.

Site Analysis and Selection

Biological Indicators

Many different methods may be used to identify areas of high wind resource. Persistent winds can cause plant deformation, and careful observation of these plants can be used to compare candidate sites and, in at least some cases, to estimate the average wind speed. The Griggs-Putnam index, shown in Figure 7.7.7 and explained in Hewson, Wade, and Baker (1979), is an example of a genus-specific correlation of tree deformation with wind speed. It should be noted that although wind-flagged trees (i.e., trees with branches bent away from a prevailing wind) may indicate that the annual average wind speed is quite strong, unflagged trees do not necessarily indicate that the winds are light. Unflagged trees may be



Wind resources[1] at 50 metres above ground level for five different topographic conditions

	Sheltered terrain[2]		Open plain[3]		At a sea coast[4]		Open sea[5]		Hills and ridges[6]	
	$m\,s^{-1}$	Wm^{-2}	$m\,s^{-1}$	Wm^{-2}	$m\,s^{-1}$	Wm^{-2}	$m\,s^{-1}$	Wm^{-2}	$m\,s^{-1}$	Wm^{-2}
	> 6.0	> 250	> 7.5	> 500	> 8.5	> 700	> 9.0	> 800	> 11.5	> 1800
	5.0-6.0	150-250	6.5-7.5	300-500	7.0-8.5	400-700	8.0-9.0	600-800	10.0-11.5	1200-1800
	4.5-5.0	100-150	5.5-6.5	200-300	6.0-7.0	250-400	7.0-8.0	400-600	8.5-10.0	700-1200
	3.5-4.5	50-100	4.5-5.5	100-200	5.0-6.0	150-250	5.5-7.0	200-400	7.0- 8.5	400- 700
	< 3.5	< 50	< 4.5	< 100	< 5.0	< 150	< 5.5	< 200	< 7.0	< 400

1. The resources refer to the power present in the wind. A wind turbine can utilize between 20 and 30% of the available resource. The resources are calculated for an air density of 1.23 $kg\,m^{-3}$, corresponding to standard sea level pressure and a temperature of 15°C. Air density decreases with height but up to 1000 m a.s.l. the resulting reduction of the power densities is less than 10%.
2. Urban districts, forest and farm land with many windbreaks (roughness class 3).
3. Open landscapes with few windbreaks (roughness class 1). In general, the most favourable inland sites on level land are found here.
4. The classes pertain to a straight coastline, a uniform wind rose and a land surface with few windbreaks (roughness class 1). Resources will be higher, and closer to open sea values, if winds from the sea occur more frequently, i.e. the wind rose is not uniform and/or the land protrudes into the sea. Conversely, resources will generally be smaller, and closer to land values, if winds from land occur more frequently.
5. More than 10 km offshore (roughness class 0).
6. The classes correspond to 50% overspeeding and were calculated for a site on the summit of a single axisymmetric hill with a height of 400 metres and a base diameter of 4 km. The overspeeding depends on the height, length and specific setting of the hill.

FIGURE 7.7.6 Map of European wind energy resources. Reproduced from Troen and Petersen, 1989. *European Wind Atlas.* (Courtesy of Rise National Laboratory, Roskilde, Denmark.

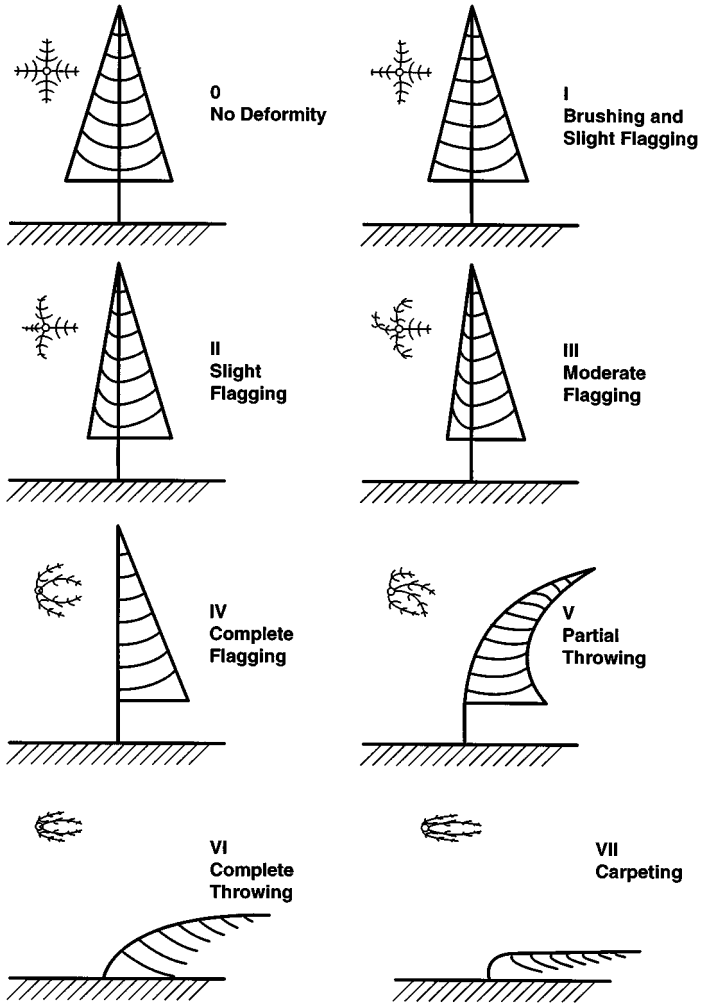


FIGURE 7.7.7 The Griggs-Putnam index of wind-induced tree deformation..

exposed to strong winds from all directions, with insufficient persistence in any one direction to cause flagging.

Effects of Topography

The effects of surrounding terrain on the wind speed at a specific site can be significant. If the proposed site is flat, the terrain texture is the area of most concern. The rougher the terrain, the more the wind flowing over it is retarded near the surface. Flat terrain, such as a large area of flat, open grassland, is the simplest type of terrain for selecting a turbine site. If there are no nearby obstacles (i.e., buildings, trees, or hills), the wind speed at a given height is nearly the same over the entire area, and the best way to increase the available power is to raise the rotor higher above the ground to take advantage of positive wind shear.

Siting turbines on elevated terrain such as hills, ridges, and cliffs takes advantage of the generally higher wind speed at increased altitudes. It may also help keep the turbine exposed to the winds above the calm, cool air that frequently accumulates in valleys and lowlands at night. In some cases, the elevated terrain may cause an acceleration of the wind and thus further improve the resource available to the turbine (see Figure 7.7.8). Depressions include such terrain features as valleys, basins, gorges, and passes. If the depressions channel the wind flow, they may provide good turbine sites. However, the power is

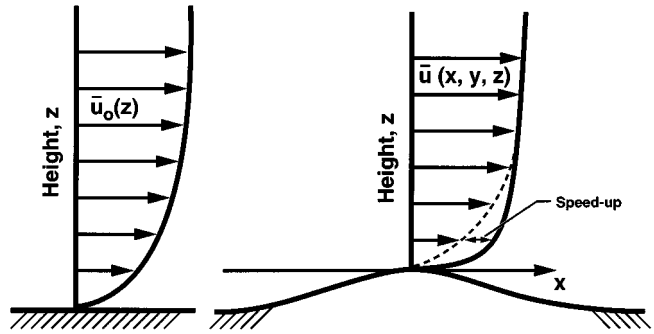


FIGURE 7.7.8 Effect of wind flow over a hill.

more likely to vary by time-of-day or by season than it is for elevated terrain. The impact of terrain on the wind resource and the best places to site turbines for various types of terrain are discussed by Wegley, et al. (1980).

Wind Speed Measurements

The amount of wind energy that can be captured at a site depends on the characteristics of the wind speed at that site. Wind characteristics (speed, direction, distribution, and shear) can vary widely over fairly short distances in either the horizontal or vertical direction. Therefore, in order to predict the energy capture of a turbine with the greatest possible accuracy, wind measurements at the precise site where the turbine will be located are needed. If the rotor, is relatively small, measurements at the hub height (or middle of the rotor) will suffice; for larger rotors, simultaneous measurements at several heights will be required to model the wind speed variation across the rotor disk.

Complete characterization of the wind resource at a specific site is a very time-consuming and expensive effort. A comprehensive site characterization normally requires measuring the wind for at least 12 months, according to Wegley et al. (1980). Even after acquiring the data, questions remain. Long-term data from the nearest airport or long-term weather recording station can help determine whether or not this was a normal wind year for the site and whether the winds were higher or lower than the long-term average (Gipe, 1993). If the costs and time requirements for on-site measurement are excessive, Wegley et al. (1980) and Gipe (1993) give suggestions on methods of using available data from nearby sites, together with a minimum of on-site data to estimate site wind speed.

In the absence of actual site data a Rayleigh distribution (Equation 7.7.3) for the proper yearly average wind speed may be used to estimate the site wind speed distribution. Wind resource estimates that are made with these approximations to the site wind speed and wind distribution will contain significantly larger errors than would be found if the measured site characteristics were used. Figure 7.7.9 illustrates the discrepancies between the measured wind speed distributions at the Amarillo, TX airport and a Bushland, TX test site 20 miles away, and a Rayleigh distribution for the same average wind speed. The measured distribution at the airport has far more energy in the 9 to 12 m/sec range, and this will yield far more energy production than that estimated with the Rayleigh distribution.

Wind Energy Capture Potential

With a wind speed distribution and a turbine power curve (power generated as a function of wind speed) properly adjusted for the local air density, the wind energy potential, or gross annual energy production for a specific site, can be estimated as:

$$\text{Energy} = 8760 \sum_{i=1}^{n} f(U_i) \cdot \Delta U_i \cdot P(U_i) \tag{7.7.4}$$

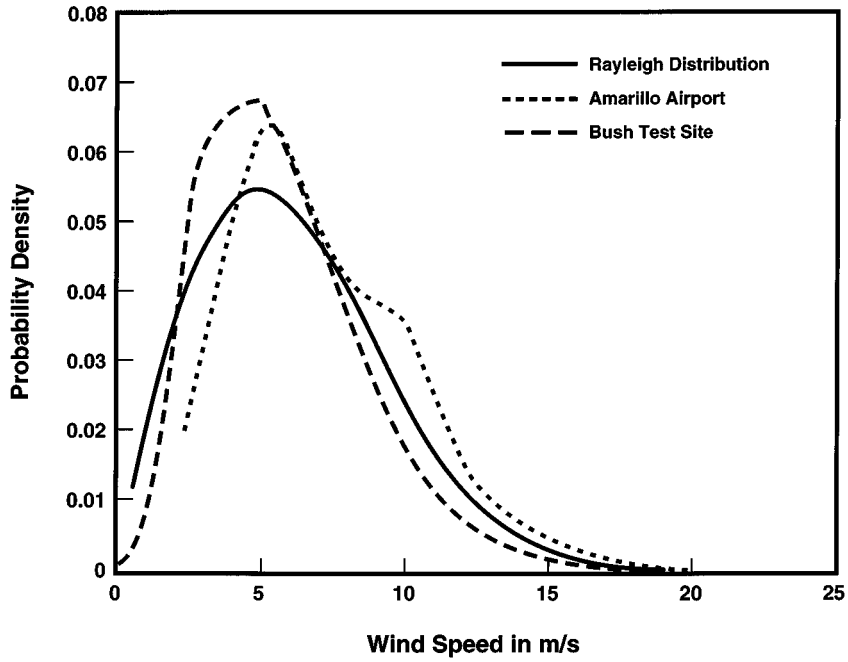


FIGURE 7.7.9 Rayleigh and measured wind frequency distributions.

where 8760 is the number of hours in a year, n is the number of wind speeds considered, $f(U_i) \cdot \Delta U_i$ is the probability of a given wind speed occurring, and $P(U_i)$ is the power produced by the turbine at wind speed U_i. If the measured wind speed distribution is used, replace $f(U_i) \cdot \Delta U_i$ by the probability for the appropriate wind speed (the height of the histogram bar for that wind speed times the wind speed increment). If the power curve is for the turbine aerodynamic or rotor power rather than the electrical power, subtract 10% to approximate the losses due to the turbine bearings and drivetrain. To estimate the net annual energy production, or amount of energy actually delivered to the grid by the turbine, subtract an additional 10% to approximate the electrical losses between the turbine and the grid, the control system losses, and the decreased performance due to dirty blades. If the turbine will be operated in an array, subtract another 5% for array-induced losses.

Environmental Impacts

Although wind turbines generate electricity without causing any air pollution or creating any radioactive wastes, they have an impact on the environment. Wind turbines require a lot of land, but most of that land (90% or so) will remain available for other uses such as farming or livestock grazing. Roads must be built to facilitate installation and maintenance of the wind turbines. Turbines generate noise, but newer turbines are much quieter than earlier models. Current industry standards call for characterization of turbine noise production and rate of decay with distance from the turbine as part of the turbine-testing process. In addition, turbines are large structures that will significantly change the landscape where they are installed, creating visual impact pollution. Another environmental issue facing the wind industry is that of bird deaths as a result of collisions with wind turbines. Preliminary studies indicate that the magnitude of the problem is site specific, and avoiding known bird migration corridors and areas of high bird concentrations when siting turbines and using tubular towers to offer fewer bird perch sites will help to minimize bird collisions with new turbine installations.

Defining Terms

Wind energy potential: The total amount of energy that can actually be extracted from the wind, taking into account the efficiency of the wind turbines and the land available for wind turbine siting.

Wind energy resource: The total amount of energy that is present in the wind.

Wind shear: The change in wind velocity with increasing height above the ground.

Wind speed distribution: The frequency of occurrence of each wind speed over the course of a year. This is site specific.

References

Elliott, D.L., Holladay, C.G., Barchet, W.R., Foote, H.P., and Sandusky, W.F. 1987. *Wind Energy Resource Atlas of the United States*, DOE/CH10094-4, Solar Energy Research Institute, Golden, CO.

Gipe, P. 1993. *Wind Power for Home & Business — Renewable Energy for the 199Os and Beyond*, Chelsea Green Publishing Company, Post Mills, VT.

Hewson, E.W., Wade, J.E., and Baker, R.W. 1979. *A Handbook on the Use of Trees as an Indicator of Wind Power Potential*, RL012227-7913, Oregon State University, Corvallis, OR.

Rohatgi, J.S. and Nelson, V. 1994. *Wind Characteristics — An Analysis for the Generation of Wind Power*, Alternative Energy Institute, West Texas A&M University, Canyon, TX.

Schwartz, M.N. and Elliott, D.L. 1995. Mexico Wind Resource Assessment Project, *Proceedings of Windpower '95*, American Wind Energy Association, Washington, D.C.

Schwartz, M.N., Elliott, D.L., and Gower, G.L. 1992. Gridded state maps of wind electric potential," *Proceedings of Windpower* '92, American Wind Energy Association, Washington, D.C.

Troen, I. and Petersen, E.L., 1989. *European Wind Atlas*, Risø National Laboratory, Roskilde, Denmark.

Wegley, H.L., Rarnsdell, J.V., Orgill, M.M., and Drake, R.L. 1980. *A Siting Handbook for Small Wind Energy Conversion Systems*, PNL2521, Pacific Northwest Laboratory, Richland, WA.

Further Information

Wind Characteristics — An Analysis for the Generation of Wind Power, by J. S. Rohatti and V. Nelson, Alternative Energy Institute, West Texas A&M University, Canyon, TX, 1994, is an excellent source for additional information on the wind resource. *Wind Turbine Technology, Fundamental Concepts of Wind Turbine Engineering*, D. Spera, editor, ASME Press, New York, 1994, contains a wealth of information on wind energy history and technology, together with extensive reference lists. Resources for European-community countries are given in the *European Wind Atlas*, by I. Troen and E. L. Petersen, Risø National Laboratory, Roskilde, Denmark, 1989.

7.8 Geothermal Energy

Joel L. Renner and Marshall J. Reed

The word *Geothermal* comes from the combination of the Greek words *gê*, meaning Earth, and *thérm*, meaning heat. Quite literally, geothermal energy is the heat of the Earth. Geothermal resources are concentrations of the Earth's heat, or geothermal energy, that can be extracted and used economically now or in the reasonable future. Currently, only concentrations of heat associated with water in permeable rocks can be exploited. Heat, fluid, and permeability are the three necessary components of all exploited geothermal fields. This section of Energy Resources will discuss the mechanisms for concentrating heat near the surface, the types of geothermal systems, and the environmental aspects of geothermal production.

Heat Flow

Temperature within the Earth increases with depth at an average of about 25°C/km. Spatial variations of the thermal energy within the deep crust and mantle of the Earth give rise to concentrations of thermal energy near the surface of the Earth that can be used as an energy resource. Heat is transferred from the deeper portions of the Earth by conduction of heat through rocks, by the movement of hot, deep rock toward the surface, and by deep circulation of water. Most high-temperature geothermal resources are associated with concentrations of heat caused by the movement of magma (melted rock) to near-surface positions where the heat is stored.

In older areas of continents, such as much of North America east of the Rocky Mountains, heat flow is generally 40 to 60 mWm^{-2} (milliwatts per square meter). This heat flow coupled with the thermal conductivity of rock in the upper 4 km of the crust yields subsurface temperatures of 90 to 110°C at 4 km depth in the Eastern United States. Heat flow within the Basin and Range (west of the Rocky Mountains) is generally 70 to 90 mWm^{-2}, and temperatures are generally greater than 110°C at 4 km. There are large variations in the Western United States, with areas of heat flow greater than 100 mWm^{-2} and areas which have generally lower heat flow such as the Cascade and Sierra Nevada Mountains and the West Coast. A more detailed discussion of heat flow in the United States is available in Blackwell et al. (1991).

Types of Geothermal Systems

Geothermal resources are hydrothermal systems containing water in pores and fractures. Most hydrothermal resources contain liquid water, but higher temperatures or lower pressures can create conditions where steam and water or only steam are the continuous phases (White et al., 1971; Truesdell and White, 1973). All commercial geothermal production is expected to be restricted to hydrothermal systems for many years because of the cost of artificial addition of water. Successful, sustainable geothermal energy usage depends on reinjection of the maximum quantity of produced fluid to augment natural recharge of hydrothermal systems.

Other geothermal systems that have been investigated for energy production are (1) geopressured-geothermal systems containing water with somewhat elevated temperatures (above normal gradient) and with pressures well above hydrostatic for their depth; (2) magmatic systems, with temperature from 600 to 1400°C; and (3) hot dry rock geothermal systems, with temperatures from 200 to 350°C, that are subsurface zones with low initial permeability and little water. These types of geothermal systems cannot be used for economic production of energy at this time.

Geothermal Energy Potential

The most recent report (Huttrer, 1995) shows that 6800 MW_e (megawatts electric) of geothermal electric generating capacity is on-line in 21 countries (Table 7.8.1). The expected capacity in the year 2000 is

TABLE 7.8.1 Installed and Projected Geothermal Power Generation Capacity

Country	1995	2000
Argentina	0.67	n/a[b]
Australia	0.17	n/a
China	28.78	81
Costa Rica	55	170
El Salvador	105	165
France	4.2	n/a
Greece[a]	0	n/a
Iceland	49.4	n/a
Indonesia	309.75	1080
Italy	631.7	856
Japan	413.705	600
Kenya	45	n/a
Mexico	753	960
New Zealand	286	440
Nicaragua	35	n/a
Philippines	1227	1978
Portugal (Azores)	5	n/a
Russia	11	110
Thailand	0.3	n/a
Turkey	20.6	125
U.S.	2816.7	3395
Totals	6797.975	9960

[a] Greece has closed its 2.0 MWe Milos pilot plant.

[b] n/a = information not available.

Source: Huttrer, G.W., in *Proceedings of the World Geothermal Congress, 1995*, International Geothermal Association, Auckland, N.Z., 1995, 3–14. With permission.

9960 MW_e. Table 7.8.2 lists the electrical capacity of U.S. geothermal fields. Additional details of the U.S. generating capacity are available in DiPippo (1995) and McClarty and Reed (1992). Geothermal resources also provide energy for agricultural uses, heating, industrial uses, and bathing. Freeston (1995) reports that 27 countries had a total of 8228 MW_t (megawatts thermal) of direct use capacity. The total energy used is estimated to be 105,710 TJ/year (terajoules per year). The thermal energy used by the ten countries using the most geothermal resource for direct use is listed in Table 7.8.3.

The U.S. Geological Survey has prepared assessments of the geothermal resources of the U.S. Muffler(1979) estimated that the identified hydrothermal resource, that part of the **identified accessible base** that could be extracted and utilized at some reasonable future time, is 23,000 MW_e for 30 years. This resource would operate power plants with an aggregate capacity of 23,000 MW_e for 30 years. The undiscovered U.S. resource (inferred from knowledge of Earth science) is estimated to be 95,000 to 150,000 MW_e for 30 years.

Geothermal Applications

In 1991, geothermal electrical production in the United States was 15,738 GWh (gigawatt hours), and the largest in the world (McLarty and Reed, 1992).

Most geothermal fields are water dominated, where liquid water at high temperature, but also under high (hydrostatic) pressure, is the pressure-controlling medium filling the fractured and porous rocks of the reservoir. In water-dominated geothermal systems used for electricity, water comes into the wells from the reservoir, and the pressure decreases as the water moves toward the surface, allowing part of the water to boil. Since the wells produce a mixture of flashed steam and water, a separator is installed

TABLE 7.8.2 U.S. Installed Geothermal Electrical Generating Capacity in MW_e

Rated State/Field	Plant Capacity	Type
California		
Casa Diablo	27	B
Coso	240	2F
East Mesa	37	2F
East Mesa	68.4	B
Honey Lake Valley	2.3	B
Salton Sea	440	2F
The Geysers	1797	S
Hawaii		
Puna	25	H
Nevada		
Beowawe	16	2F
Brady Hot Springs	21	2F
Desert Peak	8.7	2F
Dixie Valley	66	2F
Empire	3.6	B
Soda Lake	16.6	B
Steamboat	35.1	B
Steamboat	14.4	1F
Stillwater	13	B
Wabuska	1.2	B
Utah		
Roosevelt	20	1F
Cove Fort	2	B
Cove Fort	9	S

Note: S = natural dry steam, 1F = single flash, 2F = double flash, B = binary, H = hybrid flash and binary.

TABLE 7.8.3 Geothermal Energy for Direct Use by the Ten Largest Users Worldwide

Country	Flow Rate, kg/sec	Installed Power, MWt	Energy Used, TJ/year
China	8,628	1,915	16,981
France	2,889	599	7,350
Georgia	1,363	245	7,685
Hungary	1,714	340	5,861
Iceland	5,794	1,443	21,158
Italy	1,612	307	3,629
Japan	1,670	319	6,942
New Zealand	353	264	6,614
Russia	1,240	210	2,422
U.S.	3,905	1874	13,890
Total	37,050	8,664	112,441

Source: Freeston, D.H., in *Proceedings of the World Geothermal Congress, 1995*, International Geothermal Association, Auckland, N.Z., 1995, 15–26. With permission.

between the wells and the power plant to separate the two phases. The flashed steam goes into the turbine to drive the generator, and the water is injected back into the reservoir.

Many water-dominated reservoirs below 175°C used for electricity are pumped to prevent the water from boiling as it is circulated through heat exchangers to heat a secondary liquid that then drives a turbine to produce electricity. **Binary geothermal plants** have no emissions because the entire amount of produced geothermal water is injected back into the underground reservoir. The identified reserves of lower-temperature geothermal fluids are many times greater than the reserves of high-temperature fluids, providing an economic incentive to develop more-efficient power plants.

Warm water, at temperatures above 20°C, can be used directly for a host of processes requiring thermal energy. Thermal energy for swimming pools, space heating, and domestic hot water are the most widespread uses, but industrial processes and agricultural drying are growing applications of geothermal use. In 1995, the United States was using over 500 TJ/year of energy from geothermal sources for direct use (Lienau, et al., 1995). The cities of Boise, ID; Elko, NV; Klamath Falls, OR; and San Bernardino and Susanville, CA have geothermal district-heating systems where a number of commercial and residential buildings are connected to distribution pipelines circulating water at 54 to 93°C from the production wells (Rafferty, 1992).

The use of geothermal energy through ground-coupled heat pump technology has almost no impact on the environment and has a beneficial effect in reducing the demand for electricity. Geothermal heat pumps use the reservoir of constant temperature, shallow groundwater and moist soil as the heat source during winter heating and as the heat sink during summer cooling. The energy efficiency of geothermal heat pumps is about 30% better than that of air-coupled heat pumps and 50% better than electric-resistance heating. Depending on climate, advanced geothermal heat pump use in the United States reduces energy consumption and, correspondingly, power-plant emissions by 23 to 44% compared to advanced air-coupled heat pumps, and by 63 to 72% compared with electric-resistance heating and standard air conditioners (L'Ecuyer et al., 1993).

Environmental Constraints

Geothermal energy is one of the cleaner forms of energy now available in commercial quantities. Geothermal energy use avoids the problems of acid rain, and it greatly reduces greenhouse gas emissions and other forms of air pollution. Potentially hazardous elements produced in geothermal brines are removed from the fluid and injected back into the producing reservoir. Land use for geothermal wells, pipelines, and power plants is small compared with land use for other extractive energy sources such as oil, gas, coal, and nuclear. Geothermal development projects often coexist with agricultural land uses, including crop production or grazing. The average geothermal plant occupies only 400 m^2 for the production of each gigawatt hour over 30 years (Flavin and Lenssen, 1991). The low life-cycle land use of geothermal energy is many times less than the energy sources based on mining, such as coal and nuclear, which require enormous areas for the ore and processing before fuel reaches the power plant. Low-temperature applications usually are no more intrusive than a normal water well. Geothermal development will serve the growing need for energy sources with low atmospheric emissions and proven environmental safety.

All known geothermal systems contain aqueous carbon dioxide species in solution, and when a steam phase separates from boiling water, CO_2 is the dominant (over 90% by weight) **noncondensible gas.** In most geothermal systems, noncondensible gases make up less than 5% by weight of the steam phase. Thus, for each megawatt-hour of electricity produced in 1991, the average emission of carbon dioxide by plant type in the United States was 990 kg from coal, 839 kg from petroleum, 540 kg from natural gas, and 0.48 kg from geothermal flashed-steam (Colligan, 1993). Hydrogen sulfide can reach moderate concentrations of up to 2% by weight in the separated steam phase from some geothermal fields.

At The Geysers geothermal field in California, either the Stretford process or the incineration and injection process is used in geothermal power plants to keep H_2S emissions below 1 ppb (part per billion). Use of the Stretford process in many of the power plants at The Geysers results in the production and disposal of about 13,600 kg of sulfur per megawatt of electrical generation per year. Figure 7.8.1, shows a typical system used in the Stretford process at The Geysers (Henderson and Dorighi, 1989).

The incineration process burns the gas removed from the steam to convert H_2S to SO_2, the gases are absorbed in water to form SO_3^{-2} and SO_4^{-2} in solution, and iron chelate is used to form $S_2O_3^{-2}$ (Bedell and Hammond, 1987). Figure 7.8.2 shows an incineration abatement system (Bedell and Hammond, 1987). The major product from the incineration process is a soluble thiosulfate which is injected into the reservoir with the condensed water used for the reservoir pressure-maintenance program. Sulfur emissions for each megawatt-hour of electricity produced in 1991, as SO_2 by plant type in the United

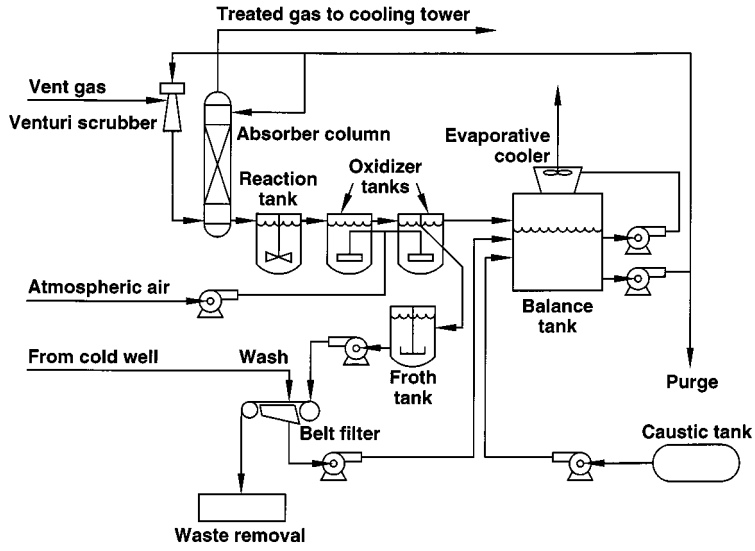


FIGURE 7.8.1 Typical equipment used in the Stretford process for hydrogen sulfide abatement at The Geysers geothermal field. (Based on the diagram of Henderson, J.M. and Dorighi, G.P., *Geotherm. Resour. Counc. Trans.*, 13, 593–595, 1989.)

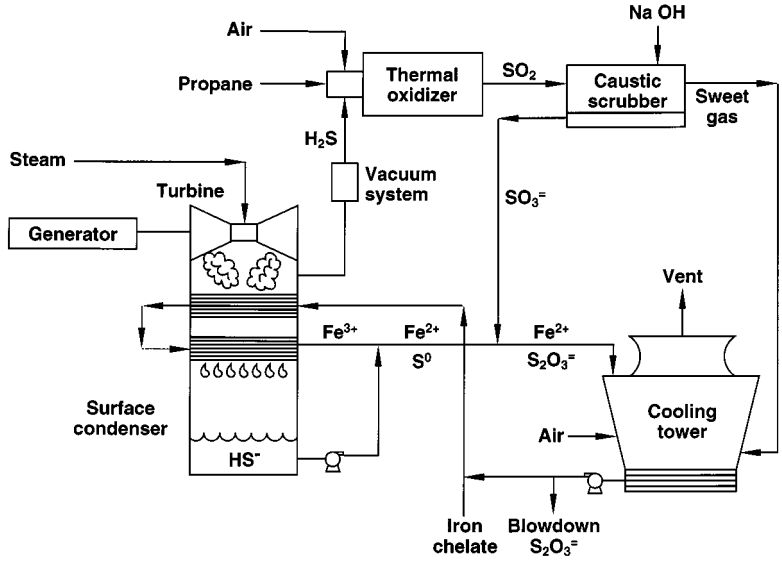


FIGURE 7.8.2 Equipment used in the incineration process for hydrogen sulfide abatement at The Geysers geothermal field. (Based on the diagram of Bedell, S.A. and Hammond, C.A., *Geotherm. Resour. Counc. Bull.*, 16(8), 3–6, 1987.)

States was 9.23 kg from coal, 4.95 kg from petroleum, and 0.03 kg from geothermal flashed-steam (Colligan, 1993). Geothermal power plants have none of the nitrogen oxide emissions that are common from fossil fuel plants.

The waters in geothermal reservoirs range in composition from 0.1 to over 25 wt% dissolved solutes. The geochemistry of several representative geothermal fields is listed in Table 7.8.4. Temperatures up to 380°C have been recorded in geothermal reservoirs in the United States, and many chemical species have a significant solubility at high temperature. For example, all of the geothermal waters are saturated

TABLE 7.8.4 Major Element Chemistry of Representative Geothermal Wells

Field	T(°C)	Na	K	Li	Ca	Mg	Cl	F	Br	SO_4	Total[a] CO_2	Total[a] SiO_2	Total[a] B	Total[a] H_2S
Reykyavik, Iceland	100	95	1.5	<1	0.5	—	31	—	—	16	58	155	0.03	—
Hveragerdi, Iceland	216	212	27	0.3	1.5	0.0	197	1.9	0.45	61	55	480	0.6	7.3
Broadlands, N. Zealand	260	1050	210	1.7	2.2	0.1	1743	7.3	5.7	8	128	805	48.2	<1
Wairekai, New Zealand	250	1250	210	13.2	12	0.04	2210	8.4	5.5	28	17	670	28.8	1
Cerro Prieto, Mexico	340	5820	1570	19	280	8	10420		14.1	0	1653	740	12.4	700
Salton Sea, California	340	50400	17500	215	28000	54	155000	15	120	5	7100	400	390	16
Roosevelt, Utah[b]	<250	2320	461	25.3	8	<2	3860	6.8	—	72	232	563	—	—

[a] Total CO_2, Sio_2 etc. is the total $CO_2 + HCO_2^- + CO_3^{2-}$ expressed as CO_2, silica + silicate as SIO_2, etc.

[b] From Wright (1991); remainder of data from Ellis and Mahon (1977).

in silica with respect to quartz. As the water is produced, silica becomes supersaturated, and, if steam is flashed, the silica becomes highly supersaturated. Upon cooling, amorphous silica precipitates from the supersaturated solution. The high flow rates of steam and water from geothermal wells usually prevent silica from precipitating in the wells, but careful control of fluid conditions and residence time is needed to prevent precipitation in surface equipment. Silica precipitation is delayed in the flow stream until the water reaches a crystallizer or settling pond. There the silica is allowed to settle from the water, and the water is then pumped to an injection well.

Operating Conditions

For electrical generation, typical geothermal wells in the United States have production casing pipe in the reservoir with an inside diameter of 29.5 cm, and flow rates usually range between 150,000 and 350,000 kg/hr of total fluid (Mefferd, 1991). The geothermal fields contain water, or water and steam, in the reservoir, and production rates depend on the amount of boiling in the reservoir and the well on the way to the surface. The Geysers geothermal field in California has only steam filling fractures in the reservoir, and, in 1987 (approximately 30 years after production began), the average well flow had decreased to 33,000 kg/hr of dry steam (Mefferd, 1991) supplying the maximum field output of 2000 MW_2. Continued pressure decline has decreased the production.

In the Coso geothermal field near Ridgecrest, CA initial reservoir conditions formed a steam cap at 400 to 500 m depth, a two-phase (steam and water) zone at intermediate depth, and a liquid water zone at greater depth. Enthalpy of the fluid produced from individual wells ranges from 840 to 2760 kJ/kg (Hirtz et al., 1993), reservoir temperatures range from 200 to 340°C, and the fluid composition flowing from the reservoir into the different wells ranges from 100% liquid to almost 100% steam. Production wells have a wide range of flow rates, but the average production flow rate is 135,000 kg/hr (Mefferd, 1991). Much of the produced fluid is evaporated to the atmosphere in the cooling towers of the power plant, and only about 65% of the produced mass is available for injection into the reservoir at an average rate of 321,000 kg/hr (Mefferd, 1991).

The Salton Sea geothermal system in the Imperial Valley of southern California has presented some of the most difficult problems in brine handling. Water is produced from the reservoir at temperatures between 300 and 350°C and total dissolved solid concentrations between 20 and 25% by weight at an average rate of 270,000 kg/hr (Mefferd, 1991). When up to 20% of the mass of brine boils during production, the salts are concentrated in the brine causing supersaturation with respect to several solid phases. Crystallizers and clarifier and thickener tanks are needed to remove solids from the injection water. Figure 7.8.3 shows the flow stream for removal of solids from the vapor and brine (Signorotti and Hunter, 1992). Other power plants use the addition of acid to lower the pH and keep the solutes in solution (Signorotti and Hunter, 1992). The output from the crystallizers and clarifiers is a slurry of brine and amorphous silica. The methods used to dewater the salt and silica slurry from operations in the Salton Sea geothermal system are described by Benesi (1992). Approximately 80% of the produced water is injected into the reservoir at an average rate of 310,000 kg/hr.

Acknowledgments

This study was supported in part by the U.S. Department of Energy, Assistant Secretary for Energy Efficiency and Renewable Energy, Geothermal Division, under DOE Idaho Operations Office Contract DE-AC07-941D13223.

Defining Terms

Binary geothermal plant: A geothermal electric generating plant that uses the geothermal fluid to heat a secondary fluid that is then expanded through a turbine.

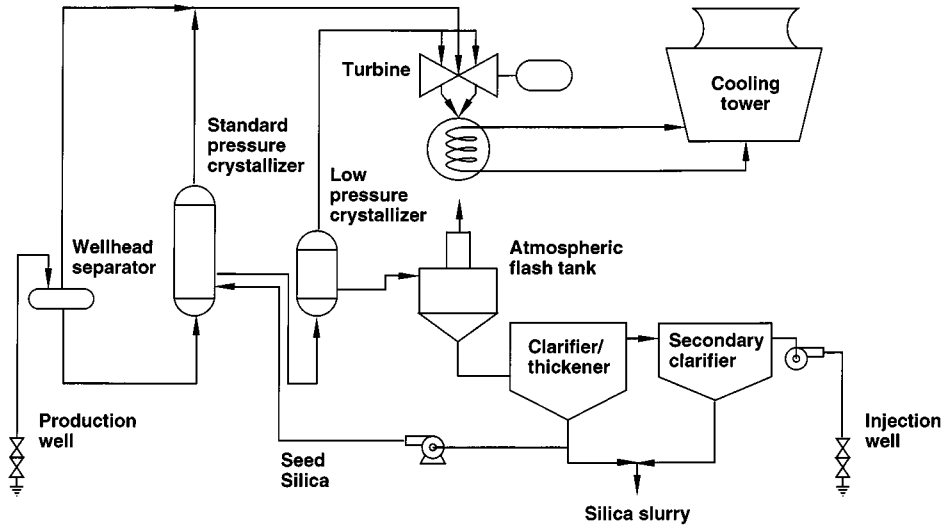


FIGURE 7.8.3 The flow stream for removal of solids from the vapor and brine in typical power plants in the Salton Sea geothermal field. (Modified from the diagram of Signorotti, V. and Hunter, C.C., *Geotherm. Resour. Counc. Bull.*, 21(9), 277–288, 1992).

Identified accessible base: That part of the thermal energy of the Earth that is shallow enough to be reached by production drilling in the foreseeable future. *Identifed* refers to concentrations of heat that have been characterized by drilling or Earth science evidence. Additional discussion of this and other resource terms can be found in Muffler (1979).

Noncondensible gases: Gases exhausted from the turbine into the condenser that do not condense into the liquid phase.

References

Bedell, S.A. and Hammond, C.A. 1987. Chelation chemistry in geothermal H_2S abatement, *Geotherm. Resour. Counc. Bull.*, 16(8), 3–6.

Benesi, S.C. 1992. Dewatering of slurry from geothermal process streams. *Geotherm. Resourc. Counc. Trans.*, 16, 577–581.

Blackwell, D.B., Steele, J.L., and Carter, L.S. 1991. Heat-flow patterns of the North American continent; a discussion of the geothermal map of North America, in *Neotectonics of North America*, D.B. Slemmons, E.R. Engdahl, M.D. Zoback, and D.B. Blackwell, Eds., Geological Society of America, Boulder, CO, 4123–1436.

Colligan, J.G. 1993. U.S. electric utility environmental statistics, in *Electric Power Annual 1991*, U.S Department of Energy, Energy Information Administration, DOE/EIA-0348(91), Washington, D.C.

DiPippo, R. 1995. Geothermal electric power production in the United States: a survey and update for 1990–1994, in *Proceedings of the World Geothermal Congress*, 1995, International Geothermal, Association, Auckland, N.Z., 353–362.

Ellis, A.J. and Mahon, W.A.J. 1977. *Chemistry and Geothermal Systems*, Academic Press, New York.

Flavin, C. and Lenssen, N. 1991. Designing a sustainable energy system, in *State of the World, 1991, A Worldwatch Institute Report on Progress Toward a Sustainable Society*, W.W. Norton and Company, New York.

Freeston, D.H. 1995. Direct uses of geothermal energy 1995 — preliminary review, in *Proceedings of the World Geothermal Congress, 1995*, International Geothermal Association, Auckland, N.Z., 15–26.

Henderson, J.M. and Dorighi, G.P. 1989. Operating experience of converting a Stretford to a Lo-Cat(R) H_2S abatement system at Pacific Gas and Electric Company's Geysers unit 15, *Geotherm. Resour. Counc. Trans.*, 13, 593–595.

Hirtz, P., Lovekin, J., Copp, J., Buck, C., and Adams, M. 1993. Enthalpy and mass flowrate measurements for two-phase geothermal production by tracer dilution techniques, in *Proceedings, 18th Workshop on Geothermal Reservoir Engineering*, Stanford University, Palo Alto, CA, SGPTR-145, 17–27.

Huttrer, G.W. 1995. The status of world geothermal power production 1990–1994, in *Proceedings of the World Geothermal Congress*, 1995, International Geothermal Association, Auckland, N.Z., 3–14.

L'Ecuyer, M., Zoi, C., and Hoffman, J.S. 1993. *Space Conditioning — The Next Frontier*, U.S. Environmental Protection Agency, EPA430-R-93-004, Washington, D.C.

Lienau, P.J., Lund, J.W., and Culver, G.G. 1995. Geothermal direct use in the United States, update 1990–1995, in *Proceedings of the World Geothermal Congress, 1995*, International Geothermal Association, Auckland, N.Z., 363–372.

McLarty, L. and Reed, M.J. 1992. The U.S. geothermal industry: three decades of growth, *Energ. Sources*, 14, 443–455.

Mefferd, M.G. 1991. *76th Annual Report of the State Oil & Gas Supervisor: 1990*, California Division of Oil & Gas, Pub. 06, Sacramento, CA.

Muffler, L.J.P., Ed. 1979. Assessment of geothermal resources of the United States — 1978, U.S. Geological Survey Circular 790, Washington, D.C.

Rafferty, K. 1992. A century of service: the Boise Warm Springs water district system. *Geotherm. Resour. Counc. Bull.*, 21(10), 339–344.

Signorotti, V. and Hunter, C.C. 1992. Imperial Valley's geothermal resource comes of age, *Geotherm. Resour. Counc. Bull.*, 21(9), 277–288.

Truesdell, A.H. and White, D.E. 1973. Production of superheated steam from vapor-dominated reservoirs, *Geothermics*, 2, 145–164.

White, D.E., Muffler, L.T.P., and Truesdell, A.H. 1971. Vapor-dominated hydrothermal systems compared with hot-water systems, *Econ. Geol.*, 66(1), 75–97.

Wright, P.M. 1991. Geochemistry, *Geo-Heat Cent. Bull.*, 13(1), 8–12.

Further Information

Geothermal education materials are available from the Geothermal Education Office, 664 Hilary Drive, Tiburon, CA 94920.

General coverage of geothermal resources can be found in the proceedings of the Geothermal Resources Council's annual technical conference, *Geothermal Resources Council Transactions*, and in the Council's *Geothermal Resources Council Bulletin*, both of which are available from the Geothermal Resources Council, P.O. Box 1350, Davis, CA 95617–1350.

Current information concerning direct use of geothermal resources is available from the Geo-Heat Center, Oregon Institute of Technology, Klamath Falls, OR 97601.

A significant amount of geothermal information is also available on a number of geothermal home pages that can be found by searching on "geothermal" through the Internet.

Goswami, D.Y.; et. al. "Energy Conversion"
Mechanical Engineering Handbook
Ed. Frank Kreith
Boca Raton: CRC Press LLC, 1999

Energy Conversion

D. Yogi Goswami
Univeristy of Florida
Lawrence Conway
Westinghouse Electric Corporation
Steven I. Freedman
Gas Research Institute
David E. Klett
North Carolina A&T State University
Elsayed M. Afify
North Carolina State University
Roger E. A. Arndt
University of Minnesota
William B. Stine
California State Polytechnic University
Anthony F. Armor
Electric Power Research Institute
Chand K. Jotshi
University of Florida
Roberto Pagano (deceased)
University of Florida
James S. Tulenko
University of Florida
Thomas E. Shannon
University of Tennessee
Dale E. Berg
Sandia National Laboratories
Carl J. Bliem (deceased)
CJB Consulting
Gregory L. Mines
Idaho National Engineering Laboratory
Kitt C. Reinhardt
Wright Laboratory, United States Air Force

Mysore L. Ramalingam
UES, Inc.
Jean-Pierre Fleurial
Jet Propulsion Laboratory/California Institute of Technology
William D. Jackson
HMJ Corporation
Desikan Bharatban
National Renewable Energy Laboratory
Frederica Zangrando
National Renewable Energy Laboratory
William W. Bathie
Iowa State University
Howard T. Odum
University of Florida

8.1 Steam Power Plant

Lawrence Conway

Introduction

This section provides an understanding, at an overview level, of the steam power cycle. References were selected for the next level of study if required. There are noteworthy omissions in the section: site selection, fuel handling, civil engineering-related activities (like foundations), controls, and nuclear power.

Thermal power cycles take many forms, but the majority are fossil steam, nuclear, simple cycle gas turbine, and combined cycle. Of those listed, conventional coal-fired steam power is predominant. This is especially true in developing third-world countries that either have indigenous coal or can import coal inexpensively. These countries make up the largest new product market. A typical unit is shown in Figure 8.1.1.

The Rankine cycle is overwhelmingly the preferred cycle in the case of steam power and is discussed first.

Topping and bottoming cycles, with one exception, are rare and mentioned only for completeness. The exception is the combined cycle, where the steam turbine cycle is a bottoming cycle. In the developed countries, there has been a move to the combined cycle because of cheap natural gas or oil. Combined cycles still use a reasonably standard steam power cycle except for the boiler. The complexity of a combined cycle is justified by the high thermal efficiency, which will soon approach 60%.

The core components of a steam power plant are boiler, turbine, condenser and feedwater pump, and generator. These are covered in successive subsections.

The final subsection is an example of the layout/and contents of a modern steam power plant.

As a frame of reference for the reader, the following efficiencies/effectivenesses are typical of modern fossil fuel steam power plants. The specific example chosen had steam conditions of 2400 psia, 1000°F main steam temperature, 1000°F reheat steam temperature: boiler thermal 92; turbine/generator thermal 44; turbine isentropic 89; generator 98.5; boiler feedwater pump and turbine combined isentropic 82; condenser 85; plant overall 34 (Carnot 64).

Nuclear power stations are so singular that they are worthy of a few closing comments. Modern stations are all large, varying from 600 to 1500 MW. The steam is both low temperature and low pressure (~600°F and ~1000 psia), compared with fossil applications, and hovers around saturation conditions

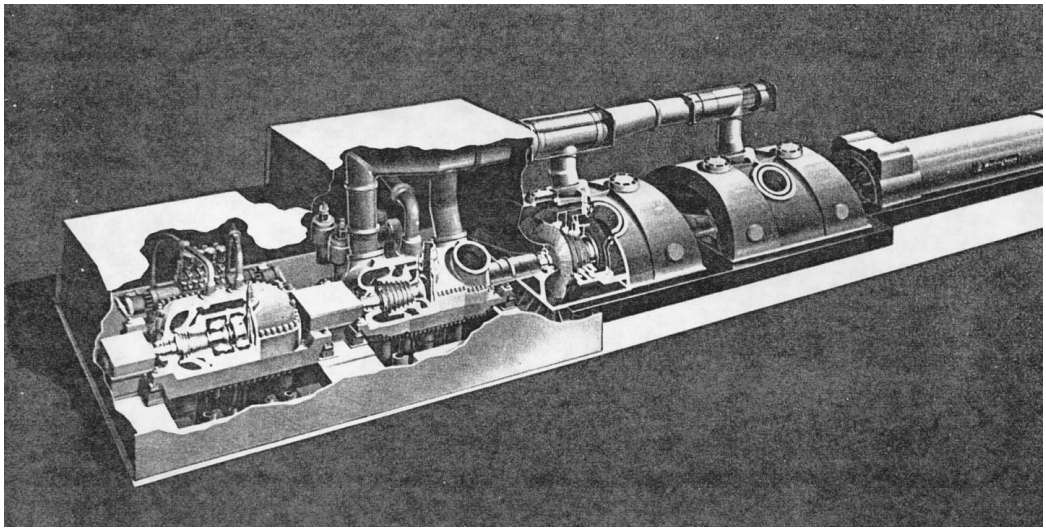


FIGURE 8.1.1 Modern steam turbine generator.

or is slightly superheated. Therefore, the boiler(s), superheater equivalent (actually a combined moisture separator and reheater), and turbines are unique to this cycle. The turbine generator thermal efficiency is around 36%.

Rankine Cycle Analysis

Modern steam power plants are based on the Rankine cycle. The basic, ideal Rankine cycle is shown in Figure 8.1.2. The ideal cycle comprises the processes from state 1:

1–2: Saturated liquid from the condenser at state 1 is pumped isentropically (i.e., $S_1 = S_2$) to state 2 and into the boiler.
2–3: Liquid is heated at constant pressure in the boiler to state 3 (saturated steam).
3–4: Steam expands isentropically (i.e., $S_3 = S_4$) through the turbine to state 4 where it enters the condenser as a wet vapor.
4–1: Constant-pressure transfer of heat in the condenser to return the steam back to state 1 (saturated liquid).

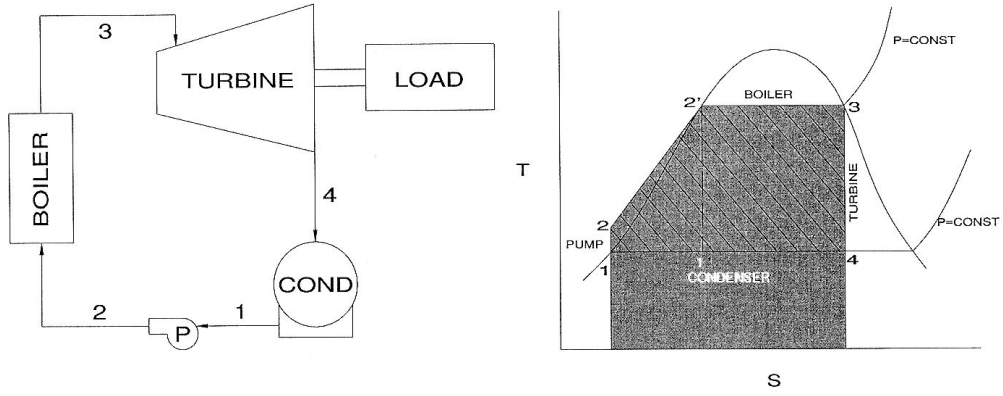


FIGURE 8.1.2 Basic Rankine cycle.

If changes in kinetic and potential energy are neglected, the total heat added to the rankine cycle can be represented by the shaded area on the T-S diagram in Figure 8.1.2, while the work done by this cycle can be represented by the crosshatching within the shaded area. The thermal efficiency of the cycle (η) is defined as the work (W_{NET}) divided by the heat input to the cycle (Q_H).

$$\eta = W_{NET}/Q_H = \left(h_3 - h_4\right)/\left(h_3 - h_2\right)$$

The Rankine cycle is preferred over the Carnot cycle for the following reasons:

The heat transfer process in the boiler has to be at constant temperature for the Carnot cycle, whereas in the Rankine cycle it is superheated at constant pressure. Superheating the steam can be achieved in the Carnot cycle during heat addition, but the pressure has to drop to maintain constant temperature. This means the steam is expanding in the boiler while heat added which is not a practical method.

The Carnot cycle requires that the working fluid be compressed at constant entropy to boiler pressure. This would require taking wet steam from point 1′ in Figure 8.1.2 and compressing it to saturated liquid condition at 2′. A pump required to compress a mixture of liquid and vapor isentropically is difficult to design and operate. In comparison, the Rankine cycle takes the saturated liquid and compresses it to boiler pressure. This is more practical and requires much less work.

The efficiency of the Rankine cycle can be increased by utilizing a number of variations to the basic cycle. One such variation is superheating the steam in the boiler. The additional work done by the cycle is shown in the crosshatched area in Figure 8.1.3.

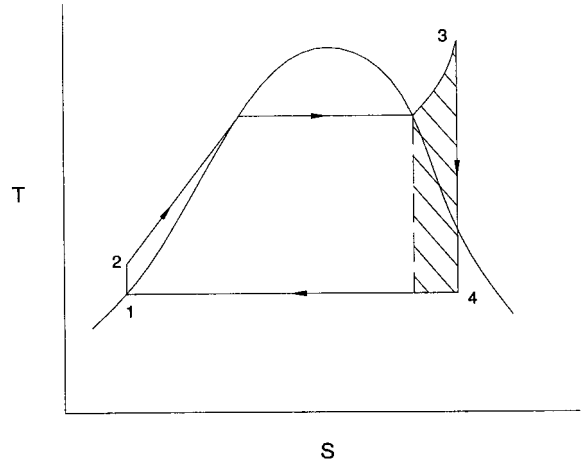


FIGURE 8.1.3 Rankine cycle with superheat.

The efficiency of the Rankine cycle can also be increased by increasing the pressure in the boiler. However, increasing the steam generator pressure at a constant temperature will result in the excess moisture content of the steam exiting the turbine. In order to take advantage of higher steam generator pressures and keep turbine exhaust moistures at safe values, the steam is expanded to some intermediate pressure in the turbine and then reheated in the boiler. Following reheat, the steam is expanded to the cycle exhaust pressure. The reheat cycle is shown in Figure 8.1.4.

Another variation of the Rankine cycle is the regenerative cycle, which involves the use of feedwater heaters. The regenerative cycle regains some of the irreversible heat lost when condensed liquid is pumped directly into the boiler by extracting steam from various points in the turbine and heating the condensed liquid with this steam in feedwater heaters. Figure 8.1.5 shows the Rankine cycle with regeneration.

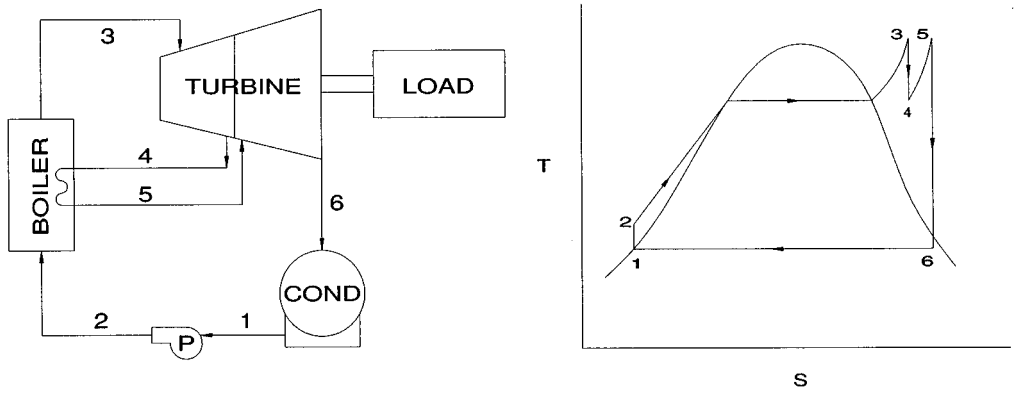


FIGURE 8.1.4 Rankine cycle with reheat.

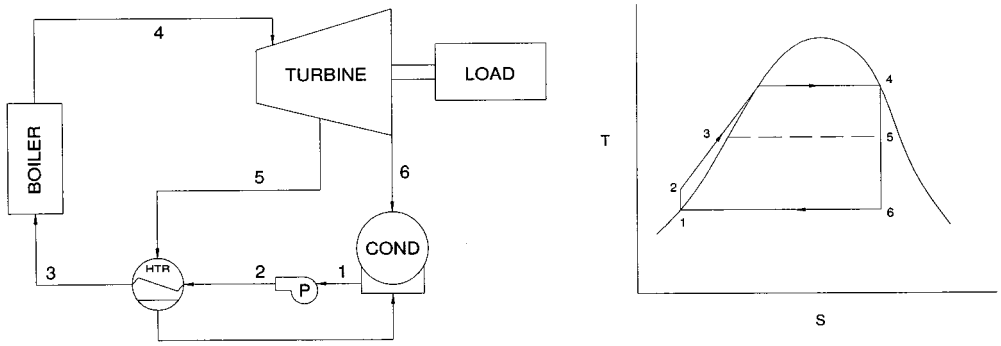


FIGURE 8.1.5 Rankine cycle with regeneration.

The actual Rankine cycle is far from ideal as there are losses associated with the cycle. They include the piping losses due to friction and heat transfer, turbine losses associated with steam flow, pump losses due to friction, and condenser losses when condensate is subcooled. The losses in the compression (pump) and expansion process (turbine) result in an increase in entropy. Also, there is lost energy in heat addition (boiler) and rejection (condenser) processes as they occur over a finite temperature difference.

Most modern power plants employ some variation of the basic Rankine cycle in order to improve thermal efficiency. For larger power plants, economies of scale will dictate the use of one or all of the variations listed above to improve thermal efficiency. Power plants in excess of 200,000 kW will in most cases have 300°F superheated steam leaving the boiler reheat, and seven to eight stages of feedwater heating.

References

Salisbury, J.K. 1950. *Steam Turbines and Their Cycles,* Reprint 1974. Robert K. Krieger Publishing, Malabar, FL.

Van Wylen, G.J. and Sonntag, R.E. 1986. *Fundamentals of Classical Thermodynamics,* 3rd ed., John Wiley & Sons, New York.

Topping and Bottoming Cycles

Steam Rankine cycles can be combined with topping and/or bottoming cycles to form binary thermodynamic cycles. These topping and bottoming cycles use working fluids other than water. Topping cycles change the basic steam Rankine cycle into a binary cycle that better resembles the Carnot cycle and improves efficiency. For conventional steam cycles, state-of-the-art materials allow peak working fluid temperatures higher than the supercritical temperature for water. Much of the energy delivered into the cycle goes into superheating the steam, which is not a constant-temperature process. Therefore, a significant portion of the heat supply to the steam cycle occurs substantially below the peak cycle temperature. Adding a cycle that uses a working fluid with a boiling point higher than water allows more of the heat supply to the thermodynamic cycle to be near the peak cycle temperature, thus improving efficiency. Heat rejected from the topping cycle is channeled into the lower-temperature steam cycle. Thermal energy not converted to work by the binary cycle is rejected to the ambient-temperature reservoir. Metallic substances are the working fluids for topping cycles. For example, mercury was used as the topping cycle fluid in the 40-MW plant at Schiller, New Hampshire. This operated for a period of time but has since been dismantled. Significant research and testing has also been performed over the years toward the eventual goal of using other substances, such as potassium or cesium, as a topping cycle fluid.

Steam power plants in a cold, dry environment cannot take full advantage of the low heat rejection temperature available. The very low pressure to which the steam would be expanded to take advantage of the low heat sink temperature would increase the size of the low-pressure (LP) turbine to such an extent that it is impractical or at least inefficient. A bottoming cycle that uses a working fluid with a vapor pressure higher than water at ambient temperatures (such as ammonia or an organic fluid) would enable smaller LP turbines to function efficiently. Hence, a steam cycle combined with a bottoming cycle may yield better performance and be more cost-effective than a stand-alone Rankine steam cycle.

Further Information

Fraas, A.P. 1982. *Engineering Evaluation of Energy Systems,* McGraw-Hill, New York.

Horlock, J.H. 1992. *Combined Power Plants, Including Combined Cycle Gas Turbine (CCGT) Plants,* Pergamon Press, Oxford.

Steam Boilers

A boiler, also referred to as a steam generator, is a major component in the plant cycle. It is a closed vessel that efficiently uses heat produced from the combustion of fuel to convert water to steam. Efficiency is the most important characteristic of a boiler since it has a direct bearing on electricity production.

Boilers are classified as either drum-type or once-through. Major components of boilers include an economizer, superheaters, reheaters, and spray attemperators.

Drum-Type Boilers

Drum-type boilers (Figure 8.1.6) depend on constant recirculation of water through some of the components of the steam/water circuit to generate steam and keep the components from overheating. Drum-type boilers circulate water by either natural or controlled circulation.

Natural Circulation. Natural circulation boilers use the density differential between water in the downcomers and steam in the waterwall tubes for circulation.

Controlled Circulation. Controlled circulation boilers utilize boiler-water-circulating pumps to circulate water through the steam/water circuit.

Once-Through Boilers

Once-through boilers, shown in Figure 8.1.7, convert water to steam in one pass through the system.

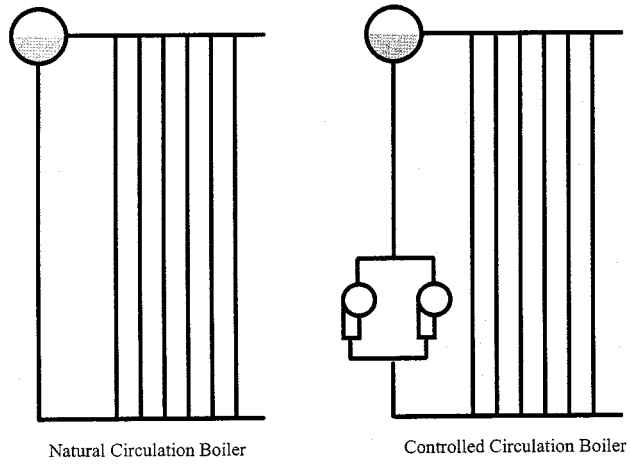


FIGURE 8.1.6 Drum boilers.

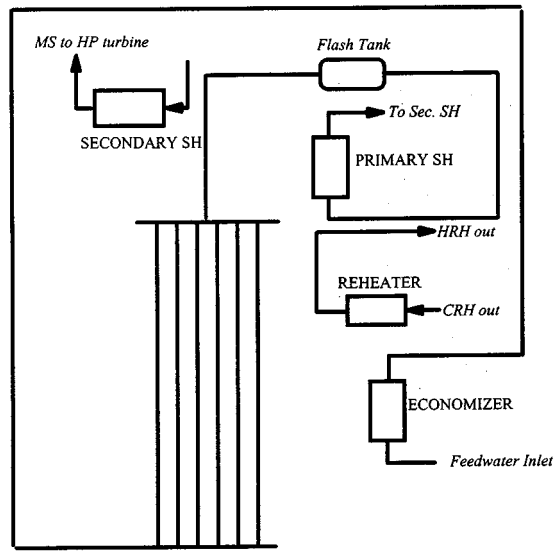


FIGURE 8.1.7 Once-through boilers.

Major Boiler Components

Economizer. The economizer is the section of the boiler tubes where feedwater is first introduced into the boiler and where flue gas is used to raise the temperature of the water.

Steam Drum (Drum Units Only). The steam drum separates steam from the steam/water mixture and keeps the separated steam dry.

Superheaters. Superheaters are bundles of boiler tubing located in the flow path of the hot gases that are created by the combustion of fuel in the boiler furnace. Heat is transferred from the combustion gases to the steam in the superheater tubes.

Superheaters are classified as primary and secondary. Steam passes first through the primary superheater (located in a relatively cool section of the boiler) after leaving the steam drum. There the steam receives a fraction of its final superheat and then passes through the secondary superheater for the remainder.

Reheaters. Reheaters are bundles of boiler tubes that are exposed to the combustion gases in the same manner as superheaters.

Spray Attemperators. Attemperators, also known as desuperheaters, are spray nozzles in the boiler tubes between the two superheaters. These spray nozzles supply a fine mist of pure water into the flow path of the steam to prevent tube damage from overheating. Attemperators are provided for both the superheater and reheater.

Steam Turbines

General

Each turbine manufacturer has unique features in their designs that impact efficiency, reliability, and cost. However, the designs appear similar to a non-steam-turbine engineer. Steam turbines for power plants differ from most prime movers in at least three ways. (1) All are extremely high powered, varying from about 70,000 to 2 million hp, and require a correspondingly large capital investment, which puts a premium on reliability. (2) Turbine life is normally between 30 and 40 years with minimal maintenance. (3) Turbines spend the bulk of their life at constant speed, normally 3600 or 1800 rpm for 60 Hz. These three points dominate the design of the whole power station, particularly of the steam turbine arrangement and materials.

In an earlier subsection it was shown that high steam supply temperatures make for more-efficient turbines. Even so, the range of steam conditions in modern service has narrowed because of these three points. Figure 8.1.8 shows the distribution of steam conditions of one manufacturer for turbines recently put in service. They are reasonably typical of the industry. This is one of the primary reasons that the steam turbines appear similar.

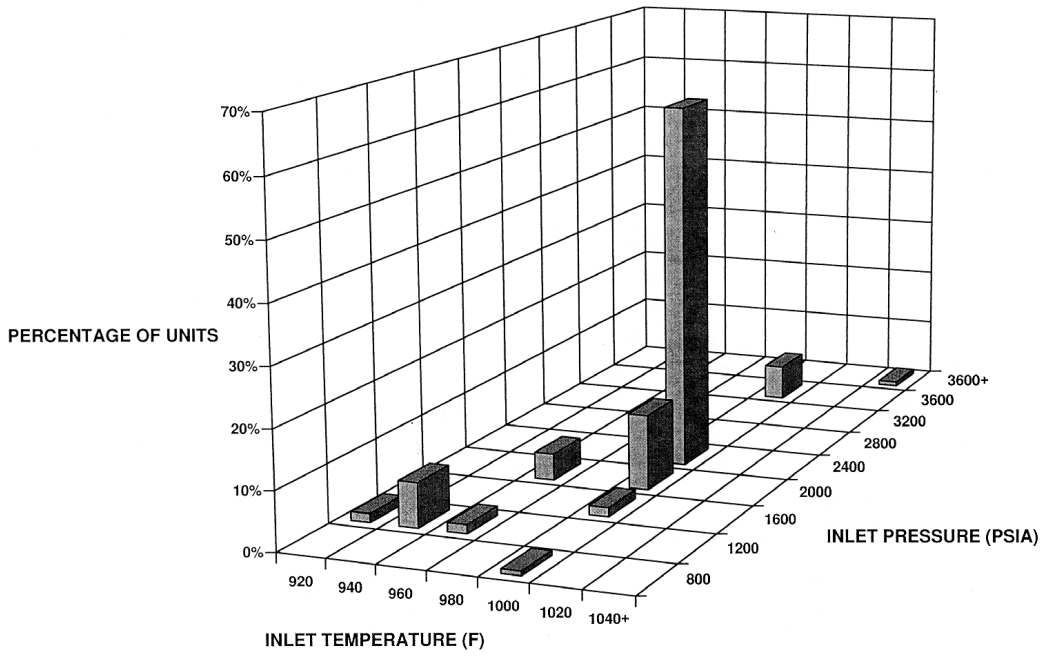


FIGURE 8.1.8 Steam turbine operating conditions.

Blading

The most highly stressed component in steam turbines are the blades. Blades are loaded by centrifugal and steam-bending forces and also harmonic excitation (from nonuniform circumferential disturbances

in the blade path). All blades are loaded by centrifugal and steam-bending loads, and, in general, blades are designed to run when the harmonic excitation is resonant with the natural modes of the blade. If harmonic excitation is permitted on very long blades, the blades become impractically big. Fortunately, as the turbine runs at constant speed, the blade modes can be tuned away from resonant conditions so that the harmonic loads are significantly reduced. This forms a split in blade design, commonly referred to as tuned and untuned blading.

Blades guide steam throughout the turbine in as smooth and collision-free a path as possible. Collisions with blades (incidence) and sudden expansions reduce the energy available for doing work. Until recently, designers would match flow conditions with radially straight blades (called parallel sided). Turbine physics does not recognize this convenience for several reasons. The most visually obvious is the difference in tangential velocity between blade hub and tip. Twisted blades better match the flow (and area) conditions. The manufacturing process was costly and this cost confined application to long blades. Now, with numerical control machine tools, twist is being spread throughout the turbine. Twisted blades are a two-dimensional adjustment for a three-dimensional steam flow. The latest blades address the full three-dimensional nature of the flow by curving in three dimensions (bowed blades). Examples of all three classes of blades are shown in Figure 8.1.9.

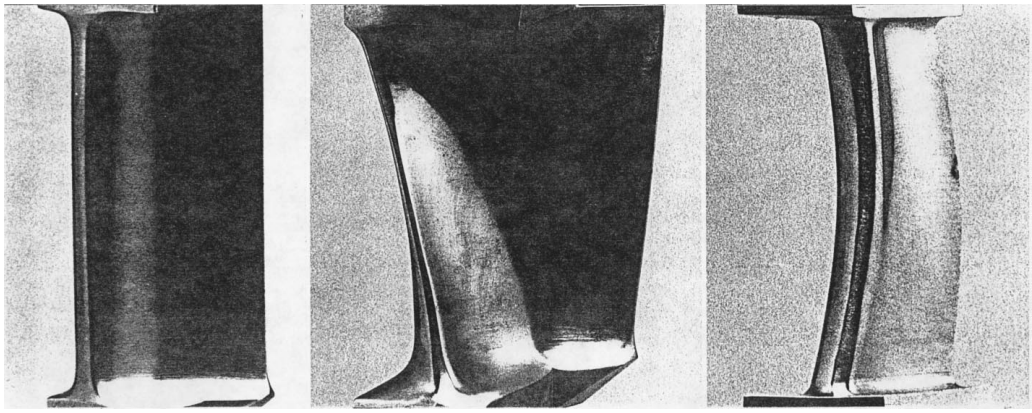


FIGURE 8.1.9 Typical steam turbine blades.

Rotors

After blades, steam turbine rotors are the second most critical component in the machine. Rotor design must account for (1) a large forging with uniform chemistry and properties in the high-strength alloys needed; (2) centrifugal force from the rotor itself and the increase from the centrifugal pull of the blades; (3) resistance to brittle fracture potentially occurring in the LP cylinder when the machine is at high speed, but the material is still not up to operational temperature; (4) creep of the high-pressure (HP) and intermediate-pressure (IP) rotors under steady high-temperature load. The life cycle is further complicated by the various transient fatigue loads occurring during load changes and start-up. Two further events are considered in rotor design: torsional and lateral vibrations caused by both harmonic steam and electrical loads. As with tuned blades, this is normally accommodated by tuning the primary modes away from running resonance.

Choosing the Turbine Arrangement

The turbine shaft would be too flexible in one piece if all the blades were to follow sequentially. It is therefore cut up into supportable lengths. The “cuts” in the shaft result in HP, IP, and LP cylinders. Manufacturers address the grouping of cylinders in many different ways, depending upon steam conditions. It is U.S. practice to combine HPs and IPs into one cylinder for the power range of about 250 to 600 MW (rare in the rest of the world). One manufacturer’s grouping, shown in Figure 8.1.10, is fairly

representative of the industry. So far, the text has discussed the steam flow as though it expanded monotonically through the turbine. This is usually not the case for two reasons. The most common steam conditions, shown in Figure 8.1.10, would cause the steam exiting the last row of blades to be very wet and cause excessive erosion. Thermal efficiency can be raised by removing the steam from the turbine, reheating, and then returning it to the blade path; this increases the "average" heat supply temperature. The turbine position for reheat is normally between the HP and IP turbines.

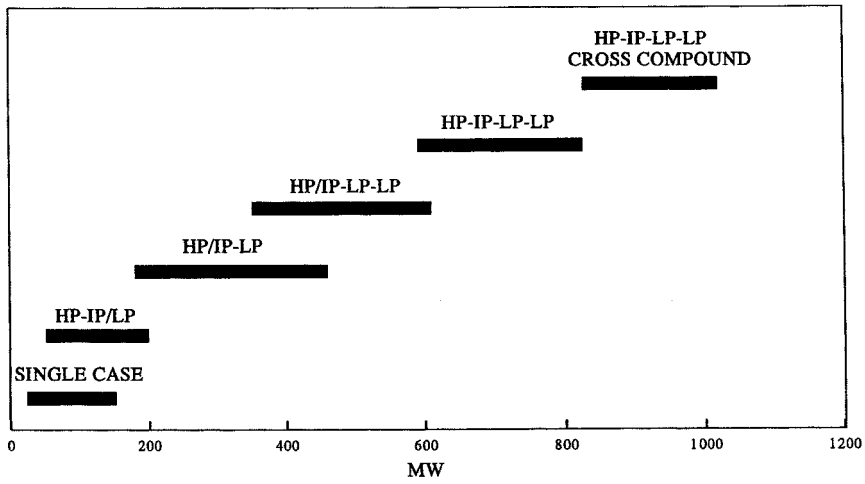


FIGURE 8.1.10 Steam turbine product combinations.

There is one further geometric arrangement. Cylinders need not be all on one shaft with a single generator at the end. A cross-compound arrangement exists in which the steam path is split into two separate parallel paths with a generator on each path. Commonly, the split will be with paths of HP-LP-generator and IP-LP-generator. Torsional and lateral vibration difficulties are more easily prevented with shorter trains, which make the foundation more compact. The primary shortcoming is the need for two generators and resultant controls.

Historically, steam turbines have been split into two classes, reaction and impulse, as explained in Basic Power Cycles. This difference in design makes an observable difference between machines. Impulse turbines have fewer, wider stages than reaction machines. As designs have been refined, the efficiencies and lengths of the machines are now about the same. For a variety of reasons, the longer blades in the LP ends are normally reaction designs. As each stage may now be designed and fabricated uniquely, the line between impulse and reaction turbines will probably disappear. Turbine blading is broadly split between machines as follows:

	Cylinder				
	HP			LP	
	Control Stage	Remainder	IP	Short Blades	End Blade(s)
Reaction turbines	Impulse	Reaction	Reaction	Reaction	Reaction
Impulse turbines	Impulse	Impulse	Impulse	Impulse	Reaction

Materials

Materials are among the most variable of all turbine parts with each manufacturer striving to improve performance by using alloying and heat-treatment techniques. It follows that accurate generalizations are difficult. Even so, the following is reasonably representative:

Item	Common Material Description								
High-temperature HP and IP blades	Moderate and cold temperature stator blades	Moderate temperature rotating blades	Cold LP rotating blades	High temperature rotors	Low temperature rotors	Hot	LP	High-temperature bolting	Cold bolting
Mod'd SS403	SS304	SS403	SS403 or 17/4 PH	1CrMoV, occasionally 12Cr	3.5 NiCrMoV	1.25Cr or 2.25Cr	Carbon, steel	SS422	B16

Cylinders and Bolting

These items are relatively straightforward, especially the LP cylinder, except for the very large sizes and precision required for the castings and fabrications. A large HP-IP cylinder has the temperature and pressure loads split between an inner and outer cylinder. In this case, finding space and requisite strength for the bolting presents a challenge for the designer.

Valves

The turbine requires many valves for speed control, emergency control, drains, hydraulics, bypasses, and other functions. Of these, there are four valves distinguished by their size and duty. They are throttle or stop, governor or control, reheat stop, and reheat interceptor.

The throttle, reheat stop, and reheat interceptor valves normally operate fully open, except for some control and emergency conditions. Their numbers and design are selected for the appropriate combination of redundancy and rapidity of action. The continuous control of the turbine is accomplished by throttling the steam through the governor valve. This irreversible process detracts from cycle efficiency. In some circumstances, the efficiency detraction is reduced by a combination of throttling and reducing the boiler pressure (normally called sliding pressure).

Further Information

Kutz, M. 1986. *Mechanical Engineers' Handbook,* John Wiley & Sons, New York.

Stodola, A. and Loewenstein, L.C. 1927. *Steam and Gas Turbines,* Reprint of 6th ed. 1945, Peter Smith, New York.

Japikse, D. and Nicholas, C.B. 1994. *Introduction to Turbomachinery,* Concepts ETI, Norwich, VT.

Heat Exchangers, Pumps, and Other Cycle Components

Heater Exchangers

Heaters. There are two classifications of condensate and feedwater heaters: the open or direct contact heater and the closed or tube-and-shell heater.

Open Heaters. In an open heater, the extraction or heating steam comes in direct contact with the water to be heated. While open heaters are more efficient than closed heaters, each requires a pump to feed the outlet water ahead in the cycle. This adds cost, maintenance, and the risk of water induction to the turbine, making the closed heater the preferred heater for power plant applications.

Closed Heaters. These employ tubes within a shell to separate the water from the heating steam (see Figure 8.1.11). They can have three separate sections where the heating of the feedwater occurs. First is the drain cooler section where the feedwater is heated by the condensing heating steam before cascading back to the next-lower-pressure heater. The effectiveness of the drain cooler is expressed as the drain cooler approach (DCA), which is the difference between the temperature of the water entering the heater and the temperature of the condensed heating steam draining from the heater shell. In the second section (condensing section), the temperature of the water is increased by the heating steam condensing around

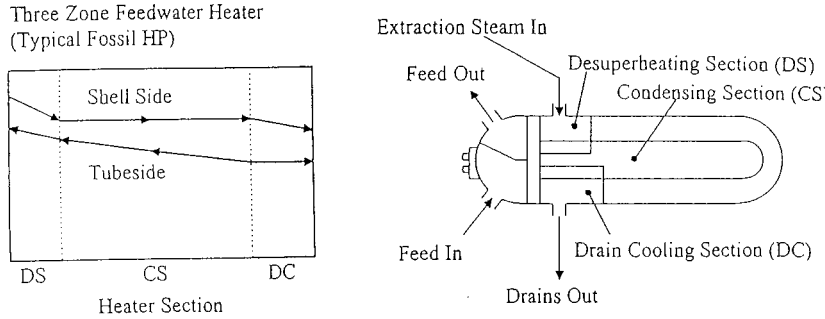


FIGURE 8.1.11 Shell and tube feedwater heater.

the tubes. In the third section (desuperheating section), the feedwater reaches its final exit temperature by desuperheating the extraction steam. Performance of the condensing and superheating sections of a heater is expressed as the terminal temperature difference (TTD). This is the difference between the saturation temperature of the extraction steam and the temperature of the feedwater exiting the heater. Desuperheating and drain cooler sections are optional depending on the location of the heater in the cycle (i.e., desuperheating is not necessary in wet extraction zones) and economic considerations.

The one exception is the deaerator (DA), which is an open heater used to remove oxygen and other gases that are insoluble in boiling water. The DA is physically located in the turbine building above all other heaters, and the gravity drain from the DA provides the prime for the boiler feed pump (BFP).

Two other critical factors considered in heater design and selection are (1) venting the heater shell to remove any noncondensable gases and (2) the protection of the turbine caused by malfunction of the heater system. Venting the shell is required to avoid air-bounding a heater, which reduces the performance or, in extreme cases, puts the heater out of service. Emergency drains to the condenser open when high water levels are present within the shell. Check valves on the heating steam line are also used, and a water detection monitor can be installed to enable operators to take prompt action when water is present.

Condenser. The steam turbines employ surface-type condensers comprising large shell-and-tube heat exchangers operating under vacuum. The condenser (1) reduces the exhaust pressure at the last-stage blade exit to extract more work from the turbine and (2) collects the condensed steam and returns it to the feedwater-heating system. Cooling water circulates from the cooling source to the condenser tubes by motor-driven pumps, which may be centrifugal, propeller, or mixed-flow type. Multiple pumps, each rated less than 100% of required pumping power, are used to allow to operation with one or more pumps out of service and operate more efficiently at part load. Cooling water is supplied from either a large heat sink water source, such as a river, or from cooling towers. The cooling in the cooling tower is assisted by evaporation of 3 to 6% of the cooling water. Airflow is natural draft (hyperbolic towers) or forced draft. The noncondensable gases are removed from the condenser by a motor-driven vacuum pump or, more frequently, steam jet air ejectors which have no moving parts.

Pumps

Condensate Pump. Condensate is removed from the hot well of the condenser and passed through the LP heater string via the condensate pump. Typically, there will be two or more vertical (larger units) or horizontal (medium and small units) motor-driven centrifugal pumps located near the condenser hot well outlet. Depending on the size of the cycle, condensate booster pumps may be used to increase the pressure of the condensate on its way to the DA.

Feedwater Booster Pump. The DA outlet supplies the feedwater booster pump which is typically a motor-driven centrifugal pump. This pump supplies the required suction head for the BFP.

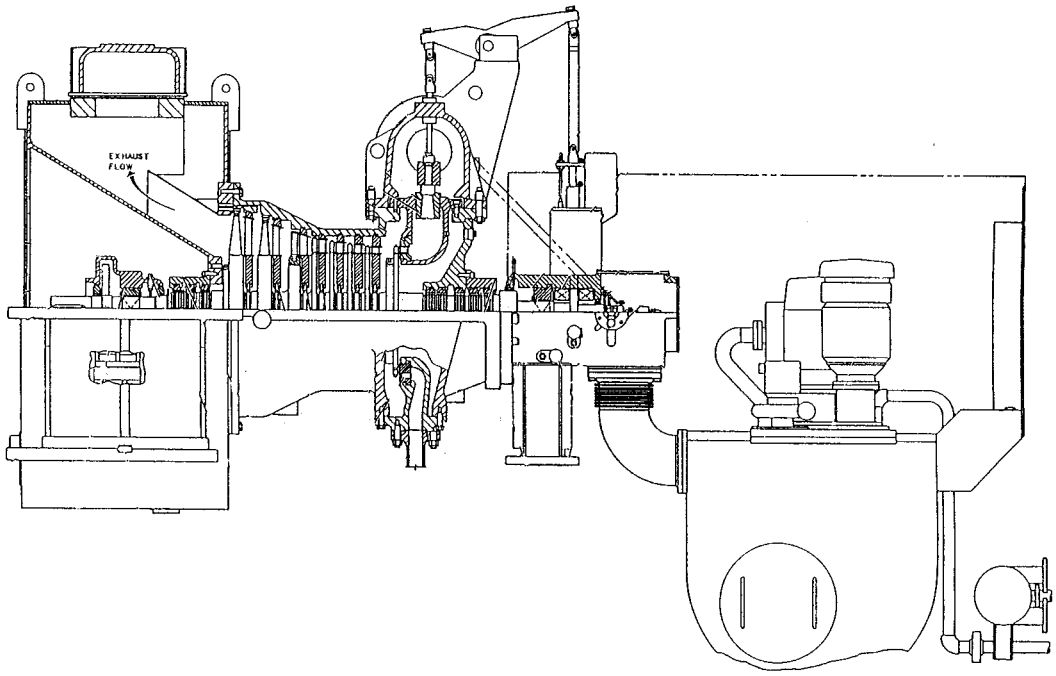


FIGURE 8.1.12 Boiler feed pump turbine.

Boiler Feed Pump. These pumps are multiple-stage centrifugal pumps, which, depending on the cycle, can be turbine or motor driven. BFP turbines (BFPT), Figure 8.1.12, are single-case units which draw steam from the main turbine cycle and exhaust to the main condenser. Typical feedpump turbines require 0.5% of the main unit power at full-load operation. Multiple pumps rated at 50 to 100% each are typically used to allow the plant to operate with one pump out of service.

Further Information

British Electricity International, 1992. *Modern Power Station Practice,* 3rd ed., Pergammon Press, Oxford.

Lammer, H.B. and Woodruff, 1967. *Steam Plant Operation,* 3rd ed., McGraw-Hill, New York.

Powell, C. 1955. *Principles of Electric Utility Operation,* John Wiley & Sons, New York.

Generators

The electric generator converts rotating shaft mechanical power of the steam turbine to three-phase electrical power at voltages of between 13.8 and 26 kV, depending upon the power rating. The generator comprises a system of ventilation, auxiliaries, and an exciter. Figure 8.1.13 shows an installed hydrogen-cooled generator and brushless exciter of about 400 MW. Large generators greater than 25 MW usually have a solid, high-strength steel rotor with a DC field winding embedded in radial slots machined into the rotor. The rotor assemblage then becomes a rotating electromagnet that induces voltage in stationary conductors embedded in slots in a laminated steel stator core surrounding the rotor (see Figure 8.1.14).

The stator conductors are connected to form a three-phase AC armature winding. The winding is connected to the power system, usually through a step-up transformer. Most steam turbines driven by fossil-fired steam use a two-pole generator and rotate at 3600 rpm in 60-Hz countries and 3000 rpm in

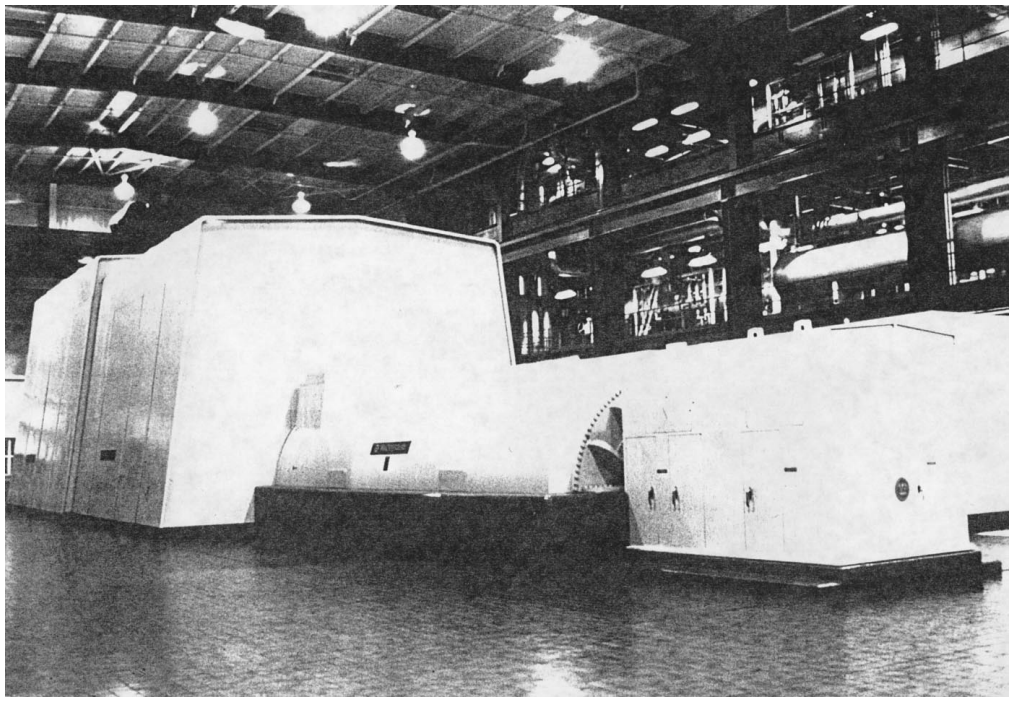


FIGURE 8.1.13 Generator and exciter.

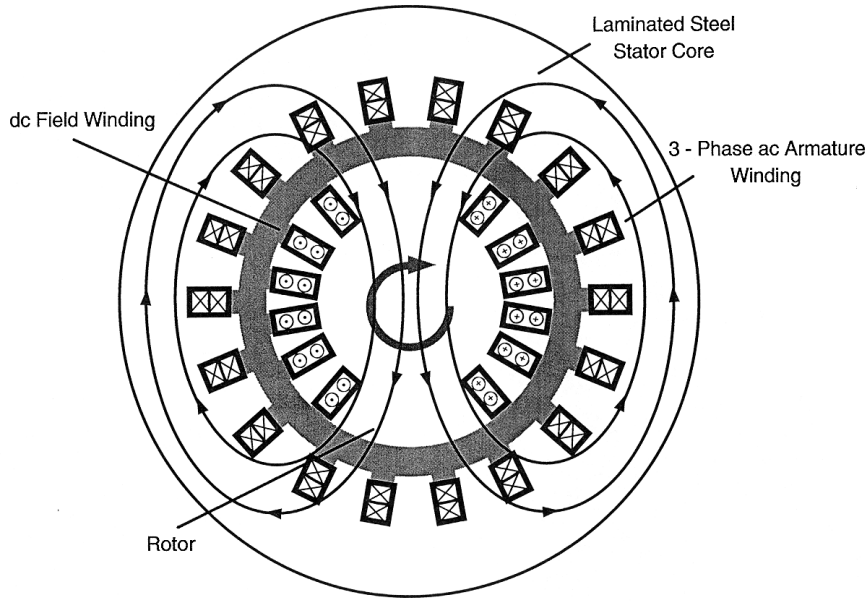


FIGURE 8.1.14 Generator magnetic paths.

50-Hz countries. Most large steam turbines driven by nuclear steam supplies use a four-pole generator and rotate at 1800 or 1500 rpm for 60 and 50 Hz, respectively.

Generator Ventilation

Cooling the active parts of the generator is so important that generators are usually classified by the type of ventilation. Air-cooled generators are used commonly up to 100 MW, though some applications exist up to 200 MW. Some use ambient air drawing air through filters and others recirculate air through air-to-water heat exchangers. Above 100 MW, most manufacturers offer hydrogen for the overall cooling, sometimes up to 1400 MW. Hydrogen has 14 times the specific heat of air and is used at lower density. This contributes to much better cooling and much lower windage and blower loss. The frame must be designed to withstand the remote circumstance of a hydrogen explosion and requires shaft seals. Hydrogen is noncombustible with purities greater than 70%. Generator purities are usually maintained well above 90%. Depending upon the manufacturer, generators with ratings above 200 to 600 MW may have water-cooled stator winding, while the remaining components are cooled with hydrogen.

Generator Auxiliaries

Large generators must have a lubrication oil system for the shaft journal bearings. Major components of this system are pumps, coolers, and a reservoir. In most cases, the turbine and generator use a combined system. For hydrogen-cooled generators, a shaft seal system and hydrogen supply system are needed. The shaft seal system usually uses oil pumped to a journal seal ring at a pressure somewhat higher than the hydrogen pressure. Its major components are pumps, coolers, and reservoir, similar to the lubrication system. The hydrogen supply system consists of a gas supply and regulators. A CO_2 supply is used to purge the generator when going from air to hydrogen or vice versa to avoid a combustible hydrogen/air mixture. The stator winding water supply again uses pumps, coolers, and a reservoir. It needs demineralizers to keep the water nonconducting and provisions to control oxygen content to avoid copper oxide corrosion which might break off and clog water passages.

Excitation

The rotor field winding must have a DC source. Many generators use rotating "collector" rings with stationary carbon brushes riding on them to transfer DC current from a stationary source, such as a thyristor-controlled "static" excitation system, to the rotor winding.

A rotating exciter, known as a brushless exciter, is used for many applications. It is essentially a small generator with a rotating rectifier and transfers DC current through the coupling into the rotor winding without the need for collectors and brushes.

Further Information

Fitzgerald, A.E., Kingsley, C.F., and Kusko, A. 1971. *Electric Machinery,* 3rd ed., McGraw-Hill, New York.

Modern Steam Power Plant — An Example

The purpose of a power plant is to generate electric power. It does so by converting chemical energy contained in fuel into thermal energy in steam; thermal energy in steam into mechanical energy in the turbine/generator; and mechanical energy in the turbine/generator into electrical energy.

Operating efficiency of a typical modern steam plant is about 35 to 45%. The primary losses result from (1) heat sink losses in the condenser, (2) boiler losses, and (3) electrical losses.

Steam plant capacities have ranges from 50 to 1600 MW; however, modern plants are being designed for more than 250 MW due to energy demands, system load requirements, and economies of scale in the larger centralized stations.

Major Steam Plant Components

Steam plants comprise three major components (1) boiler, (2) turbine, and (3) main steam condenser. The boiler and turbine are covered in earlier subsections and neither will be repeated here. A graphic of the entire fluid or work cycle is shown in Figure 8.1.15.

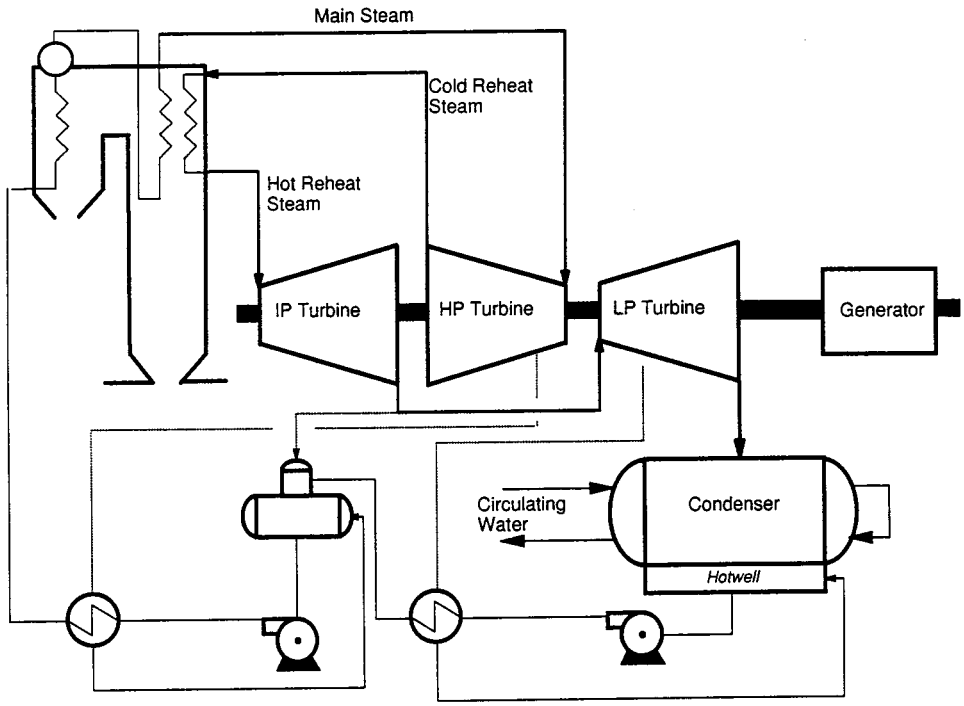


FIGURE 8.1.15 Steam power plant schematic.

Condenser. The condenser (also discussed earlier) is a large heat exchanger that takes the LP turbine exhaust steam and converts it back to water. The steam passes over a bundle of tubes located in the condenser and is cooled by the circulating water which passes through the tubes. The steam is condensed into water drops and collected in the condenser hot well. The condensate is delivered from the condenser hot well through the condensate and feedwater systems and back to the boiler where it becomes steam again.

Fuels. Coal, oil, and gas are used to fuel fossil plants. Although coal possesses the highest carbon content, it also possesses the highest sulfur, nitrogen, and ash content, thereby requiring air pollution-control equipment. Controlling these pollutants requires the installation of scrubbers for sulfur control; overfire air or gas recirculation for in-furnace nitrous oxides (NO_x) control; selective catalytic reduction (SCR) for post-combustion NO_x control; electrostatic precipitators (ESP) or baghouse for fly ash control; and pneumatic, hydraulic, or mechanical ash-handling systems for bottom ash removal. Fuel oil and natural gas are chiefly composed of compounds of hydrogen and carbon (hydrocarbons) with very low percentages of sulfur, nitrogen, and ash and do not require pollution-control equipment.

Power Plant System

Power plants comprise the following main systems: fuel handling, air handling, gas handling, main steam, reheat steam, auxiliary steam, extraction steam, condensate, feedwater, circulating water, and air removal.

Fuel-Handling System. The fuel-handling system consists of delivery, transfer, and processing. Fuel is delivered to the plant from the fuel source by truck, boat, rail (coal and oil), or pipeline (oil and gas). Once delivered, the fuel is transferred from the delivery point to various locations throughout the fuel-handling system. For coal and oil, the fuel is either transferred to storage or sent directly to the boiler. For gas, the fuel is directly transferred to the boiler without any storage. Prior to delivery to the boiler for burning, the fuel must be processed so that it will readily mix with air and burn completely. Coal must be broken down into smaller pieces by breakers and crushers. Oil requires steam, air, or mechanical atomization. Gas requires no processing.

Air/Gas-Handling Systems. Steam plants are classified as either pressurized or balance draft. Pressurized systems include forced-draft fans to provide the necessary air for fuel, an air heater to transfer heat from the exit gas to the inlet air, and a wind box where the air is stored and then directed to the individual burner ports. A balanced draft system includes all of the components of the pressurized systems and induced-draft fans to exhaust the combustion products from the boiler, thus maintaining the furnace under slightly negative pressure.

Main Steam System. The main steam system controls and regulates the flow of high-temperature, HP steam generated in the boiler furnace as it moves from the boiler to the turbine. The components in this system include main steam piping, safety valves, main steam stop valve, steam chest, and turbine control valve.

Reheat Steam System. The reheat steam system improves overall plant efficiency by increasing the energy of steam that has been exhausted from the HP turbine. Steam from the hot reheat steam system is delivered to the IP turbine. The components of the reheat steam system are cold reheat piping, the reheater section of the boiler, hot reheat piping, safety relief valves, reheat stop valves, reheater desuperheater, and intercept valves.

Auxiliary Steam System. The auxiliary steam system directs and regulates auxiliary steam from the cold reheat line to the auxiliary steam users. The auxiliary steam system users are typically steam/air preheating coils, outdoor freeze protection/heat tracing, deaerating heater pegging, turbine-driven BFP testing, turbine seals, and plant heating.

Extraction Steam System. The extraction steam system directs and regulates the flow of the extraction steam from the turbine to the feedwater heaters, BFPT, and auxiliary steam system. The extraction steam heats the feedwater that flows through the heaters, thus improving overall plant efficiency. In large steam plants, six to eight stages of feedwater heating are typical.

Condensate System. The condensate system consists of condensate pumps, LP feedwater heaters, and DA. The condensate pumps remove condensate from the main condenser hot well, increase condensate pressure, and deliver it through the LP heaters to the DA. During this process, the condensate is heated, deaerated, and chemically treated.

Feedwater System. The feedwater system consists of BFPs, HP feedwater heaters, piping, and valves. The boiler feedwater pumps deliver water from the DA storage tank, through the HP heaters, and into the boiler. Feedwater is supplied at sufficient quantities and pressure to satisfy unit demands during startups, shutdowns, and normal operation. The BFP is also the primary source of spray water for the superheater and reheater desuperheaters for control of the main and reheat steam temperatures, respectively.

BFPs can either be turbine or motor driven. Booster pumps may be required to provide additional net positive suction head (NPSH) to the main and start-up BFPs for plants designed with a low DA setting.

Circulating Water System. The circulating water system pumps cooling water through the condenser tubes at sufficient capacity, temperature, and pressure to absorb the latent heat in the LP exhaust steam.

Circulating water systems are classified as once-through systems, when a large water source is available, or recirculating systems employing cooling towers. In once-through systems, circulating water pumps deliver water from the plant water supply (river, lake, or ocean) through the condenser tubes to absorb latent heat in the exhaust steam. Water flows through the system once and is returned to its source. The major parts of this system are screens, pumps, expansion joints, valves, and piping. In recirculating systems, the cooling tower cools the heated circulating water from the main condenser by exposing it to air. The cooled water is stored in a basin located below the tower and is then circulated back through the system by the circulating water pumps.

Air Evacuation System. The air evacuation system removes air and noncondensable gases in the main steam condenser and helps maintain the vacuum created by the volume reduction of the condensing steam during normal operation. The system also establishes a normal vacuum in the condenser prior to turbine start-up.

Further Information

Baumeister, T. and Marks, L.S. 1958. *Standard Handbook for Mechanical Engineers,* 8th ed., McGraw-Hill, New York.

Singer, J.G., Ed. 1991. *Combustion Fossil Power,* 4th ed., Combustion Engineering, Inc., Windsor, CN.

8.2 Gas Turbines

Steven I. Freedman

Overview

Gas turbines are steady-flow power machines in which a gas (usually air) is compressed, heated, and expanded for the purpose of generating power. The term *turbine* is the component which delivers power from the gas as it expands; it is also called an expander. The term *gas turbine* refers to a complete power machine. The term gas turbine is often shortened to simply turbine, which can lead to confusion with the term for an expander.

The basic thermodynamic cycle on which the gas turbine is based is known as the Brayton cycle. Gas turbines may deliver their power in the form of torque or one of several manifestations of pneumatic power, such as the thrust produced by the high-velocity jet of an aircraft propulsion gas turbine engine.

Gas turbine machines vary in size from large, 250,000-hp utility machines, to small automobile, truck, and motorcycle turbochargers producing as little as 5 hp.

Gas turbines are used in electric power generation, propulsion, and compressor and pump drives. The most efficient power generation systems in commercial service are gas turbine combined cycle plants with power-to-fuel energy efficiencies of more than 50% (higher heating value basis) or 55% (lower heating value basis). Systems five points higher in efficiency are now under development and are being offered commercially, and systems of even higher efficiency are considered feasible.

History

The fourth quarter of the 19th century was one of great innovation in power machinery. Along with the spark-ignited gasoline engine, the compression-ignited diesel engine, and the steam turbine, engineers applied their skills to several hot-air engines. Charles Curtis received the first U.S. patent for a complete gas turbine on June 24, 1895. Aegidius Elling built the first gas turbine in 1903, which produced 11 hp.

The first commercial stationary gas turbine engineered for power generation was a 4000-kW machine built by the Brown Boverei Company in Switzerland in 1939.

Aviation provided the impetus for gas turbine development in the 1930s. In Germany, Hans von Ohain's first engine ran in March 1937. Frank Whittle's first engine ran in England in April 1937. The first airplane flight powered by a gas turbine jet engine was in Germany on August 27, 1939. The first British airplane powered by a gas turbine flew on May 15, 1941.

A Swiss railway locomotive using a gas turbine was first run in 1941. The first automobile powered by a gas turbine was a British Rover, which ran in 1950. And, in 1956, a gas turbine-powered Plymouth car drove over 3000 miles on a coast-to-coast exhibition trip in the United States.

Fuels and Firing

The first heat engines were external combustion steam engines. The combustion products never came in contact with the working fluid, so ash, corrosive impurities, and contaminants in the fuel or exhaust did not affect the internal operation of the engine. Later, internal combustion (piston) engines were developed. In these engines, a mixture of air and fuel burned in the space enclosed by the piston and cylinder walls, thereby heating the air. The air and combustion products formed the working fluid, and contacted internal engine parts.

Most gas turbines in use today are internal combustion engines and consequently require clean fuels to avoid corrosion and erosion of critical turbine components. Efforts were made to develop gas turbines rugged enough to burn residual or crude oil. However, due to the higher efficiencies obtainable by burning extremely clean fuel at higher temperatures, there is little current interest in using liquid fuel other than (clean) distillate oil in gas turbines. Interest in the use of residual oil is now centered on gasifying and cleaning these fuels prior to use.

A few externally heated gas turbines have been built for use with heavy oil, coal, nuclear reactor, radioisotope, and solar heat sources. However, none of these has become commercial. The added cost and pressure drop in the externally fired heater make externally fired gas turbines expensive. Because the working fluid temperature cannot be greater than that of the walls of the fired heater, externally fired gas turbines are substantially less efficient than modern internal combustion gas turbines with internally cooled blades.

The only internal combustion coal-fired gas turbine entering commercial service is the pressurized fluidized bed (PFB) combustion system. In the PFB, air discharged from the compressor of the turbine is used to fluidize a bed of limestone or dolomite in which coal is burned. The bed is maintained at modest temperature so that the ash in the coal does not form sticky agglomerates. Fortuitously, this temperature range also minimizes NO_x formation and allows capture of sulfur dioxide (SO_2) in the bed. Bed temperature is maintained in the desired range by immersed boiler tubes. Carryover fly ash is separated from gaseous combustion products by several stages of cyclone inertial separators and, in some cases, ceramic filters. The power turbine is modified to accommodate the combustion products, which after mechanical cleanup may still contain particles as large as 3 to 5 μm.

The most common gas turbine fuels today are natural gas and distillate oil. To avoid hot corrosion by alkali metal sulfates, the total sodium and potassium content of the fuel is typically limited to less than 5 ppm. Liquid fuels may also contain vanadium, which also causes corrosion. Fuels must be ash-free because particles larger than 3 to 5 μm rapidly erode blades and vanes. Experimental prototype gas turbines using pulverized coal pressurized combustors have not demonstrated adequate life. Hybrid systems — in which the moderate-temperature coal combustion products are mechanically cleaned and heated to higher temperature by use of a clean fuel such as natural gas or distillate oil — are the subject of ongoing development.

Efficiency

The term *efficiency* is applied not only to complete power generation machines but also to the individual compression, expansion, and combustion processes that make up the gas turbine operating cycle. Different definitions of efficiency apply in each case. In an **expansion process**, the **turbine efficiency** is the ratio of the actual power obtained to the maximum power that could have been obtained by expanding the gas reversibly and adiabatically between the same initial and final pressures.

Gas turbines typically involve high-speed gas flows, so appreciable differences exist between the static pressure and temperature and the total (or stagnation) pressure and temperature. Care must be taken in interpreting data to be sure that the pressure condition — static or stagnation — at each component interface is properly used.

Irreversible losses in one stage of an expansion process show up as heat (increased temperature) in later stages and add to the power delivered by such stages. Hence, a distinction exists between the polytropic efficiency (used to describe the efficiency of a process of differential pressure change) and the adiabatic (complete pressure change) efficiency. The efficiency of compressors and turbines based on their inlet and outlet pressures is called the isentropic or adiabatic efficiency. Unfortunately, both terms are reported in the literature, and confusion can exist regarding the meaning of the term *efficiency* .

Combustion efficiency in well-engineered and well-built internal combustion gas turbines is almost always close to 100%. The combustion losses appear as carbon monoxide, unburned hydrocarbons, and soot, which are typically below 100 ppm, with clean fuels.

The **gas turbine or engine efficiency** is the ratio of the net power produced to the energy in the fuel consumed. The principal gas turbine fuels are liquid and gaseous hydrocarbons (distillate oil and natural gas) which have high hydrogen content. Consequently, the term *engine efficiency* needs to be qualified as to whether it is based on the higher or the lower heat content of the fuel (the difference between the two being the latent heat of condensation of the water vapor in the products of combustion). Utility fuel transactions are traditionally based on higher heating values, and most engine publications presume the lower heating value of the fuel as the efficiency basis.

Engineers analyze gas turbine machines to evaluate improvements in component performance, in higher temperature and pressure ratio designs, and in innovative cycles. Ideal case cycle calculations generally assume the following:

- Air (with either constant or temperature-dependent specific heats) is the working fluid in both turbine and compressor (with equal mass flows);
- Air is the working fluid in both turbine and compressor but with the turbine mass flow greater by the amount of fuel used.

Components are modeled with or without frictional pressure drops, and heat transfer effectiveness may be ideal (unity) or actual, depending on the purpose of the analysis. Use of compressor air for cooling of high-temperature structure, nozzles, and blades are modeled in varying degrees of complexity. Two-dimensional temperature profiles or pattern factors exist. Component inlet and exit total pressure losses should be included in cycle analyses.

Gas Turbine Cycles

Gas turbine cycles are usually plotted on temperature-entropy (T-S) coordinates. Readers unfamiliar with entropy are referred to the chapter on thermodynamics. The T-S plot is useful in depicting cycles because in an adiabatic process — as is the case for turbines and compressors — the power produced or consumed is the product of the mass flow and the enthalpy change through the process. Thus, temperature difference, which is found on a T-S plot, is proportional to the power involved. Additionally, the heat exchange in a process involving zero power — such as a combustor or heat exchanger — is the product of the absolute temperature and the entropy change. On a T-S chart, the area under a process line for a combustor or heat exchanger is the heat exchanged.

The slope of a constant-pressure line on a T-S diagram is proportional to the absolute temperature. Consequently, lines of constant pressure become steeper, and diverge as the temperature increases. This illustrates that more work is obtained expanding a gas between fixed pressures at higher temperatures than at lower temperatures. Figure 8.2.1 shows a comparison of the process of an ideal and an actual simple cycle gas turbine on a T-S diagram. The increased compressor power consumption and the decreased turbine power generation in the actual cycle are shown to provide an understanding of the differences that component efficiencies make on machine performance.

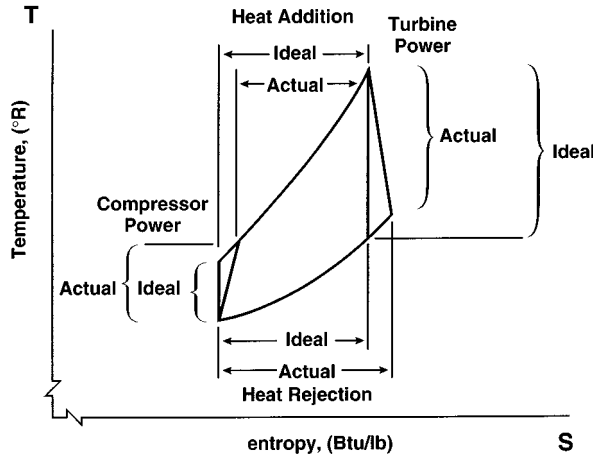


FIGURE 8.2.1 T-S diagram for a simple cycle illustrating the differences in compressor and turbine power for ideal (100% efficient) and actual components.

The incremental amount of power produced per differential pressure change in the gas is given by

$$d(\text{Power/mass flow}) = -RT\, dp/p$$

Two phenomena are illustrated by this equation. First, power is proportional to the absolute temperature of the gas. Second, power is proportional to the percent change in pressure. This latter point is important in understanding the effect of pressure losses in cycle components. In heat exchangers, the proper measure of power lost is the percent pressure drop.

Cycle Configurations

The basic Brayton cycle consists of a compressor, a combustor or burner, and an expander. This configuration is known as the simple cycle. In idealizing the actual cycle, combustion is replaced by constant-pressure heat addition, and the cycle is completed by the assumption that the exhaust to ambient pressure could be followed by a zero-pressure-loss cooling to inlet conditions.

A T-S diagram of the simple cycle gas turbine with an upper temperature limit set by metallurgical conditions is illustrated in Figure 8.2.2 for cycles of low, medium, and high pressure ratios. The heat addition is only by fuel combustion, simplified here to be without mass addition or change in specific heat of the working fluid.

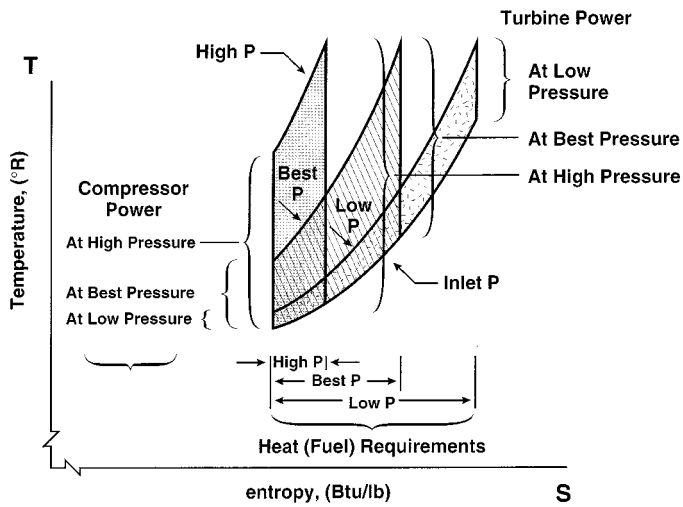


FIGURE 8.2.2 T-S diagram illustrating the power and heat (fuel) requirements at low, best, and high cycle pressures.

It is seen that the low-pressure-ratio cycle requires a large heat addition, which leads to a low efficiency, and the high-pressure-ratio cycle has turbine power output barely greater than the compressor power requirement, thereby leading to low net output and low efficiency. At intermediate pressure ratios, the turbine power output is substantially higher than the compressor power requirement, and the heat addition is modest in comparison with the difference between the turbine and compressor powers. There is an optimum pressure ratio for maximum efficiency, which is a function mainly of the maximum gas temperature in the machine, and to a lesser extent, by the component efficiencies, internal pressure losses, and the isentropic exponent. There is another optimum pressure ratio for maximum specific power (power per unit mass flow).

As the achievable turbine inlet temperature increases, the optimum pressure ratios (for both maximum efficiency and maximum specific power) also increase. The optimum pressure ratio for maximum specific power is at a lower pressure level than that for maximum efficiency for all cycles not employing a recuperator. For cycles with a recuperator, the reverse is true: maximum efficiency occurs at a lower

pressure ratio than maximum specific power. Heavy-duty utility and industrial gas turbines are typically designed to operate near the point of maximum specific power, which approximates lowest equipment cost, while aeroderivative gas turbines are designed to operate near the point of maximum efficiency, approximating highest thrust. Figure 8.2.3 shows a performance map (efficiency as a function of power per unit of air flow) for a simple cycle gas turbine for two turbine inlet temperatures. It is seen that at higher temperature, both the efficiency and specific power increase, as well as the optimum pressure ratios for both the maximum efficiency and maximum specific power conditions.

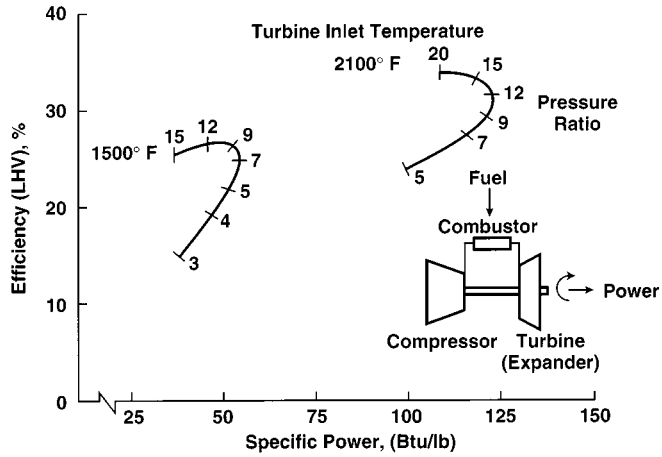


FIGURE 8.2.3 Performance map of a simple cycle gas turbine.

Aircraft gas turbines operate at temperatures above the limit of turbine materials by using blades and vanes with complex internal cooling passages. The added cost is economically justified because these machines can command high prices in the aircraft propulsion marketplace. Aeroderivative engines have higher pressure ratios, higher efficiencies, and lower exhaust temperatures than heavy-duty machines. The stationary power markets served by aeroderlvative gas turbines are principally pipeline compressor stations and oil/gas production wells. Aeroderivative gas turbines also are economically advantageous for intermediate-duty applications.

Components Used in Complex Cycles

Recuperators and **regenerators** recover heat from the turbine exhaust and use it to preheat the air from the compressor before it enters the combustor, thereby saving fuel. This heat transfer is shown in Figure 8.2.4. While recuperators and regenerators are quite similar thermodynamically, they are totally different in design. Recuperators are conventional heat exchangers in which hot and cold gases flow steadily on opposite sides of a solid (usually metal) wall.

Regenerators are periodic-flow devices. Fluid streams flow in opposite directions through passages in a wheel with heat storage walls. The wheel rotates, transferring heat from one stream to the other. Regenerators usually use a nest of very small parallel passages oriented axially on a wheel which rotates between hot and cold gas manifolds. Such regenerators are sometimes used in industrial processes for furnace heat recovery, where they are referred to as heat wheels. Because regenerators are usually more compact than recuperators, they are used in automotive gas turbines (under development). The difficulty in using regenerators on gas turbines intended for long life is that the two gas streams are at very different pressures. Consequently, the seals between the manifolds and the wheel must not leak excessively over the maintenance overhaul interval of the engine. If they do, the power loss due to seal leakage will compromise engine power and efficiency. Figure 8.2.5 shows a performance map for the regenerative

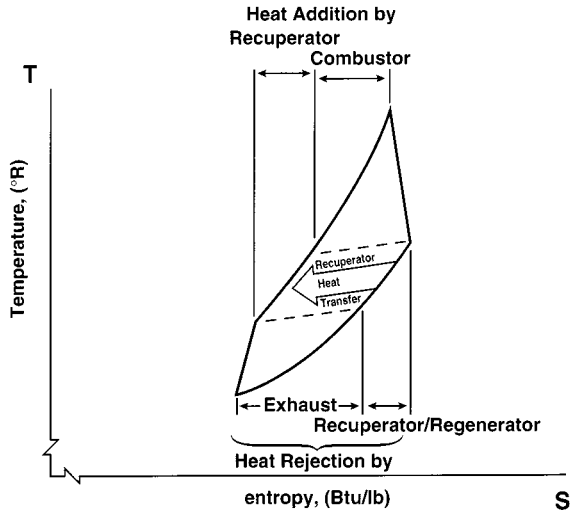


FIGURE 8.2.4 T-S diagram illustrating the heat transfer from the turbine exhaust to the compressor discharge accomplished by a recuperator/regenerator.

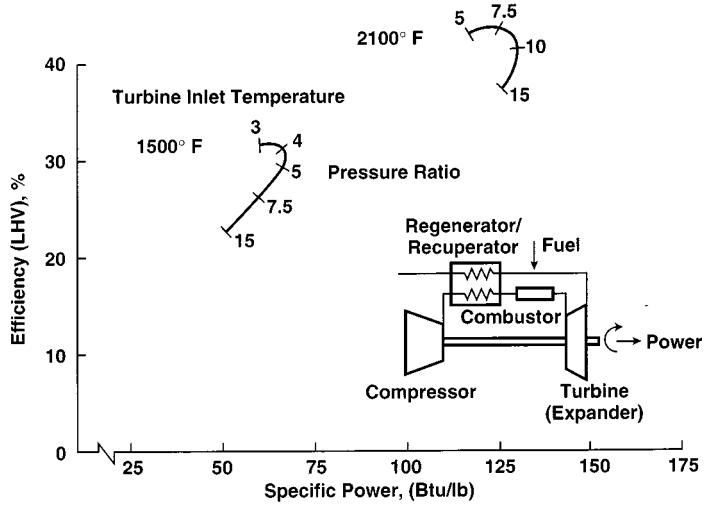


FIGURE 8.2.5 Performance map of a regenerative cycle gas turbine.

gas turbine cycle for two temperatures. It is seen that as the temperature increases, the efficiency, specific power, and optimum pressure ratio all increase.

Current research on the recovery of gas turbine exhaust heat includes examination of thermochemical recuperation, where exhaust heat is used to effect a chemical reaction (reforming) of the fuel with steam, thereby increasing the heating value of the fuel. Although this process is feasible, research is underway to determine if it is practical and economic.

Industrial process compressors frequently use **intercoolers** to reduce compressor power when the compressor has a high pressure ratio and operates for a large number of hours per year. When analyzing cycles with intercoolers, the added pressure drops in the compressor interstage entrance and exit diffuser and scroll and the pressure drop in the intercooler itself should be included.

In a similar manner, turbine reheat can be used to increase the power output of a large-pressure-ratio turbine. This is the thermodynamic principle in turbojet afterburner firing. Turbine reheat increases

power, but decreases efficiency unless the turbine exhaust heat is used for additional power generation, as is the case with a combined cycle, or is used with a recuperator to preheat combustor inlet air.

Intercoolers and reheat burners increase the temperature difference between the compressor and turbine discharges, thereby increasing the opportunity to use a recuperator to preheat the burner air with exhaust heat. An intercooled recuperated (ICR) machine is at present in development. The efficiency decrease at part load of an ICR gas turbine is much less than of conventional simple cycle machines.

Small gas turbines have uncooled turbine blades as a result of the difficulty in manufacturing extremely small cooling passages in small blades. This results in low efficiencies, making it difficult for such turbines to compete with high-volume production (low-cost) reciprocating (piston) engines. The low-pressure-ratio recuperated cycle has greater efficiency, although at higher cost. The recuperated cycle is finding favor in programs for small (under 300-kW) gas turbines used for stationary power. The recuperated cycle is efficient enough in comparison with piston engines (Otto cycles) to be of interest to automotive power plant engineers. Current designs of automotive gas turbines (AGT) have ceramic turbines and combustors and use ceramic regenerators made from a spirally wound corrugated structure with gas passages about a millimeter in diameter.

Because of their compact size, low emissions, and light weight, gas turbines are also being considered for hybrid engine-battery vehicles. Proponents are pursuing the low-pressure-ratio recuperated gas turbine as the way to obtain high efficiency and low emissions in a compact power plant.

An ingenious gas turbine cycle is the closed cycle in which the working fluid is sealed in the system. Heat is added to the fluid with an externally fired heater and extracted from the fluid through heat exchangers. The working fluid may be any gas, and the density of the gas may be varied — to vary the power delivered by the machine — by using a gas storage cylinder connected to the compressor discharge and inlet. The gas storage system is at an intermediate pressure so that it can discharge gas into the lowest pressure point in the cycle and receive gas from the highest pressure point in the cycle. About ten such units were built between 1938 and 1968; however, in spite of its sophistication, the added cost and low efficiency prevented this system from becoming economic.

The exhaust from a gas turbine is quite hot and can be used to raise steam, which can then be used to generate additional power with a steam turbine. Such a compound gas turbine-steam turbine system is referred to as a **combined cycle**. Figure 8.2.6 shows a schematic diagram of the equipment in a combined cycle. Because the exhaust of heavy-duty machines is hotter than that of aeroderivative machines, the gain in combined cycle system efficiency through use of the steam bottoming cycle described above is greater for heavy-duty machines than for aeroderivatives. Indeed, heavy-duty machines are designed with two criteria in mind: achieving lowest cost for peaking (based on the simple cycle configuration) and achieving highest efficiency in combined cycle configuration for baseload use. The optimum pressure ratios for these two system configurations are very close. Steam bottoming cycles used in combined cycles usually use steam at multiple pressure levels to increase efficiency.

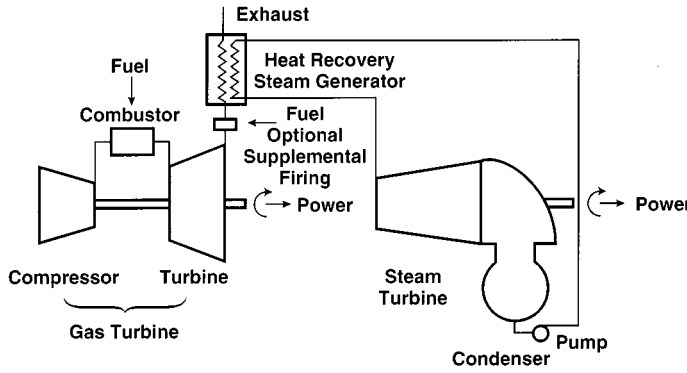


FIGURE 8.2.6 Combined (Brayton-Rankine) cycle.

Another system in which the power and efficiency of a gas turbine is increased through the use of steam is the **steam-injected gas turbine**. Figure 8.2.7 shows a schematic diagram of a steam-injected gas turbine cycle. Here the turbine exhaust flows into a heat recovery steam generator (HRSG) operating at a pressure somewhat higher than the compressor discharge pressure. The steam is introduced into the gas turbine at the combustor. The steam-air mixture then passes into the turbine, where the augmented mass flow increases the power produced by the turbine. Additional fuel is required by the combustor because the steam must be heated from the HRSG delivery temperature to the combustor discharge temperature. Typical turbines can accommodate only a limited additional mass flow — from 5 to 15%, depending on the design of the original gas turbine. Steam-injected gas turbines enable the host to use the steam for industrial purposes, space heating, or for the generation of additional power.

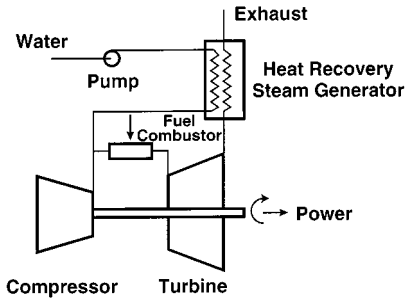


FIGURE 8.2.7 Steam-injected gas turbine.

A group of cycles under consideration for development involve the use of **adiabatic saturators** to provide steam at compressor discharge pressure to augment greatly the mass flow through the turbine, and consequently increase cycle power and efficiency. In the adiabatic saturator, water flows in a countercurrent path to the compressor discharge air in a mass transfer tower. Such equipment is often used in the chemical processing industries. The saturated air is preheated in a turbine exhaust heat recuperator. This cycle is called the **humid air turbine**, or HAT, cycle. The HAT cycle is particularly useful in using the low-temperature heat generated in coal-gasification-fueled gas turbine power plants. As the mass flow through the turbine is significantly augmented, engineers can no longer use the expansion turbine which was matched to the compressor in a conventional simple cycle gas turbine.

Figure 8.2.8 shows performance maps for the gas turbine cycles of major interest for a turbine inlet temperature typical of new products. Intercooling increases the specific power appreciably when compared with a simple cycle; however, such improvement requires an increase in pressure ratio. Recuperated cycles have considerably higher efficiency than similar cycles without recuperation. The effect of pressure ratio on the performance of recuperated cycles is opposite to that of similar cycles without recuperation. For recuperated cycles, the pressure ratio for maximum efficiency is considerably lower than for maximum specific power. Performance maps such as these are used in screening cycle alternatives for improved performance. Individual curves are generated for specific component performance values for use as a guide in developing new or improved machines.

Upper Temperature Limit

Classically, gas turbine engineers often spoke of a metallurgical limit in reference to maximum turbine inlet temperature. Later, turbine vane and blade cooling became standard on large machines. This situation creates a temperature difference between the combustion products flowing through the turbine and the turbine blade wall. Thus, because heat can be removed from the blades, the turbine can be operated with a combustion gas temperature higher than the metallurgical limit of the blade material. As a rule, the blades and vanes in new large gas turbines contain complex internal passages, through which up to 20% of compressor discharge air is directed. The cooling air first flows through internal convective cooling

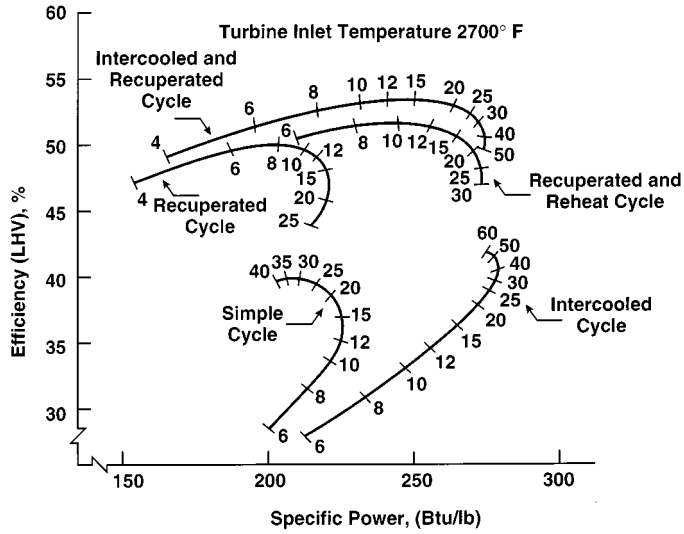


FIGURE 8.2.8 Specific power (Btu/lb).

passages, then through impingement passages, where the air is directed at the blade and vane walls, and finally through small holes in the blade, where it is used to provide a low-temperature film over the blade surface. This film cooling of the surface reduces heat transfer to the blade.

The design of blade and vane cooling passages is an extremely competitive endeavor because greater cooling enables use of higher combustion temperatures without exceeding the metallurgical limit of the blade material. However, a balance between air flow for cooling and air flow for power must be achieved; the cooling air flowing within a blade drops in pressure without producing any power within that stage (although it is available for power in later stages). In the newest gas turbines, blade cooling, the difference between turbine inlet gas temperature and blade metal temperature, is around 1000°F.

Some of the latest large gas turbines being introduced to the market in the 1997 to 2000 period are being offered for combined cycle application with closed-circuit **steam cooling** of selected hot section parts. Steam cooling reduces the need for air cooling, so that more of the compressor discharge air can be used for NO_x reduction in the combustor and for power generation. The heat transferred to the steam increases the efficiency of the bottoming cycle. The additional combustion products which flow through the high-pressure portions of the turbine generate substantially more power, thereby increasing both the power output and the efficiency of the machine. With more air for combustion, the fuel can be burned as a leaner mixture, with either less NO_x produced, or, as is preferred, with higher-temperature gases going to the turbine and the same NO_x (or a combination of these benefits).

Materials

The high-technology parts of a gas turbine are its hot section parts: blades, vanes, combustors and transition pieces. Gas turbine power, efficiency, and economics increase with the temperature of the gas flowing through the turbine blade passages. It is in the fabrication of these hot section parts that manufacturers are most competitive. Materials are selected to survive in serviceable condition for over 50,000 hr and associated numbers of thermal cycles. Ceramic coatings protect materials from oxidation and corrosion and provide thermal insulation, permitting higher gas temperatures.

Gas turbine alloys are frequently referred to as superalloys because of their extremely high strength at high temperatures. These superalloys are nickel based (such as IN 738), cobalt based (such as FSX-414), or with a nickel-iron base such as Inconel 718. Nickel resists oxidation and is creep resistant, but is subject to corrosive sulfidation. Alloy and manufacturing advancements have been led by the needs of military aircraft engines. Coating developments for corrosion resistance have been led by the needs

of stationary power for overhaul intervals as large as 50,000 hr. The developmental needs of automotive gas turbines have led to significant advances in strength and reliability of high-temperature ceramic components, including radial inflow turbines. Ceramic materials, principally silicon nitride, are expected to enter service soon in small gas turbines.

Combustion

Gas turbine combustors appear to be simple in design, yet they solve several difficult engineering challenges. Until relatively recently, gas turbine combustors employed a (turbulent) diffusion flame design approach, which created the most compact flame. European heavy-duty gas turbine manufacturers — with substantial interest in burning heavy fuel oils — preferred large, off-engine combustors, often called silo combustors because of their appearance, in order to obtain lower flame velocities and longer residence times. American heavy-duty gas turbine manufacturers use compact on-engine combustors and design for gaseous and clean (distillate) liquid fuels. Aeropropulsion gas turbines require the smallest possible frontal area and use only clean liquid fuels; hence, they use on-engine combustors.

Recently, stationary engines have been required to reduce NO_x emissions to the greatest extent possible, and combustors on stationary gas turbines first modified their diffusion flame combustors and employed water and steam injection to quench flame hot spots. Most recently, designs changed to the lean-premixed process. With the improved blade cooling, materials, and coatings now in use, the material limits on turbine inlet temperature and the NO_x emission limits on combustor temperature appear to be converging on a combustion-temperature asymptote around 2700°F (1482°C). This may be increased if catalytic combustors prove practical.

Mechanical Product Features

In view of the need to achieve all the performance features described above, one must keep in mind that a gas turbine is a high-speed dynamic machine with numerous machine design, materials, and fabrication features to consider. Major issues include the following: critical shaft speed considerations, bearing rotational stability, rotor balancing, thrust bearing design, bearing power loss, oil lubrication system, oil selection, air filter design and minimization of inlet and exhaust diffuser pressure drops, instrumentation, controls, diagnostic systems, scheduled service and inspection, overhaul, and repair. All of these topics must be addressed to produce a cost-effective, reliable, long-lived, practical gas turbine product that will satisfy users while also returning to investors sufficient profit for them to continue to offer better power generation products of still higher performance.

Defining Terms

Adiabatic saturator: A combined heat-and-mass-exchanger whereby a hot gas and a volatile liquid pass through a series of passages such that the liquid is heated and evaporates into the gas stream.

Combined cycle: An arrangement of a gas turbine and a stream turbine whereby the heat in the exhaust from the gas turbine is used to generate steam in a heat recovery boiler which then flows through a steam turbine, thereby generating additional power from the gas turbine fuel.

Combustion efficiency: Ratio of rate of heat delivered in a device which burns fuel to the rate of energy supplied in the fuel.

Expansion process: Process of power generation whereby a gas passes through a machine while going from a condition of high pressure to one of low pressure, usually the machine produces power.

Gas turbine or engine efficiency: The ratio of the net power delivered (turboexpander power minus compressor and auxiliary power) to the rate of energy supplied to the gas turbine or engine in the form of fuel, or, in certain cases such as solar power, heat.

Humid air turbine: A gas turbine in which the flow through the expander is augmented by large amounts of steam generated by use of an adiabatic saturator.

Intercooler: A heat exchanger used to cool the flow between sections of a compressor such that the high pressure section acts on a stream of reduced volumetric flow rate, thereby requiring less overall power to compress the stream to the final pressure.

Recuperator: A heat exchanger in which the hot and cold streams pass on opposite sides of a wall through which heat is conducted.

Regenerator: A heat exchanger in which the hot and cold streams alternately heat and cool a wall whose temperature rises and falls, thereby transferring heat between the streams.

Steam cooling: A process in which steam is used as the heat transfer fluid to cool a hot component.

Steam-injected gas turbine: A system in which the gas turbine flow is augmented by steam, thereby generating additional power.

Turbine efficiency: Ratio of the power delivered in an expansion process employing a turbine as the expander to the maximum power which could be produced by expanding the gas in a reversible adiabatic (isentropic) process from its initial pressure and temperature to its final pressure to the actual power.

Further Information

Wilson, D.G. 1984. *The Design of High-Efficiency Turbomachinery and Gas Turbines*, MIT Press, Cambridge, MA.

Kerrebrock, J. 1992. *Aircraft Engines and Gas Turbines*, MIT Press, Cambridge, MA.

Boyce, M.P. 1982. *Gas Turbine Engineering Handbook,* Gulf Publishing, Houston, TX.

Sawyer's Gas Turbine Engineering Handbook, Vol. 1: *Theory and Design,* Vol. 2: *Section and Application*, Vol. 3: *Accessories and Support*, Turbomachinery International Publications, Norwalk, CT, 1985.

Appendix

Equations for gas turine calculations based on the use of a perfect gas as the working fluid.

Perfect gas law	$pv = RT$
Gas constant	$R = \check{R}$/molecular weight
For air (molecule weight of 28.97)	$R = 286.96$ J/kg · K $= 0.06855$ Btu/lb$_m$ · °R $= 53.32$ ft · lb$_f$/lb$_m$ · °R
Universal gas constant	$\check{R} = 8313$ J/kg · mol · K $= 1.986$ Btu/lb · mol · °R $= 1545$ ft · lb$_f$/lb · mol · °R
Relationships of properties	$c_p = c_v + R$
Isentropic exponet	$\gamma = c_p/c_v$ (air, $\gamma = 1.4$)
	$(\gamma - 1)/\gamma = R/c_p$
Isentropic process	$pv^\gamma = \text{constant}$
	$P_2/P_1 = (T_2/T_1)^{\gamma/(\gamma-1)}$
Prolytropic process	$pv^n = \text{constant}$
	$P_2/P_1 = (T_2/T_1)^{n/(n-1)}$
Pressure ratio	$r = P_2/P_1$
Ratio of stagnation T° and p° to static T and p	$\frac{T^\circ}{T} = 1 + \frac{\gamma - 1}{2} M^2$
	$\frac{p^\circ}{p} = \left(1 + \frac{\gamma - 1}{2} M^2\right)^{\gamma/(\gamma-1)}$
Mach number	$M = V \Big/ \sqrt{g_c \gamma RT}$
Gravitational constant	$g_c = ma/F$
Subscripts	t = turbine
	c = compressor

f = fuel
i = inlet
e = exit

Cycle efficiency:

$$\eta = \frac{\dot{m}_t \Delta h_t - \dot{m}_c \Delta h_c}{\dot{m}_f HV}$$

where HV = heating value of fuel.
For specific heat independent of temperature and small mass flow of fuel in comparison to air:

$$\eta = \frac{\Delta T_t - \Delta T_c}{\Delta T_b}$$

Isentropic efficiency (finite pressure ratio):

$$\eta_t = \Delta T \text{ actual} / \Delta T \text{ isentropic}$$

$$\eta_t = \frac{1 - T_e/T_i}{1 - r^{(\gamma-1)/\gamma}}$$

or

$$\eta_t = \frac{1 - r^{(n-1)/n}}{1 - r^{(\gamma-1)/\gamma}}$$

and

$$\eta_c = \Delta T \text{ isentropic} / \Delta T \text{ actual}$$

$$\eta_c = \frac{r^{(\gamma-1)/\gamma} - 1}{T_e/T_i - 1}$$

or

$$\eta_c = \frac{r^{(\gamma-1)/\gamma} - 1}{r^{(\eta-1)/\eta} - 1}$$

Polytropic efficiency (differential pressure ratio):

$$\eta_t = \frac{(n-1)/n}{(\gamma-1)/\gamma}$$

and

$$\eta_c = \frac{(\gamma-1)/\gamma}{(n-1)/n}$$

Relationships between isentropic and polytropic efficencies:

$$\eta_{s,c} = \frac{r^{(\gamma-1)/\gamma} - 1}{r^{(\gamma-1)/\gamma \eta_{pc}} - 1}$$

$$\eta_{s,t} = \frac{1 - r^{(\gamma-1)/\gamma \eta_{p,t}}}{1 - r^{(\gamma-1)/\gamma}}$$

$$\eta_{p,c} = \frac{\ln r^{(\gamma-1)/\gamma}}{\ln\left[\dfrac{r^{(\gamma-1)/\gamma} - 1}{\eta_{s,c}} + 1\right]}$$

$$\eta_{p,t} = \frac{\ln\left[1 - \eta_{s,t}\left(1 - r^{(\gamma-1)/\gamma}\right)\right]}{\ln r^{(\gamma-1)/\gamma}}$$

8.3 Internal Combustion Engines

David E. Klett and Elsayed M. Afify

Introduction

This section discusses the two most common reciprocating internal combustion (IC) engine types in current use; the **spark ignition** (SI) and the **compression ignition** (CI or diesel) **engine**. Both the Stirling engine (technically, an external combustion engine) and the gas turbine engine are covered in other sections of this chapter. Space limitations do not permit detailed coverage of the very broad field of IC engines. For a more detailed treatment of SI and CI engines and for information on variations, such as the Wankel rotary engine and the Miller cycle engine (a variation on the reciprocating four-stroke SI engine introduced in production by Mazda in 1993), the reader is referred to the several excellent textbooks on the subject, technical papers, and other sources that are included in the list of references and the Further Information section.

Basic SI and CI engines have not fundamentally changed since the early 1900s with the possible exception of the introduction of the Wankel rotary SI engine in the 1960s (Norbye, 1971). However, major advances in the areas of materials, manufacturing processes, electronic controls, and computer-aided design have led to significant improvements in dependability, longevity, thermal efficiency, and emissions during the past decade. Electronic controls, in particular, have played a major role in efficiency gains in SI automotive engines through improved control of the fuel injection and ignition systems that control the combustion process. Electronic control of diesel fuel injection systems is also becoming more common and is producing improvements in diesel emissions and fuel economy.

This section presents the fundamental theoretical background of IC engine function and performance, including **four-stroke** and **two-stroke** SI and CI **engines**. Sections on combustion, emissions, fuels, and intake pressurization (**turbocharging** and **supercharging**) are also included.

Engine Types and Basic Operation

IC engines may be classified by a wide variety of characteristics, the primary ones being SI vs. CI, four-stroke vs. two-stroke, and reciprocating vs. rotary. Other possible categories of classification include intake type (naturally aspirated vs. turbocharged or supercharged), number of cylinders, cylinder arrangement (in-line, vee, opposed), cooling method (air vs. water), fueling system (injected vs. carbureted), valve gear arrangement (overhead cam vs. pushrod), type of **scavenging** for two-stroke engines (cross, loop, or uniflow), and type of injection for diesel engines (direct vs. indirect).

Four-Stroke SI Engine

Figure 8.3.1 is a cross-section schematic of a four-stroke SI engine. The SI engine relies on a spark plug to ignite a volatile air-fuel mixture as the piston approaches **top dead center** (**TDC**) on the compression stroke. This mixture may be supplied from a carburetor, a single throttle-body fuel injector, or by individual fuel injectors mounted in the intake port of each cylinder. One combustion cycle involves two revolutions of the crankshaft and thus four strokes of the piston, referred to as the intake, compression, power, and exhaust strokes. Intake and exhaust valves control the flow of mixture and exhaust gases into and out of the cylinder, and an ignition system supplies a spark-inducing high voltage to the spark plug at the proper time in the cycle to initiate combustion. On the intake stroke, the intake valve opens and the descending piston draws a fresh combustible charge into the cylinder. During the compression stroke, the intake valve closes and the fuel-air mixture is compressed by the upward piston movement. The mixture is ignited by the spark plug, typically somewhat before TDC. The rapid **premixed homogeneous combustion** process causes a sharp increase in cylinder temperature and pressure that forces the piston down for the power stroke. Near **bottom dead center** (BDC) the exhaust valve opens and the cylinder pressure drops rapidly to near atmospheric. The piston then returns to TDC, expelling the

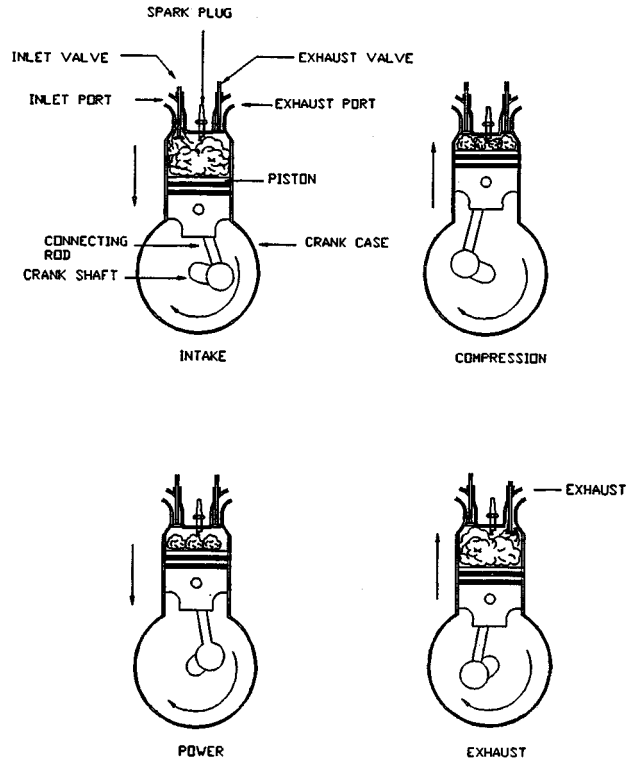


FIGURE 8.3.1 Schematic diagram of four-stroke SI engine.

exhaust products. At TDC, the exhaust valve closes and the intake valve opens to repeat the cycle again. Figure 8.3.2 is a cutaway drawing of a modern high-performance automotive SI engine. This is a fuel-injected normally aspirated aluminum alloy V-8 engine of 4.6 L displacement with dual overhead cams for each cylinder bank. Peak power output is 228 kw at 5800 rpm.

Two-Stroke SI Engine

The two-stroke SI engine completes a combustion cycle for every revolution of the crankshaft by essentially overlapping the power and exhaust functions in one downward stroke and the intake and compression processes in one upward stroke. A single-cylinder, crankcase-scavenged, two-stroke SI engine is illustrated schematically in Figure 8.3.3. The operation is as follows. On the upward stroke, the piston first covers the transfer port and then the exhaust port. Beyond this point the fresh charge is compressed and ignited near TDC. During the upward stroke, the negative pressure created in the crankcase below the piston draws in a fresh charge of fuel-air mixture through a one-way valve. On the downward power stroke, the mixture in the crankcase is pressurized. The piston uncovers the exhaust port and the high-pressure exhaust gases exit. Near BDC the transfer port is uncovered and the pressurized mixture flows from the crankcase into the cylinder and the cycle repeats. Since the crankcase is part of the induction system, it does not contain oil, and lubrication is accomplished by mixing oil with the fuel. With the cross-flow scavenging configuration illustrated in Figure 8.3.3, there will be a certain degree of mixing of the fresh charge with the combustion products remaining in the cylinder and some loss of fresh charge out the exhaust port.

Since two-stroke engines produce twice the power impulses of four-stroke engines for the same rpm, a two-stroke engine has a higher **power density** and is thus smaller and lighter than a four-stroke engine of equal output. The disadvantages of the two-stroke engine have historically been lower fuel efficiency and higher exhaust emissions because of the overlapping of the intake and exhaust processes and the

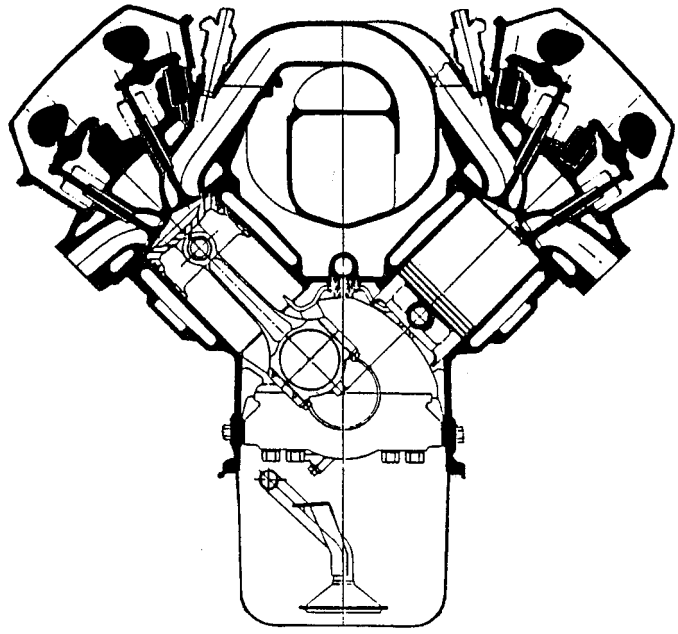


FIGURE 8.3.2 Ford 4.6-L aluminum V-8 SI engine. (Courtesy of Ford Motor Company.)

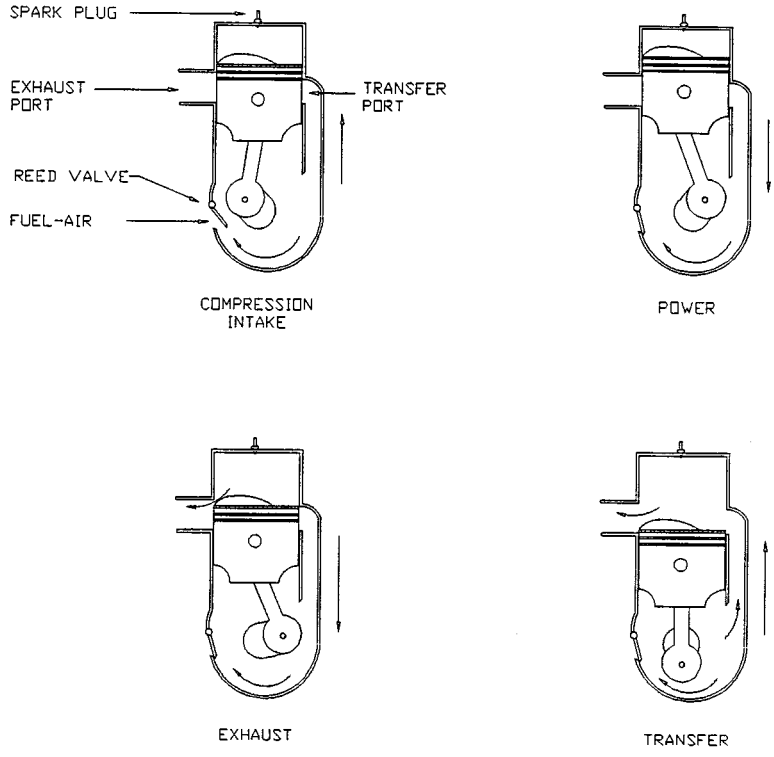


FIGURE 8.3.3 Schematic drawing of two-stroke SI engine.

loss of some fresh intake mixture with the exhaust products. For this reason, two-stroke SI engines have largely been confined to small-displacement applications, such as motorcycles, outboard marine engines, and small equipment. Several manufacturers have addressed these shortcomings in recent years and have achieved significant improvements in two-stroke engine fuel economy and emissions (Blair, 1988). The Orbital OCP (Orbital Combustion Process) engine, illustrated in Figure 8.3.4, is a modern two-stroke engine that utilizes direct injection of fuel into the cylinder in conjunction with a high-turbulence combustion chamber design and an electronically controlled exhaust port scavenging control valve to achieve very favorable fuel economy and significantly reduced levels of NO_x and hydrocarbon emissions. This engine, in three and six cylinder versions, is currently being used in automotive and marine applications.

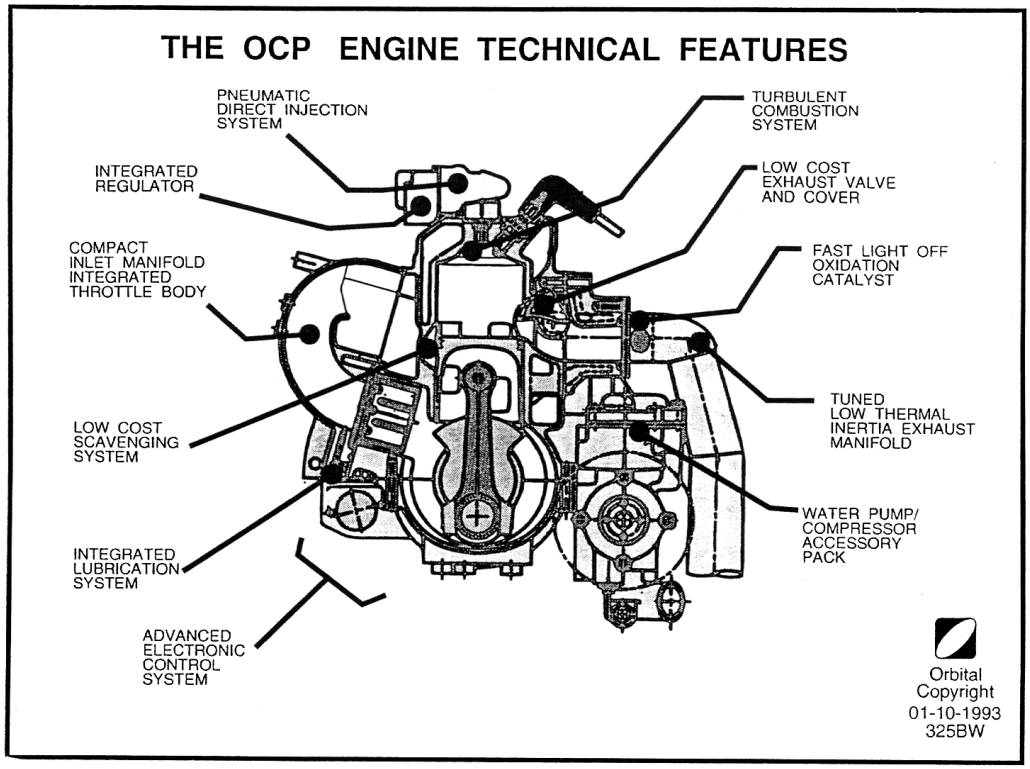


FIGURE 8.3.4 Orbital OCP two-stroke SI engine. (Courtesy of Orbital Engine Company.)

Compression Ignition Engine

The basic valve and piston motions are the same for the CI, or diesel, engine as discussed above for the SI engine. The CI engine relies on the high temperature and pressure of the cylinder air resulting from the compression process to cause **autoignition** of the fuel, which is injected directly into the combustion chamber of **direct injection (DI)** engines or into the prechamber of **indirect injection (IDI)** engines, when the piston approaches TDC on the compression stroke. The compression ratios are typically much higher for CI than for SI engines to achieve the high air temperatures required for autoignition, and the fuels used must have favorable autoignition qualities.

The time period between the start of fuel injection and the occurrence of autoignition is called the **ignition delay period**. Long ignition delay periods allow more time for fuel vaporization and fuel-air mixing, resulting in objectionable diesel knock when this larger premixed charge ignites. Combustion chambers and fuel injection systems must be designed to avoid extended ignition delay periods. Diesel

engines may be classified as DI or IDI. In DI engines, the combustion chamber consists of a bowl formed in the top of the piston and the fuel is injected into this volume. The injector tip generally has from four to eight holes to form multiple spray cones. Two variations are illustrated in Figure 8.3.5. The quiescent chamber engine utilizes a large-diameter shallow bowl shape that produces low **swirl** and low turbulence of the air during compression. Fuel is injected at high pressure through a multihole nozzle, and mixing of the fuel and air relies primarily on the energy of the injected fuel to cause air entrainment in the spray cone and diffusion of vaporized fuel into the air. This system is suited to large slow-speed engines that are operated with significant excess air.

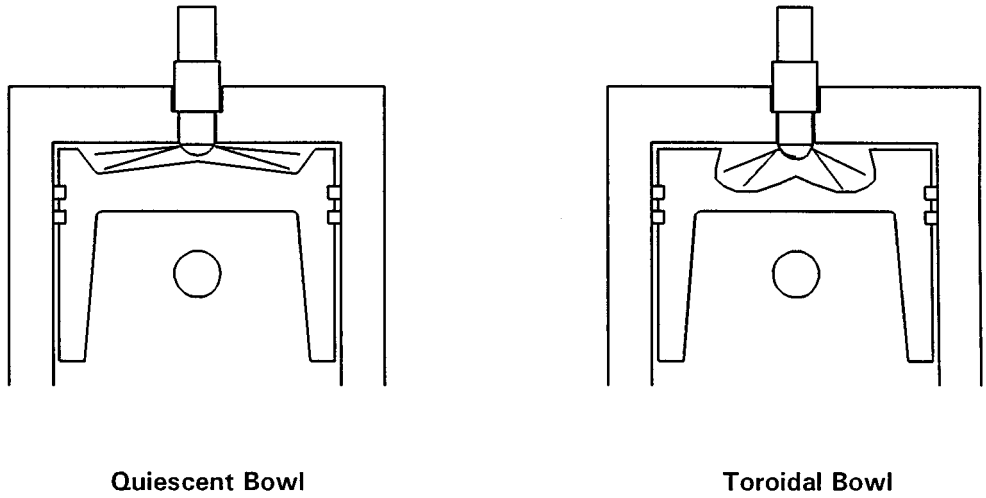


FIGURE 8.3.5 Examples of DI diesel combustion chamber design.

The toroidal bowl combustion chamber is used in conjunction with intake ports and/or valve shrouds designed to produce air swirl to enhance fuel-air mixing. The **swirl ratio** is defined by swirl ratio = swirl speed (rpm)/engine speed (rpm). The swirl velocity component is normal to the fuel spray direction and tends to promote mixing in the regions between the individual spray cones. This system makes better use of the available air and is utilized extensively in moderate-speed engines such as over-the-road truck engines. DI does not lend itself well to high-speed operation, as the time available for proper mixing and combustion is less. Diesel engines for passenger car applications are generally designed for higher-speed operation to produce higher specific output, and typically utilize IDI combustion systems, two of which are illustrated in Figure 8.3.6.

IDI systems make use of small prechambers incorporated in the cylinder head to promote rapid mixing of fuel and air and shorten the ignition delay period. Swirl chambers are designed to produce a strong vortex in the prechamber during compression. The fuel is sprayed into the chamber through a single-hole nozzle and the high vorticity promotes rapid mixing and short ignition delay periods. Precombustion chambers do not attempt to generate an orderly vortex motion within the chamber, but instead rely on a high level of turbulence, created by the rush of air into the chamber during compression, to promote mixing. Both types of prechambers generally include a lining of low-conductivity material (ceramic) to increase the surface temperature to promote further fuel evaporation. Prechambers can be used in small-displacement diesel engines to achieve operating speeds up to 5000 rpm. Disadvantages of the IDI system include poor cold-start characteristics due to high heat-transfer rates from the compressed air to the chamber wall that result from the high velocities and turbulence levels in the chamber. **Glow plugs** are often installed in each prechamber to heat the air to improve cold starting. Higher compression ratios are also used for IDI engines to further improve cold starting. The compression ratios, typically 18 to 24, are higher than the optimum for fuel efficiency (due to decreased mechanical efficiency), and IDI

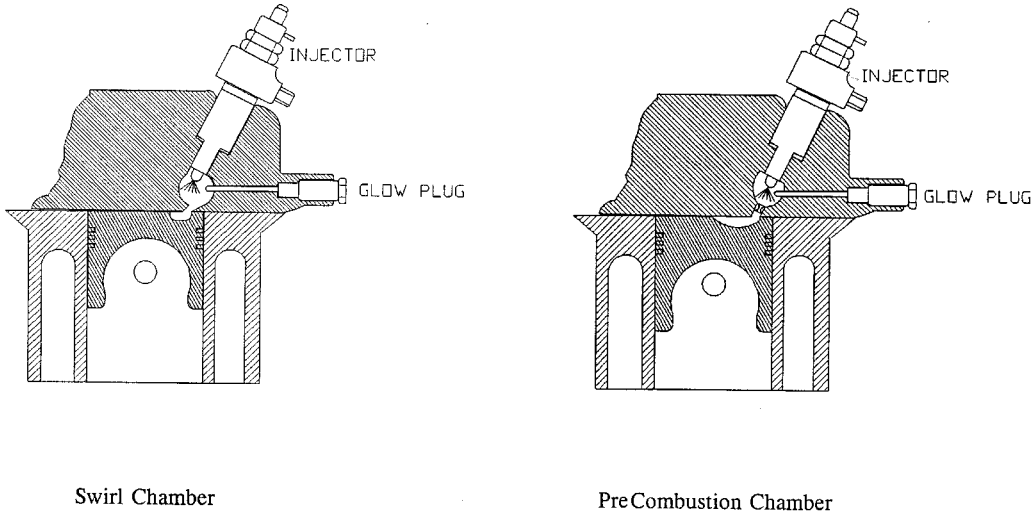


FIGURE 8.3.6 Two examples of IDI combustion chambers.

engines are typically less efficient than larger, slower, DI engines. The use of IDI is generally restricted to high-speed automotive engines, with displacements in the range of 0.3 to 0.8 liter per cylinder, and some degree of fuel economy is sacrificed in the interest of improved driveability.

CI engines are produced in both two-stroke and four-stroke versions. Since the fuel is injected directly into the combustion chamber of CI engines just prior to TDC, two-stroke CI engines do not suffer the same emission and efficiency shortcomings as do two-stroke SI engines. Hence, they are available in much larger displacements for high-power-requirement applications such as locomotive and ship propulsion and electric power generation systems. Two-stroke CI engines are generally of the DI type as the use of IDI in a two-stroke engine would lead to aggravated cold-start problems.

Air Standard Power Cycles

The actual operation of IC engines is idealized at a very basic level by the air standard power cycles (ideal thermodynamic models for converting heat into work on a continuous basis). The following simplifying assumptions are common to the air standard cycles: (1) the working substance is air, (2) the air is assumed to behave as an ideal gas with constant specific heats, (3) heat is added to the cycle from an external source, and (4) expansion and compression processes not involving heat transfer occur isentropically. The air standard cycles, while grossly oversimplified in terms of the complex processes occurring within actual engines, are nevertheless useful in understanding some fundamental principles of SI and CI engines. The simplified models also lend insight into important design parameters, e.g., **compression ratio**, that govern theoretical maximum cycle thermal efficiencies.

Constant-Volume Heat Addition — Ideal Otto Cycle

The theory of operation of the SI engine is idealized by the Otto cycle which assumes that heat is added to the system at constant volume. Constant-volume heat addition is approximated in the SI engine by virtue of the combustion process taking place rapidly when the piston is near TDC. A *P-V* diagram for the Otto cycle is illustrated in Figure 8.3.7. The cycle consists of the following processes: 1 → 2 isentropic compression, 2 → 3 constant-volume heat addition, 3 → 4 isentropic expansion, and 4 → 1 constant-volume heat rejection. The constant-volume heat rejection process is approximated in SI engines by the exhaust valve opening near BDC and the rapid blow down of exhaust gases.

Thermal efficiency for a power cycle is defined as the ratio of work output to heat input per cycle,

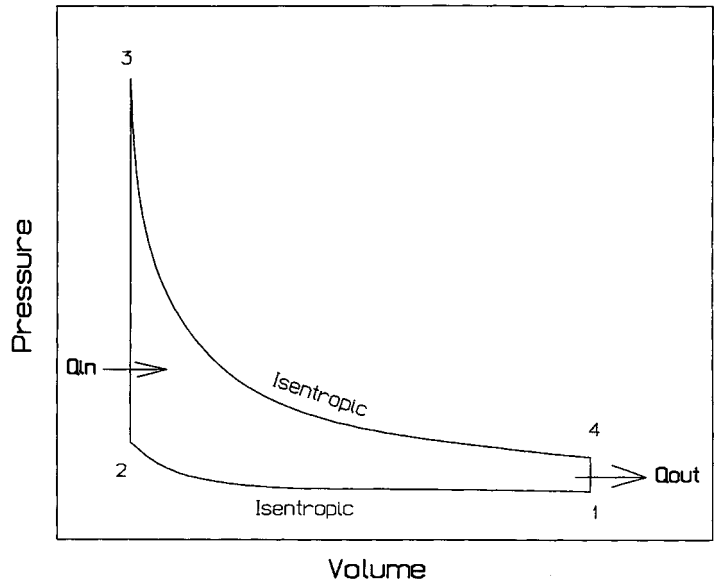


FIGURE 8.3.7 Schematic pressure-volume diagram for the ideal Otto cycle.

$$\eta = \frac{W_{out}}{Q_{in}} \tag{8.3.1}$$

For the Otto cycle, the basic efficiency expression can be manipulated into the form

$$\eta = 1 - \frac{1}{r^{\gamma - 1}} \tag{8.3.2}$$

where γ is the ratio of specific heats ($\gamma = C_p/C_v$) and r is the compression ratio, or ratio of the maximum to minimum cycle volumes ($r = V_1/V_2$). In actual IC engines, the minimum cycle volume is referred to as the **clearance volume** and the maximum cycle volume is the **cylinder volume**. The ideal Otto cycle efficiency for air, with $\gamma = 1.4$, is shown plotted in Figure 8.3.8. The theoretical efficiency of the constant volume heat addition cycle increases rapidly with compression ratio, up to about $r = 8$. Further increases in compression ratio bring moderate gains in efficiency. Compression ratios in practical SI engines are limited because of autoignition (knock) and high NO_x emission problems that accompany high compression ratios. Production SI automotive engines typically have compression ratios in the range 8 to 10, whereas high-performance normally aspirated racing engines may have compression ratios as high as 14, but they require the use of special fuels to avoid autoignition.

Constant-Pressure Heat Addition — Ideal Diesel Cycle

The air standard diesel cycle is the idealized cycle underlying the operation of CI or diesel engines. The diesel cycle, illustrated by the *P-V* diagram in Figure 8.3.9, consists of the following processes: 1 → 2 isentropic compression from the maximum to the minimum cycle volume, 2 → 3 constant-pressure heat addition during an accompanying increase in volume to V_3, 3 → 4 isentropic expansion to the maximum cycle volume, and 4 → 1 constant-volume heat rejection.

Actual diesel engines approximate constant-volume heat addition by injecting fuel for a finite duration which continues to burn and release heat at a rate that tends to maintain the pressure in the cylinder over a period of time during the expansion stroke. The efficiency of the ideal diesel cycle is given by

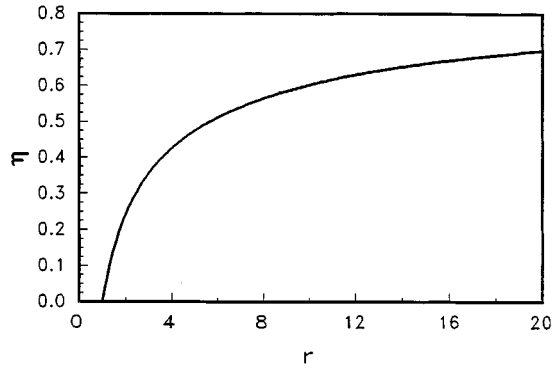


FIGURE 8.3.8 Efficiency of the ideal Otto cycle.

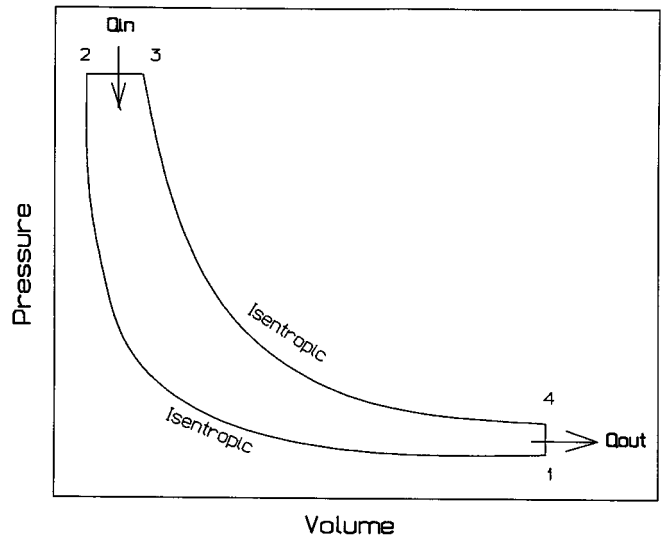


FIGURE 8.3.9 Schematic pressure-volume diagram of ideal diesel cycle.

$$\eta = 1 - \frac{1}{r^{\gamma-1}}\left[\frac{r_c^{\gamma} - 1}{\gamma\left(r_c - 1\right)}\right] \tag{8.3.3}$$

The efficiency of the ideal diesel cycle depends not only on the compression ratio, r, but also on the **cut-off ratio**, $r_c = V_3/V_2$, the ratio of the volume when heat addition ends to the volume when it begins. Equation (8.3.3) is shown plotted in Figure 8.3.10 for several values of r_c and for $\gamma = 1.4$. An r_c value of 1 is equivalent to constant-volume heat addition, i.e., the Otto cycle. The efficiency of the ideal Diesel cycle is less than the efficiency of the ideal Otto cycle for any given compression ratio and any value of the cut-off ratio greater than 1. The fact that CI engines, by design, operate at much higher compression ratios than SI engines (generally between 12 and 24) accounts for their typically higher operating efficiencies relative to SI engines.

Actual Cycles

IC engines do not operate on closed thermodynamic cycles, such as the air standard power cycles, but rather on open mechanical cycles, and heat addition occurs neither at constant volume or constant

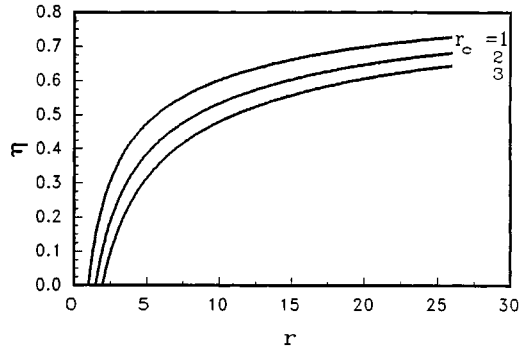


FIGURE 8.3.10 Efficiency of the ideal diesel cycle.

pressure. Figure 8.3.11 is a schematic representation of an **indicator diagram** (pressure-volume history) of a four-stroke IC engine; it could be either SI or CI. The pressure changes during the intake and exhaust strokes are exaggerated in the diagram. The **indicated work** performed per cycle can be calculated by taking the integral of *PdV* for the complete cycle. The **indicated mean effective pressure**, **imep**, is defined as the ratio of the net indicated work output to the **displacement volume**:

$$\text{imep} = \frac{\text{indicated work output per cycle}}{\text{displacement volume}} \tag{8.3.4}$$

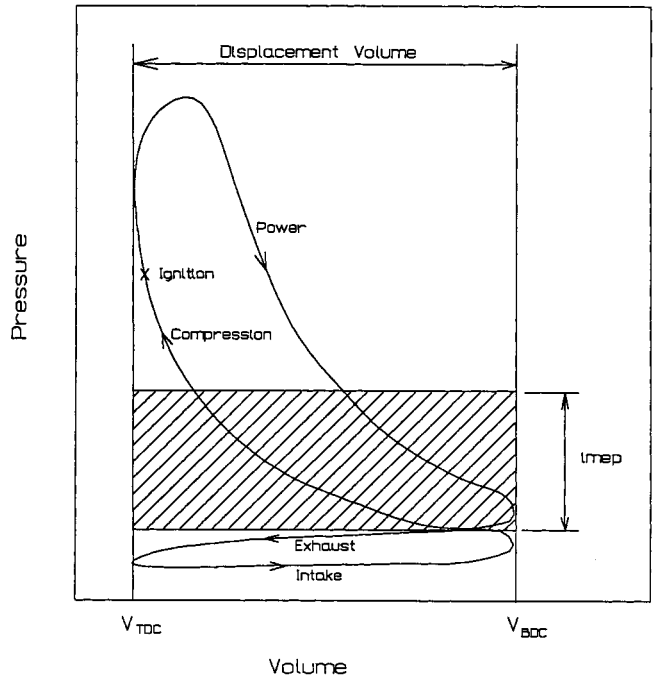


FIGURE 8.3.11 Schematic indicator diagram.

The shaded area in Figure 8.3.11 thus represents the net indicated work output per cycle. During intake and exhaust, the negative work performed represents pumping losses and acts to decrease the net work output of the engine. The magnitude of the pumping losses depends on the flow characteristics of the intake and exhaust systems including the valves, ports, manifolds, piping, mufflers, etc. The more

restrictive these passages, the higher will be the pumping losses. SI engines control power output by throttling the intake air. Thus, under partial-load conditions, the pressure drop resulting from the air throttling represents a significant increase in pumping loss with a corresponding decrease in operating efficiency. SI engines are therefore less efficient at partial-load operation than at full load. The power level of CI engines, on the other hand, is controlled by varying the amount of fuel injected, as opposed to throttling the intake air, making them significantly more efficient than SI engines under partial-load conditions.

Brake work (or power) is the actual work (or power) produced at the output shaft of an engine, as measured by a dynamometer. The brake work will be less than the indicated work due to friction losses and any parasitic power requirements for oil pumps, water pumps, etc. The **brake mean effective pressure**, **bmep**, is defined as

$$\text{bmep} = \frac{\text{brake work output per cycle}}{\text{displacement volume}} \tag{8.3.5}$$

The mechanical efficiency can then be defined as

$$\eta_m = \frac{\text{brake work (power)}}{\text{indicated work (power)}} = \frac{\text{bmep}}{\text{imep}} \tag{8.3.6}$$

Engine thermal efficiency can be determined from the ratio of power output to rate of fuel energy input, or

$$\eta_t = \frac{\text{Power}}{m_f Q_c} \tag{8.3.7}$$

where m_f is the rate of fuel consumption per unit time and Q_c is the heat of combustion per unit mass of fuel. The thermal efficiency in Equation (8.3.7) could be either indicated or brake depending on the nature of the power used in the calculation. Uncertainty associated with variations of energy content of fuels may present a practical difficulty with determining engine thermal efficiency. In lieu of thermal efficiency, **brake specific fuel consumption** (**bsfc**), is often used as an efficiency index.

$$\text{bsfc} = \frac{\text{fuel consumption rate (kg/hr)}}{\text{brake power (kW)}} \tag{8.3.8}$$

The efficiency of engines operating on the same fuel may be directly compared by their bsfc.

Volumetric efficiency, η_v, is an important performance parameter for four-stroke engines defined as

$$\eta_v = \frac{m_{\text{actual}}}{m_d} \tag{8.3.9}$$

where m_{actual} is the mass of intake mixture per cycle at inlet conditions (pressure and temperature near the inlet port) and m_d is the mass of mixture contained in the displacement volume at inlet conditions. For SI engines the mixture mass includes both air and fuel; for CI engines only air is present during intake. With the intake mixture density determined at inlet conditions, η_v accounts for flow losses associated with the intake ports, valves, and cylinder. Sometimes, for convenience, the mixture density is taken at ambient conditions. In this case, η_v is called the overall volumetric efficiency and includes the flow performance of the entire intake system. Since a certain minimum amount of air is required for complete combustion of a given amount of fuel, it follows that the maximum power output of an engine

is directly proportional to its air-flow capacity. Thus, while not affecting in any way the thermal efficiency of the engine, the volumetric efficiency directly affects the maximum power output for a given displacement, and thus can affect the efficiency of the overall system in which the engine is installed because of the effect on system size and weight. Volumetric efficiency is affected primarily by intake and exhaust valve geometry, valve lift and timing, intake port and manifold design, mixing of intake charge with residual exhaust gases, engine speed, ratio of inlet pressure to exhaust back pressure, and heat transfer to the intake mixture from warmer flow passages and combustion chamber surfaces. For further information on the fundamentals of IC engine design and operation see Taylor (1985), Ferguson (1986), Heywood (1988), and Stone (1993).

Combustion in IC Engines

Combustion in Spark Ignition Engines

Background. In SI engines, combustion of the fuel-air mixture is initiated by a spark generated between the electrodes of a spark plug. The intake and compression strokes are designed to prepare the mixture for combustion by completely vaporizing the fuel and heating the mixture to just below its autoignition temperature. This is one reason, in addition to controlling emissions, for the current practice of limiting the maximum compression ratio of SI engines to about 10:1. Near the end of compression, the mixture is well conditioned for combustion and the spark is discharged to initiate the combustion process. For best fuel economy, the combustion process must be completed as close as possible to TDC. This requires that the spark timing be controlled for varying operating speed and load conditions of the engine. Fuel metering and control, according to the engine load requirements, and with minimum variation from cylinder to cylinder and cycle to cycle, is essential for good fuel economy, power output, and emission control of the engine.

Both carburetors and fuel injection systems are used for fuel-metering control. Because of the superior control capabilities of fuel injection systems, they are nearly universally used today in production automotive applications. Carburetors are used for applications with less-stringent emission requirements, e.g., small engines for lawn and garden equipment. Figure 8.3.12 illustrates the effect of **fuel-air ratio** on the indicated performance of an SI engine. The **equivalence ratio** (γ) is defined by the ratio Fuel-Air_{actual}/Fuel-$Air_{stoichiometric}$. Rich mixtures have fuel-air ratios greater than stoichiometric ($\gamma > 1$) and lean mixtures have fuel-air ratios less than stoichiometric ($\gamma < 1$). Optimum fuel economy, coinciding with maximum thermal efficiency, is obtained at part throttle with a lean mixture as a result of the fact that the heat release from lean mixtures suffers minimal losses from dissociation and variation of specific heat effects when compared with stoichiometric and rich fuel-air ratios. Maximum power is obtained at full throttle with a slightly rich mixture, an indication of the full utilization of the air inside the cylinders. Idling, with a nearly closed throttle, requires a rich mixture because of the high percentage of exhaust gas residuals that remains in the cylinders. The fuel-air mixture requirement under transient operation, such as acceleration, requires a rich mixture to compensate for the reduced evaporation caused by the sudden opening of the throttle. Cold starting also requires a rich mixture to ensure the vaporization of sufficient amounts of the highly volatile components in the fuel to achieve proper ignition.

Normal Combustion Process. The combustion processes in SI engines can be divided into two categories, normal and abnormal. The normal combustion process occurs in three stages: initiation of combustion, flame propagation, and termination of combustion. Combustion normally starts across the spark plug gap when the spark is discharged. The fuel molecules in and around the spark discharge zone are ignited and a small amount of energy is released. The important criterion for the initial reaction to be self-sustaining is that the rate of heat release from the initial combustion be larger than the rate of heat transfer to the surroundings. The factors that play an important role in making the initial reaction self-sustaining, and thereby establishing a flame kernel, are the ignition energy level, the spark plug gap, the fuel-air ratio, the initial turbulence, and the condition of the spark plug electrodes.

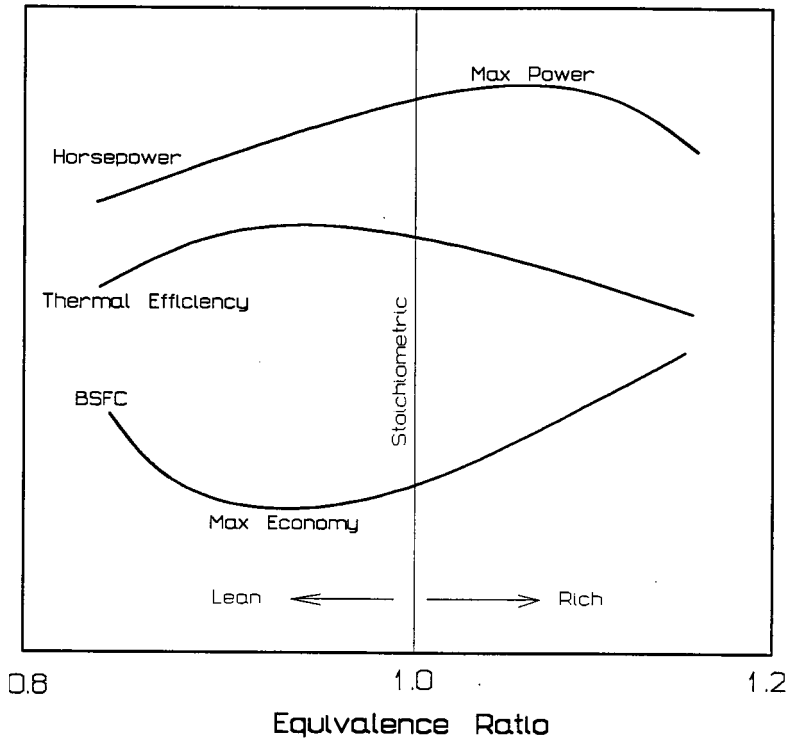


FIGURE 8.3.12 Effect of fuel-air mixture on indicated performance of an SI engine.

After a flame kernel is established, a thin spherical flame front advances from the spark plug region progressively into the unburned mixture zone. Flame propagation is supported and accelerated by two processes. First, the combined effect of the heat transfer from the high-temperature flame region and the bombardment by the active radicals from the flame front into the adjacent unburned zone raises the temperature and accelerates the rate of reactivity of the unburned mixture region directly adjacent to the flame front. This helps condition and prepare this zone for combustion. Second, the increase in the temperature and pressure of the burned gases behind the flame front will cause it to expand and progressively create thermal compression of the remaining unburned mixture ahead of the flame front. It is expected that the flame speed will be low at the start of combustion, reach a maximum at about half the flame travel, and decrease near the end of combustion. Overall, the flame speed is strongly influenced by the degree of turbulence in the combustion chamber, the shape of the combustion chamber, the mixture strength, the type of fuel, and the engine speed.

When the flame front approaches the walls of the combustion chamber, the high rate of heat transfer to the walls slows down the flame propagation and finally the combustion process terminates close to the walls because of surface quenching. This leaves a thin layer of unburned fuel close to the combustion chamber walls which shows up in the exhaust as unburned hydrocarbons.

Abnormal Combustion. Abnormal combustion may occur in SI engines associated with two combustion phenomena: **knock** and **surface ignition**. Knock occurs near the end of the combustion process if the end portion of the unburned mixture, which is being progressively subjected to thermal compression, autoignites prematurely before the flame front reaches it. As a result of the sudden energy release, a violent pressure wave propagates back and forth across the combustion chamber, causing the walls or other parts of the engine to vibrate, producing a sharp metallic noise called knock. If knock persists for a period of time, the high rate of heat transfer caused by the traveling high pressure and temperature wave may overheat the spark plug electrode or ignite carbon deposits that may be present in the

combustion chamber, causing uncontrolled combustion and preignition. As a result, loss of power and serious engine damage may occur.

Knock is sensitive to factors that increase the temperature and pressure of the end portion of the unburned mixture, as well as to fuel composition and other time factors. Factors that increase the probability of knock include (1) increasing the temperature of the mixture by increasing the charge intake temperature, increasing the compression ratio, or turbo/supercharging; (2) Increasing the density of the mixture by turbo/supercharging or increasing the load; (3) advancing the spark timing; (4) increasing the time of exposure of the end portion of the unburned mixture to autoignition conditions by increasing the length of flame travel or decreasing the engine speed and turbulence; and (5) using low-octane fuel and/or maximum power fuel-air ratios. Engine design factors that affect knock in SI engines include the shape of the combustion chamber and the location of the spark plug and inlet and exhaust valves relative to the location of the end portion of the unburned mixture.

Surface ignition is the ignition of the unburned mixture by any source in the combustion chamber other than the normal spark. Such sources could include overheated exhaust valves or spark plug electrodes, glowing carbon deposits, or other hot spots. Surface ignition will create secondary flame fronts which cause high rates of pressure rise resulting in a low-pitched, thudding noise accompanied by engine roughness. Severe surface ignition, especially when it occurs before spark ignition, may cause serious structural and/or component damage to the engine.

Combustion in Compression Ignition Engines

Unlike the SI engine, in which the charge is prepared for combustion as a homogeneous mixture during the intake and compression strokes, fuel preparation for combustion in CI engines occurs in a very short period of time called the ignition delay period. During this period, the fuel injected into the high-temperature air near the end of the compression stroke undergoes two phases of transformation. A physical delay period, during which the fuel is vaporized, mixed with the air, and raised in temperature, is followed by a chemical delay period during which fuel cracking and decomposition occur which leads to autoignition and combustion of the fuel.

The combustion process is **heterogeneous** and involves two modes, usually identified as premixed combustion and diffusion combustion. Premixed combustion occurs early in the process when the fuel which has evaporated and mixed with air during the ignition delay period ignites. This mode is characterized by uncontrolled combustion and is the source of combustion noise since it is accompanied by a high rate of heat release which produces a high rate of pressure rise. When the premixed fuel-air mixture is depleted, diffusion combustion takes over, characterized by a lower rate of heat release and producing controlled combustion during the remainder of the process. Figure 8.3.13 depicts the different stages of the combustion process in CI engines.

The ignition delay period plays a key role in controlling the time duration of the two modes of combustion. Prolonging the ignition delay period, either through engine design factors or variations in operating conditions, will generate a larger portion of premixed fuel-air mixture and will thus tend to increase the premixed combustion mode duration and decrease the diffusion mode duration. This may lead to higher peak cylinder pressure and temperature which may improve thermal efficiency and reduce CO and **unburned hydrocarbon (UHC)** emissions at the expense of increased emissions of oxides of nitrogen ($\mathbf{NO_x}$). Large increases in the ignition delay period will cause high rates of pressure rise during the premixed combustion and may lead to objectionable diesel knock. Reducing the ignition delay period causes the premixed combustion duration to decrease while increasing the diffusion combustion duration. A large reduction in ignition delay may lead to loss of power, decrease in thermal efficiency, and possible deterioration of exhaust emissions. Several factors related to the fuel-air mixture temperature and density, engine speed, combustion chamber turbulence, injection pressure, rate of injection, and fuel composition influence the duration of the ignition delay period.

Knock in CI Engines. As the combustion process in CI engines is triggered by autoignition of the fuel injected during the ignition delay period, factors that prolong the ignition delay period will increase the

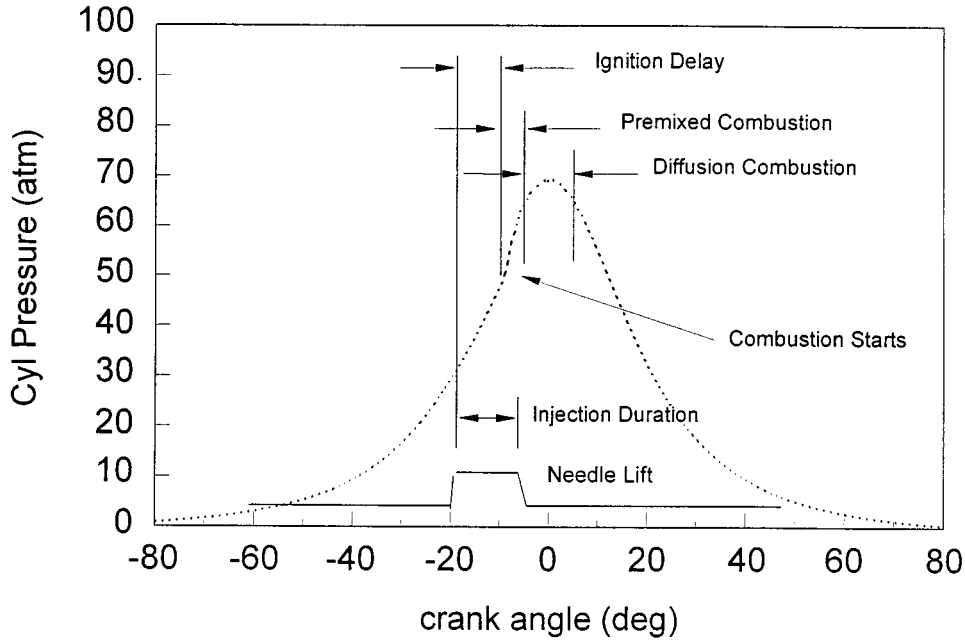


FIGURE 8.3.13 Combustion process in a CI engine.

premixed combustion duration causing very high rates of energy release and thus high rates of pressure rise. As a result, diesel knock may occur. The phenomenon is similar to knock in SI engines except that it occurs at the beginning of the combustion process rather than near the end, as observed in SI combustion. Factors that reduce the ignition delay period will reduce the possibility of knock in diesel engines. Among them are increasing the compression ratio, supercharging, increasing combustion chamber turbulence, increasing injection pressure, and using high-**cetane-number** (**CN**) fuel. For a more detailed discussion of the combustion process in IC engines, see Henein (1972), Lenz (1992), and Keating (1993).

Exhaust Emissions

Harmful Constituents

The products of combustion from IC engines contain several constituents that are considered hazardous to human health, including CO, UHCs NO_x, and **particulates** (from diesel engines). These emission products are discussed briefly below followed by a description of the principal schemes for their reduction.

Carbon Monoxide. CO is a colorless, odorless, and tasteless gas that is highly toxic to humans. Breathing air with a small volumetric concentration (0.3%) of CO in an enclosed space can cause death in a short period of time. CO results from the incomplete combustion of hydrocarbon fuels. One of the main sources of CO production in SI engines is the incomplete combustion of the rich fuel mixture that is present during idling and maximum power steady state conditions and during such transient conditions as cold starting, warm-up, and acceleration. Fuel maldistribution, poor condition of the ignition system, very lean combustion, and slow CO reaction kinetics also contribute to increased CO production in SI engines. CO production is not as significant in CI engines since these engines are always operated with significant excess air.

Unburned Hydrocarbons. When UHCs combine with NO_x (see below) in the presence of sunlight, ozone and photochemical oxidants form that can adversely affect human health. Certain UHCs are also

considered to be carcinogenic. The principal cause of UHC in SI engines is incomplete combustion of the fuel-air charge, resulting in part from flame quenching of the combustion process at the combustion chamber walls, and engine misfiring. Additional sources in four-stroke engines may include fuel mixture trapped in crevices of the top ring land of the piston and outgassed fuel during the expansion (power) stroke that was absorbed into the lubricating oil film during intake. In two-stroke SI engines, the scavenging process often results in a portion of the fresh mixture exiting the exhaust port before it closes, resulting in large UHC emissions.

The presence of UHC in CI engines is related to the heterogeneous nature of the fuel-air mixture. Under certain conditions, fuel-air mixtures that lie outside the flammability limits at both the lean and rich extremes can exist in portions of the combustion chamber and escape combustion, thus contributing significantly to UHC in the exhaust. Fuel injected near the end of the combustion process, and fuel remaining in the nozzle **sac volume** at the end of injection, both contribute to UHC emission in CI engines. Engine variables that affect UHC emissions include the fuel-air ratio, intake air temperature, and cooling water temperature.

Oxides of Nitrogen. Nitric oxide (NO) is formed from the combination of nitrogen and oxygen present in the intake air under the high-temperature conditions that result from the combustion process. As the gas temperature drops during the expansion stroke, the reaction is frozen, and levels of NO persist in the exhaust products far in excess of the equilibrium level at the exhaust temperature. In the presence of additional oxygen in the air, some NO transforms to nitrogen dioxide (NO_2), a toxic gas. The NO and NO_2 combined are referred to as oxides of nitrogen or NO_x. The production of NO_x is in general aggravated by conditions that increase the peak combustion temperature. In SI engines the most important variables that affect NO_x production are the air/fuel ratio, spark timing, intake air temperature, and amount of residual combustion products remaining in the cylinder after exhaust. In CI engines, ignition delay, which affects the degree of premixed combustion, plays a key role in NO_x formation. A larger premixed combustion fraction will produce higher combustion temperatures and higher levels of NO_x.

Particulates. Particulates are a troublesome constituent in the exhaust from CI engines. Particulates are defined by the US Environmental Protection Agency (EPA) as any exhaust substance (other than water) that can be trapped on a filter at a temperature of 325 K or below. Particulates trapped on a filter may be classified as soot plus an organic fraction of hydrocarbons and their partial oxidation products. Soot consists of agglomerates of solid uncombusted carbon particles. Particulates are of concern because their small size permits inhalation and entrapment in the lung walls, making them potential lung carcinogens.

Soot is formed in CI engines under conditions of heavy load when the gas temperature is high and the concentration of oxygen is low. Smoke production is affected by such parameters as fuel CN, rate of fuel injection, inlet air temperature, and the presence of secondary injection.

Control of Emissions from IC Engines

Figure 8.3.14 depicts the relative concentrations of CO, NO_x, and UHC in the exhaust products of an SI engine as a function of the fuel-air mixture. Lean mixture combustion, which promotes good thermal efficiency, also results in low UHC and CO production but causes high levels of NO_x emission. Increasing the fuel/air ratio to reduce NO_x results in increased CO and UHC emission. Approaches to reduce total emissions fall under two categories; the first concentrates on engine design and fuel modifications and the second involves treatment of exhaust gases after leaving the engine. In SI engines, the first approach focuses on addressing engine variables and design modifications which improve in-cylinder mixing and combustion in an effort to reduce CO and UHC emissions. To reduce NO_x, attention is focused on factors that reduce peak combustion temperature and reduce the oxygen available in the flame front. Design and operating parameters that have been implemented or modified for decreased emissions include compression ratio reduction, increased coolant temperature, modification of the combustion chamber shape to minimize surface-to-volume ratio and increase turbulence, improvement of intake manifold design for better charge distribution, use of fuel injection instead of carburetors for better mixture control,

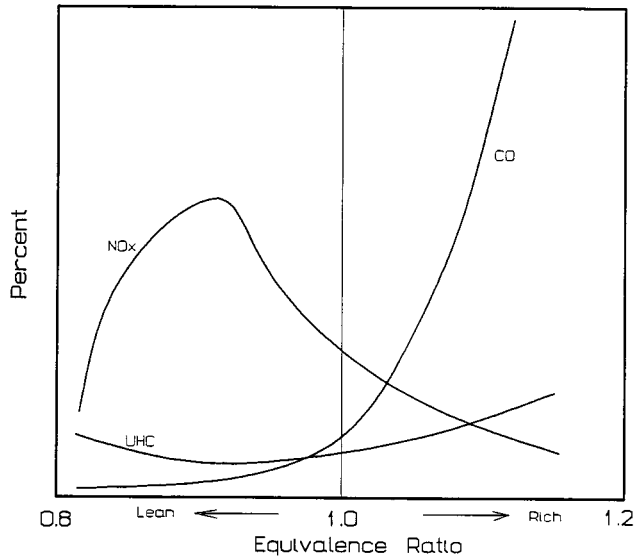


FIGURE 8.3.14 Emission levels from an SI engine vs. fuel-air mixture.

use of exhaust gas recirculation to reduce NO_x by lowering combustion temperatures, positive crankcase ventilation to reduce UHC, and increased aromatic content in gasoline.

Engine modifications that have been implemented to reduce emissions from CI engines include modifications to the combustion chamber shape to match the air swirl pattern and fuel spray pattern for better mixing and complete combustion, use of exhaust gas recirculation to limit NO_x production, use of higher injection pressure for better atomization to reduce soot and UHC, and the use of precise injection timing with electronic control.

In the second approach, several devices have been developed for after treatment of exhaust products. A thermal reactor may be used to oxidize UHC and CO. These typically consist of a well-insulated volume placed close to the exhaust manifold, with internal baffles to increase the gas residence time and an air pump to supply fresh oxygen for the oxidation reactions. Thermal reactors are ineffective for NO_x reduction and thus have limited application.

Catalytic converters utilize a catalyst, typically a noble metal such as platinum, rhodium, or palladium, deposited on a ceramic substrate to promote reactions at lower temperatures. Two types are in use, oxidation converters and reduction converters. Oxidation catalytic converters use the excess air available in lean mixtures (or supplied from an external air pump) to oxidize CO and UHC emissions. Reduction catalytic converters operate with low levels of oxygen to cause reduction of NO_x. Sometimes, dual catalytic converters are employed to treat all three pollutants with a reducing converter, to reduce NOx, placed upstream of an oxidation converter for treating CO and UHC. This arrangement requires that the engine be operated with a rich mixture which decreases fuel economy.

Three-way catalytic converters are a recent development that permits treatment of NO_x, CO, and UHC in a single device, thus reducing size and weight of the exhaust system. Proper operation of a three-way catalyst requires very nearly stoichiometric combustion. If the combustion is too lean, NO_x is not adequately reduced, and if it is too rich, UHC and CO are not adequately oxidized. There is a narrow band for equivalence ratio from about 0.999 to 1.007 within which conversion efficiency is 80% or better for all three pollutants (Kummer, 1980). Maintaining engine operation within this narrow mixture band requires a closed-loop fuel-metering system that utilizes an oxygen sensor placed in the exhaust system to monitor excess oxygen and control the fuel injection to maintain near stoichiometric combustion.

Reduction catalytic converters cannot be used with CI engines to reduce NO_x since they normally run lean with significant amounts of excess oxygen in the exhaust. Thus, engine design factors must be relied on to keep NO_x as low as possible. Soot emission may be reduced by after treatment using a

device called a trap oxidizer. A trap oxidizer filters particulate matter from the exhaust stream and oxidizes it, usually with the aid of a catalyst for reducing the oxidation temperature. They have been used on small, high-speed automotive diesel engines, but their application to larger, slower-speed engines is limited because of the higher level of particulate production and the lower exhaust temperature. For additional information on emissions see Henein (1972), Obert (1973), and *SAE Surface Vehicle Emissions Standards Manual* (1993).

Fuels for SI and CI Engines

Background

The primary distinguishing factor between SI and CI engines is the fundamental difference in the combustion process. SI engines rely on homogeneous, premixed combustion, while CI engines are designed for heterogeneous combustion with a premixed combustion period followed by a diffusion combustion period. The differences in the combustion process call for quite different qualities in the fuels to achieve optimum performance.

By far the most common fuel for SI engines is gasoline, although other fuels can be used in special circumstances including alcohol, natural gas, and propane. Even such low-grade fuels as wood gas and coal gas have been used to fuel SI engines during wartime when conventional fuels were in short supply. Diesel fuel is the predominant fuel for CI engines, but they too can be designed to operate on a variety of other fuels, such as natural gas, bio-gas, and even coal slurries. This discussion is confined to gasoline and diesel fuel, both of which are distilled from crude oil.

Crude oil is composed of several thousand different hydrocarbon compounds, which upon heating are vaporized at different temperatures. In the distillation process, different "fractions" of the original crude are separated according to the temperatures at which they vaporize. The more volatile fraction, naphtha, is followed in order of increasing temperature of vaporization by fractions called distillate, gas oil, reduced crude, and residual oil. These fractions may be further subdivided into light, middle, and heavy classifications. Light virgin naphtha can be used directly as gasoline, although it has relatively poor antiknock quality. The heavier fractions can be chemically processed through coking and catalytic cracking to produce additional gasoline. Diesel fuel is derived from the light to heavy virgin gas oil fraction and from further chemical processing of reduced crude.

Gasoline

Gasoline fuels are mixtures of hydrocarbon compounds with boiling points in the range of 32 to 215°C. The two most important properties of gasoline for SI engine performance are volatility and octane rating. Adequate volatility is required to ensure complete vaporization, as required for homogeneous combustion, and to avoid cold-start problems. If the volatility is too high, however, vapor locking in the fuel delivery system may become a problem. Volatility may be specified by the distillation curve (the distillation temperatures at which various percentages of the original sample have evaporated). Higher-volatility fuels will be characterized by lower temperatures for given fixed percentages of evaporated sample, or conversely, by higher percentages evaporated at or below a given temperature. Producers generally vary the volatility of gasoline to suit the season, increasing the volatility in winter to improve cold-start characteristics and decreasing it in summer to reduce vapor locking.

The octane rating of a fuel is a measure of its resistance to autoignition or knocking; higher-octane fuels are less prone to autoignition. The octane rating system assigns the value of 100 to iso-octane (C_8H_{18}, a fuel that is highly resistant to knock) and the value 0 to *n*-heptane (C_7H_{16}, a fuel that is prone to knock). Two standardized methods are employed to determine the octane rating of fuel test samples, the research method and the motor method; see ASTM Standards Part 47 — Test Methods for Rating Motor, Diesel and Aviation Fuels (ASTM, 1995). Both methods involve testing the fuel in a special variable-compression-ratio engine (cooperative fuels research or CFR engine). The test engine is operated on the fuel sample and the compression ratio is gradually increased to obtain a standard knock intensity reading from a knock meter. The octane rating is obtained from the volumetric percentage of iso-octane

in a blend of iso-octane and *n*-heptane that produces the same knock intensity at the same compression ratio. The principal differences between the research method and the motor method are the higher operating speed, higher mixture temperature, and greater spark advance employed in the motor method. Ratings obtained by the research method are referred to as the **research octane number** (RON), while those obtained with the motor method are called the motor octane number (MON). MON ratings are lower than RON ratings because of the more stringent conditions, i.e., higher thermal loading of the fuel. The octane rating commonly advertised on gasoline pumps is the **antiknock index**, $(R + M)/2$, which is the average of the values obtained by the two methods. The typical range of antiknock index for automotive gasolines currently available at the pump is 87 to 93. In general, higher compression SI engines require higher-octane fuels to avoid autoignition and to realize full engine performance potential from engines equipped with electronic control systems incorporating a knock sensor.

Straight-run gasoline (naphtha) has a poor octane rating on the order of 40 to 50 RON. Higher-octane fuels are created at the refinery by blending with higher-octane components produced through alkylation wherein light olefin gases are reacted with isobutane in the presence of a catalyst. Iso-octane, for example, is formed by reacting isobutane with butene. Aromatics with double carbon bonds shared between more than one ring, such as naphthalene and anthracene, serve to increase octane rating because the molecules are particularly difficult to break.

Additives are also used to increase octane ratings. In the past, a common octane booster added to automotive fuels was lead alkyls, either tetraethyl or tetramethyl lead. For environmental reasons, lead has been removed from automotive fuels in most countries. It is, however, still used in aviation fuel. Low-lead fuel has a concentration of about 0.5 g/L which boosts octane rating by about five points. The use of leaded fuel in an engine equipped with a catalytic converter to reduce exhaust emissions will rapidly deactivate the catalyst (typically a noble metal such as platinum or rhodium), quickly destroying the utility of the catalytic converter. Octane-boosting additives that are in current use include the oxygenators methanol, ethanol, and methyl tertiary butyl ether (MTBE).

RON values of special-purpose high-octane fuels for racing and aviation purposes can exceed 100 and are arrived at through an extrapolation procedure based on the knock-limited indicated mean effective pressure (klimep). The klimep is determined by increasing the engine intake pressure until knock occurs. The ratio of the klimep of the test fuel to that for iso-octane is used to extrapolate the octane rating above 100.

Diesel Fuels

Diesel fuels are blends of hydrocarbon compounds with boiling points in the range of 180 to 360°C. Properties that are of primary importance for CI fuels include the density, viscosity, cloud point, and ignition quality (CN). Diesel fuel exhibits a much wider range of variation in properties than does gasoline. The density of diesel fuels tends to vary according to the percentages of various fractions used in the blend. Fractions with higher distillation temperatures tend to increase the density. Variations in density result in variations in volumetric energy content and hence fuel economy, since fuel is sold by volume measure. Higher-density fuel will also result in increased soot emission. Viscosity is important to proper fuel pump lubrication. Low-viscosity fuel will tend to cause premature wear in injection pumps. Too high viscosity, on the other hand, may create flow problems in the fuel delivery system. Cloud point is the temperature at which a cloud of wax crystals begins to form in the fuel. This property is critical for cold-temperature operation since wax crystals will clog the filtration system. ASTM does not specify maximum cloud point temperatures, but rather recommends that cloud points be no more than 6°C above the tenth percentile minimum ambient temperature for the region for which the fuel is intended; see ASTM D 975 (ASTM 1995).

CN provides a measure of the autoignition quality of the fuel and is the most important property for CI engine fuels. The CN of a fuel sample is obtained through the use of a CI CFR engine in a manner analogous to the determination of octane rating. The test method for CN determination is specified in standard ASTM D 613. *n*-Cetane (same as hexadecane, $C_{16}H_{34}$) has good autoignition characteristics and is assigned the cetane value of 100. The bottom of the cetane scale was originally defined in terms

of α-methyl naphthalene ($C_{11}H_{10}$) which has poor autoignition characteristics and was assigned the value 0. In 1962, for reasons of availability and storability, the poor-ignition-quality standard fuel used to establish the low end of the cetane scale was changed to heptamethylnonane (HMN), with an assigned CN of 15. The CN of a fuel sample is determined from the relative volumetric percentages of cetane and HMN in a mixture that exhibits the same ignition delay characteristics as the test sample using the relation

$$\text{CN} = \%\ n\text{-cetane} + 0.15\,(\%\ \text{HMN}) \tag{8.3.10}$$

ASTM standard D 976 (ASTM, 1995) provides the following empirical correlation for calculating the **cetane index** of straight petroleum distillate fuels (no additives) as an approximation to the measured CN:

$$\text{Cetane Index} = 454.74 - 1641.416D + 774.74D^2 - 0.554B + 97.803\,(\log B)^2 \tag{8.3.11}$$

where D = density at 15°C (g/mL) and B = mid-boiling temperature (°C).

ASTM standard D 975 (ASTM, 1995) establishes three classification grades for diesel fuels (No. 1-D, No. 2-D, and No. 4-D) and specifies minimum property standards for these grades. No. 1-D is a volatile distillate fuel for engines that must operate with frequent changes in speed and load. No. 2-D is a lower-volatility distillate fuel for industrial and heavy mobile service engines. No. 4-D is a heavy fuel oil for low- and medium-speed engines. Nos. 1-D and 2-D are principally transportation fuels, while No. 4-D is for stationary applications. The ASTM minimum CN for No. 1-D and No. 2-D is 40, and for No. 4-D the minimum is 30. Typical CNs for transportation fuels lie in the range 40 to 55. Use of a low-cetane fuel aggravates diesel knock because of the longer ignition delay period which creates a higher fraction of premixed combustion.

Antiknock quality (octane number) and ignition quality (CN) are opposing properties of distillate fuels. The CN increases with decreasing octane rating of various fuels. Gasoline, with good antiknock quality, has a CN of approximately 10, while a diesel fuel with a CN of 50 will have an octane number of about 20. Thus, gasoline is not a suitable fuel for CI engines because of its poor autoignition quality, and diesel fuel is inappropriate for use in SI engines as a result of its poor antiknock quality. For additional information on fuels for IC engines see Owen and Coley (1995) and the *SAE Fuels and Lubricants Standards Manual* (1993).

Intake Pressurization — Supercharging and Turbocharging

Backround

Pressurizing the intake air (or mixture) by means of a compressor may be used to boost the specific power output of both SI and CI engines. Supercharging generally refers to the use of compressors that are mechanically driven from the engine crankshaft, while turbocharging refers to compressors powered by a turbine which extracts energy from the exhaust stream. Increasing the intake pressure increases the density and hence the mass flow rate of the intake mixture, allowing an increase in the fueling rate, thereby producing additional power. The mere process of increasing the cylinder pressure results in increased work output per cycle, as illustrated in the *P-V* diagram in Figure 8.3.15 which compares supercharged and naturally aspirated, air standard Otto cycles having the same compression ratio. The work done for the compressed intake cycle (Area 1,2,3,4,1 and Area 5,6,7,1,5) is greater than that for the naturally aspirated cycle (Area 1′,2′,3′,4′,1′) due to the boost of the intake pressure. Positive-displacement superchargers are capable of producing higher boost pressures than turbochargers, which are nearly always centrifugal-type fans. From a practical standpoint, the maximum useful boost pressure from either system is limited by the onset of autoignition in SI engines, and by the permissible mechanical and thermal stresses in CI engines.

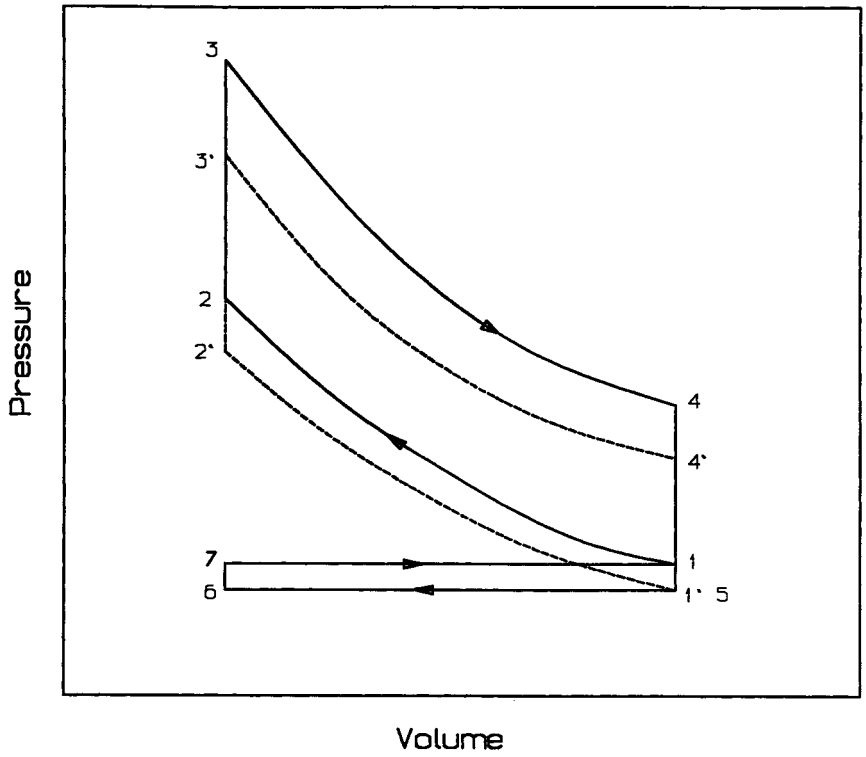


FIGURE 8.3.15 Comparison of supercharged and naturally aspirated Otto cycle.

Supercharging

The principal applications of supercharging of SI engines are in high-output drag-racing engines and in large aircraft piston engines to provide high specific output at takeoff and to improve power output at high altitudes. For diesel applications, supercharging is used mainly in marine and land-transportation applications. It is common to use either supercharging or turbocharging to improve the scavenging process in two-stroke diesel engines. Figure 8.3.16 is a schematic of an engine with a mechanically driven supercharger. Superchargers may be belt, chain, or gear driven from the engine crankshaft.

Two types of superchargers are in use: the positive displacement type (Roots blower) and the centrifugal type. Roots blowers may be classified as: (1) straight double lobe, (2) straight triple lobe, and (3) helix triple lobe (twisted 60%). The helix-triple-lobe-type runs quieter than the others and is generally recommended, especially for diesel engines operating under high torque at various speed conditions. The centrifugal-type, because of its high capacity and small weight and size, is best suited for applications where power and volumetric efficiency improvement are required at high engine speed, e.g., with aircraft engines. A centrifugal blower will also survive a backfire more readily than a Roots blower in SI applications. Since superchargers are directly driven from the engine output shaft, there is no inherent lag in the rate of pressure increase with engine speed, as is typically the case with turbochargers.

Turbocharging

Turbochargers utilize a centrifugal compressor that is directly connected to a turbine which extracts energy from the exhaust gases of the engine and converts it to the shaft work necessary to drive the compressor. Turbocharging is widely used to increase power output in automotive and truck applications of four-stroke SI and CI engines and to improve scavenging of two-stroke CI engines.

There are three methods of turbocharging: the constant-pressure, the pulse, and the pulse converter methods. In the constant-pressure method, as illustrated in Figure 8.3.17, the exhaust pressure is main-

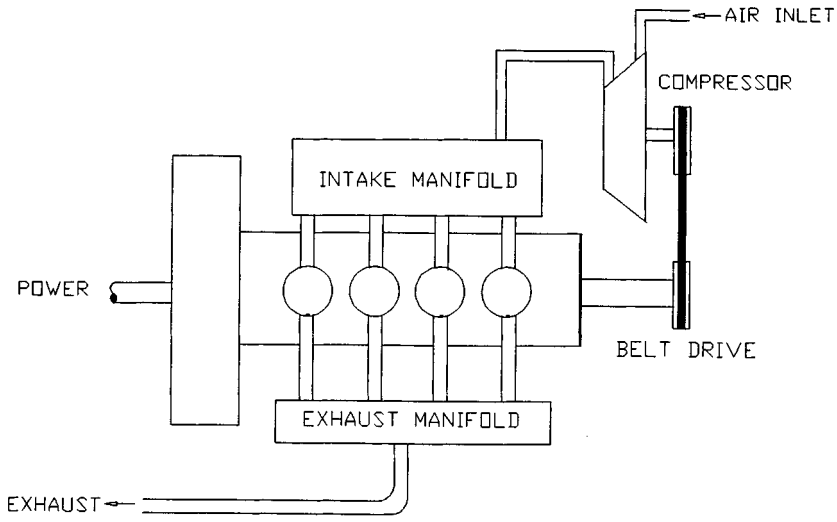


FIGURE 8.3.16 Schematic diagram of supercharged engine.

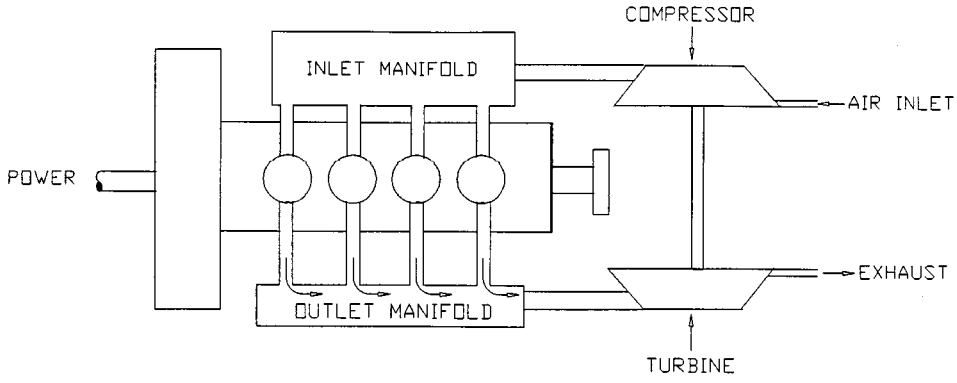


FIGURE 8.3.17 Schematic diagram of a constant-pressure turbocharger.

tained at a nearly constant level above atmospheric. To accomplish this, the exhaust manifold must be large enough to damp out the pressure fluctuations caused by the unsteady flow characteristic of the engine exhaust process. In this method, the turbine operates efficiently under steady-flow conditions; however, some engine power is lost because of the increased backpressure in the exhaust manifold.

The pulse turbocharger, as illustrated in Figure 8.3.18, utilizes the kinetic energy generated by the exhaust blow-down process in each cylinder. This is accomplished by using small exhaust lines grouped together in a common manifold to receive the exhaust from the cylinders which are blowing down sequentially. In this method, the pressure at the turbine inlet tends to fluctuate, which is not conducive to good turbine efficiency. This is offset to a large degree, however, by improved engine performance as a result of the lower exhaust backpressure relative to the constant-pressure method. The pulse converter method represents a compromise between the previous two techniques. In principle, this is accomplished by converting the kinetic energy in the blow-down process into a pressure rise at the turbine by utilizing one or more diffusers. Details of the different methods of turbocharging may be found in Watson and Janota (1982).

Recent advances in turbocharging technology have focused mainly on (1) improvement of the turbine transient response (turbo-lag), (2) improvement of the torque-speed characteristics of the engine, and (3) increasing the power output by increasing the boost pressure and using charge cooling (intercooling).

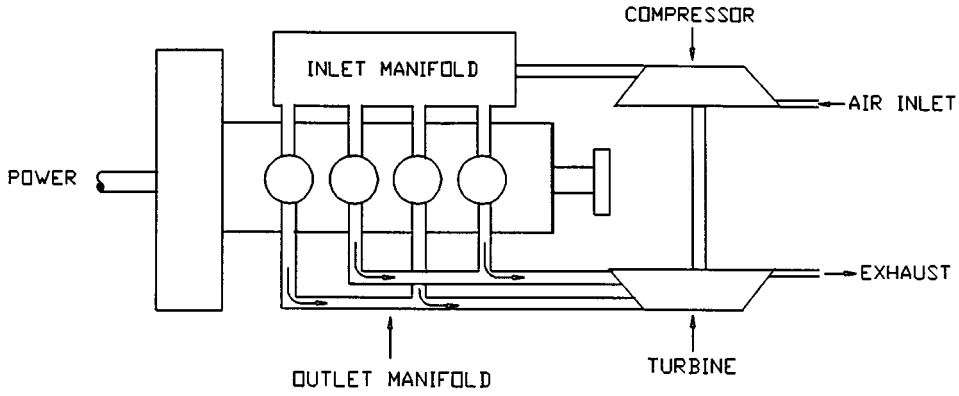


FIGURE 8.3.18 Schematic diagram of a pulse turbocharger.

The use of ceramic materials in fabricating turbine rotors improves the turbine transient response since they are lighter in weight and have less rotational inertia. Ceramic rotors also have greater thermal operating range because of their lower thermal expansion. The use of variable-geometry turbochargers can improve the low-speed torque characteristics of the engine and help reduce the transient response time. This is due to the ability of the variable-geometry turbocharger to change its internal geometry to accommodate low flow rates at low engine speeds and higher-volume flow rates at high engine speeds. However, since the geometry of the turbine rotor remains unchanged while the internal geometry varies, the turbine efficiency will be reduced for all internal geometries other than the optimum design geometry. In response to increased demand for diesel engines with high boost pressure and with size constraints, advances in the aerothermodynamics of axial/radial flow and of two-stage turbochargers, and also in the design of compressor and turbine blades, have allowed high boost pressure at improved overall turbocharger efficiency.

Charge cooling by means of a heat exchanger (intercooler) between the compressor and the intake ports is effective in reducing NO_x emissions and improving the power output of turbocharged diesel engines and in reducing the probability of knock in SI engines. There are two types of charge cooling in use, air-to-air and air-to-water. Air-to-water cooling is used in marine applications, where a source of cool water is available, while air-to-air intercoolers are used for automotive and truck applications.

Defining Terms

Antiknock index: The average of the two octane numbers obtained by the research method and the motor method.

Autoignition: The ability of a fuel-air mixture to spontaneously ignite under conditions of high temperature and pressure.

Bottom dead center (BDC): Piston located at its lowest position in the cylinder. Cylinder volume is maximum at BDC.

Brake mean effective pressure (bmep): Ratio of brake work output per cycle to the displacement volume.

Brake specific fuel consumption (bsfc): The ratio of fuel consumption rate in kg/hr to the engine output in kw.

Brake work: Work produced at the output shaft of an IC engine as measured by a dynamometer.

Cetane index: An approximation to the measured cetane number determined from an empirical relationship specified in ASTM D 976.

Cetane number: A measure of the autoignition quality of a fuel important for proper performance of CI engines determined experimentally through use of a CI CFR test engine.

Clearance volume: Combustion chamber volume remaining above the piston at TDC.

Compression ignition (CI) engine: Air alone is compressed in the cylinder and fuel is injected near TDC. Combustion results from autoignition of the fuel-air mixture due to the high temperature of the air.

Compression ratio: The ratio of the cylinder volume at BDC to the volume at TDC.

Cut-off ratio: Ratio of cylinder volume at the end of heat addition to the volume at the start of heat addition in the ideal diesel cycle.

Cylinder volume: Volume above piston at BDC. Equals displacement volume plus clearance volume.

Direct injection (DI): Method of fuel injection in low- and medium-speed CI engines wherein fuel is injected into the main combustion chamber which is formed by a bowl in the top of the piston.

Displacement volume: Difference in cylinder volume between TDC and BDC.

Equivalence ratio: Actual fuel-air ratio divided by stoichiometric fuel-air ratio.

Four-stroke engine: Entire cycle completed in two revolutions of the crankshaft and four strokes of the piston.

Fuel-air ratio: Ratio of mass of fuel to mass of air in the cylinder prior to combustion.

Glow plug: Electric heater installed in prechamber of an IDI diesel engine to aid cold starting.

Heterogeneous combustion: Refers to the mixture of liquid fuel droplets and evaporated fuel vapor and air mixture that is present in CI engine combustion chambers prior to ignition.

Ignition delay period: Period between start of injection and onset of autoignition in a CI engine.

Indicated mean effective pressure (imep): Ratio of net indicated work output of an IC engine to the displacement volume.

Indicated work: Work output of an IC engine cycle determined by an area calculation from an indicator diagram.

Indicator diagram: Pressure-volume trace for an IC engine cycle. Area enclosed by diagram represents work.

Indirect injection (IDI): Method of fuel injection used in high-speed CI engines wherein the fuel is injected into a precombustion chamber to promote fuel-air mixing and reduce ignition delay.

Knock: In SI engines: the noise that accompanies autoignition of the end portion of the uncombusted mixture prior to the arrival of the flame front. In CI engines: The noise that accompanies autoignition of large premixed fractions that are generated during prolonged ignition delay periods. Knock is detrimental to either type of engine.

NO_x: Harmful oxides of nitrogen (NO and NO_2) appearing in the exhaust products of IC engines.

Octane number: Antiknock rating for fuels important for prevention of autoignition in SI engines.

Particulates: Any exhaust substance, other than water, that can be collected on a filter. Harmful exhaust product from CI engines.

Power density: Power produced per unit of engine mass.

Premixed homogeneous combustion: Fuel and air are mixed in an appropriate combustible ratio prior to ignition process. This is the combustion mode for SI engines and for the initial combustion phase in CI engines.

Sac volume: Volume of nozzles below the needle of a diesel fuel injector that provides a source of UHC emissions in CI engines.

Scavenging: The process of expelling exhaust gases and filling the cylinder with fresh charge in two-stroke engines. This is often accomplished in SI engines by pressurizing the fresh mixture in the crankcase volume beneath the piston and in CI engines by using a supercharger or turbocharger.

Spark ignition (SI) engine: Homogeneous charge of air-fuel mixture is compressed and ignited by a spark.

Stroke: Length of piston movement from TDC to BDC, equal to twice the crankshaft throw.

Supercharging: Pressurizing the intake of an IC engine using a compressor that is mechanically driven from the crankshaft.

Surface ignition: A source of autoignition in SI engines caused by surface hot spots.

Swirl: Circular in-cylinder air motion designed into CI engines to promote fuel-air mixing.

Swirl ratio: Ratio of rotational speed of in-cylinder air (rpm) to engine speed (rpm).

Top dead center (TDC): Piston located at its uppermost position in the cylinder. Cylinder volume (above the piston) is minimum at TDC.

Turbocharging: Pressurizing the intake of an IC engine with a compressor that is driven by a turbine which extracts energy from the exhaust gas stream.

Two-stroke engine: Entire cycle completed in one revolution of the crankshaft and two strokes of the piston.

Unburned hydrocarbons (UHC): Harmful emission product from IC engines consisting of hydrocarbon compounds that remain uncombusted.

Volumetric efficiency: Ratio of the actual mass of air intake per cycle to the displacement volume mass determined at inlet temperature and pressure.

References

ASTM, 1995. *Annual Book of ASTM Standards*. American Society for Testing and Materials, Philadelphia.

Blair, G.P. Ed. 1988. *Advances in Two Stroke Cycle Engine Technology.* Society of Automotive Engineers, Inc., Warrendale, PA.

Ferguson, C.R. 1986. *Internal Combustion Engines, Applied Thermosciences*. John Wiley & Sons, New York.

Henein, N.A. 1972. *Emissions from Combustion Engines and Their Control*. Ann Arbor Science Publishers, Ann Arbor, MI.

Heywood, J.B. 1988. *Internal Combustion Engine Fundamentals*. McGraw-Hill, New York.

Keating, E.L. 1993. *Applied Combustion*. Marcel Dekker, New York.

Kummer, J.T. 1980. Catalysts for automobile emission control. *Prog. Energy Combust. Sci.* 6:177–199.

Lenz, H.P. 1992. *Mixture Formation in Spark-lgnition Engines*. Springer-Verlag, New York.

Norbye, J.P. 1971. *The Wankel Engine*. Chilton Press, Philadelphia.

Obert, E.F. 1973. *Internal Combustion Engines and Air Pollution,* 3rd ed. Harper & Row, New York.

Owen, K. and Coley, T. 1995. *Automotive Fuels Reference Book,* 2nd ed. Society of Automotive Engineers, Inc., Warrendale, PA.

SAE Fuels and Lubricants Standards Manual. 1993. Society of Automotive Engineers, Inc., Warrendale, PA.

SAE Surface Vehicle Emissions Standards Manual. 1993. Society of Automotive Engineers, Inc., Warrendale, PA.

Stone, R. 1993. *Introduction to Internal Combustion Engines,* 2nd ed. Society of Automotive Engineers, Inc., Warrendale, PA.

Taylor, C.F. 1985. *The Internal Combustion Engine in Theory and Practice,* 2nd ed. Vol. I and II. MIT Press, Cambridge, MA.

Watson, N. and Janota, M.S. 1982. *Turbocharging the Internal Combustion Engine,* John Wiley & Sons, New York.

Further Information

The textbooks on IC engines by Ferguson (1986), Heywood (1988), Obert (1973), Stone (1993), and Taylor (1985) listed under the references all provide excellent treatments of this subject. The book by Stone, in particular, is up-to-date and informative. The *Handbook of Engineering* (1966) by CRC Press, Boca Raton, FL, contains a chapter on IC Engines by A. Kornhauser. The Society of Automotive Engineers (SAE) publishes transactions, proceedings, and books related to all aspects of automotive engineering, including IC engines. Two very comprehensive handbooks distributed by SAE are the *Bosch Automotive Handbook,* and the *SAE Automotive Handbook*. For more information contact: SAE Publications, 400 Commonwealth Drive, Warrendale, PA, 15096–0001. (412)776-4970.

8.4 Hydraulic Turbines

Roger E. A. Arndt

Introduction

A hydraulic turbine is a mechanical device that converts the potential energy associated with a difference in water elevation **(head)** into useful work. Modern hydraulic turbines are the result of many years of gradual development. Economic incentives have resulted in the development of very large units (exceeding 800 mW in capacity) with efficiencies that are sometimes in excess of 95%.

The emphasis on the design and manufacture of very large turbines is shifting to the production of smaller units, especially in developed nations, where much of the potential for developing large-baseload plants has been realized. At the same time, the escalation in the cost of energy has made many smaller sites economically feasible and has greatly expanded the market for smaller turbines. The increased value of energy also justifies the cost of refurbishment and increasing the capacity of older facilities. Thus, a new market area is developing for updating older turbines with modern replacement **runners** having higher efficiency and greater capacity.

General Description

Typical Hydropower Installation

As shown schematically in Figure 8.4.1, the hydraulic components of a hydropower installation consist of an intake, penstock, guide vanes or distributor, turbine, and **draft tube**. Trash racks are commonly provided to prevent ingestion of debris into the turbine. Intakes usually require some type of shape transition to match the passageway to the turbine and also incorporate a gate or some other means of stopping the flow in case of an emergency or for turbine maintenance. Some types of turbines are set in an open flume; others are attached to a closed-conduit penstock.

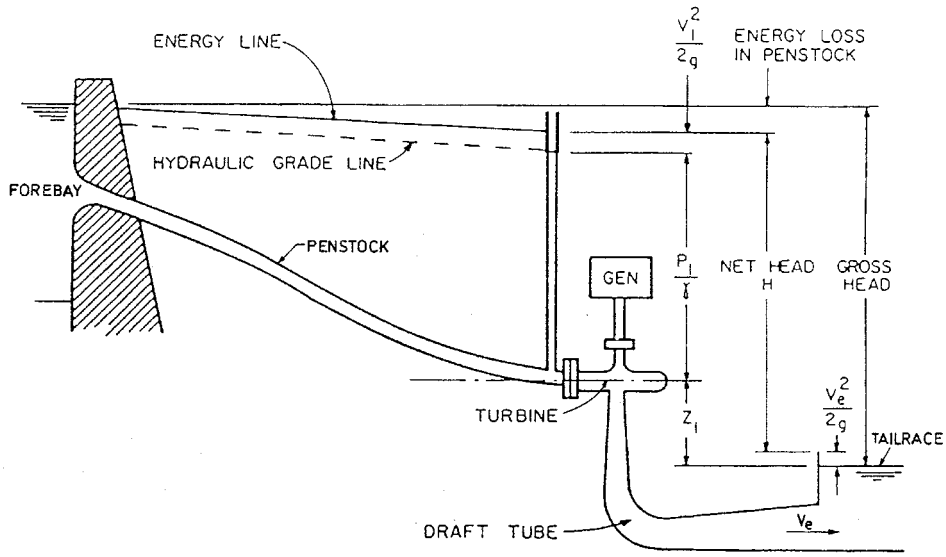


FIGURE 8.4.1 Schematic of a hydropower installation.

Turbine Classification

There are two types of turbines, denoted as impulse and reaction. In an impulse turbine, the available head is converted to kinetic energy before entering the **runner**; the power available is extracted from the flow at approximately atmospheric pressure. In a reaction turbine, the runner is completely submerged and both the pressure and the velocity decrease from inlet to outlet. The velocity head in the inlet to the turbine runner is typically less than 50% of the total head available.

Impulse Turbines. Modern impulse units are generally of the Pelton type and are restricted to relatively high-head applications (Figure 8.4.2). One or more jets of water impinge on a wheel containing many curved buckets. The jet stream is directed inward, sideways, and outward, thereby producing a force on the bucket, which in turn results in a torque on the shaft. All kinetic energy leaving the runner is "lost." A draft tube is generally not used since the runner operates under approximately atmospheric pressure and the head represented by the elevation of the unit above tailwater cannot be utilized.* Since this is a high-head device, this loss in available head is relatively unimportant. As will be shown later, the Pelton wheel is a low-**specific-speed** device. Specific speed can be increased by the addition of extra nozzles, the specific speed increasing by the square root of the number of nozzles. Specific speed can also be increased by a change in the manner of inflow and outflow. Special designs such as the Turgo or cross-flow turbines are examples of relatively high specific speed impulse units (Arndt, 1991).

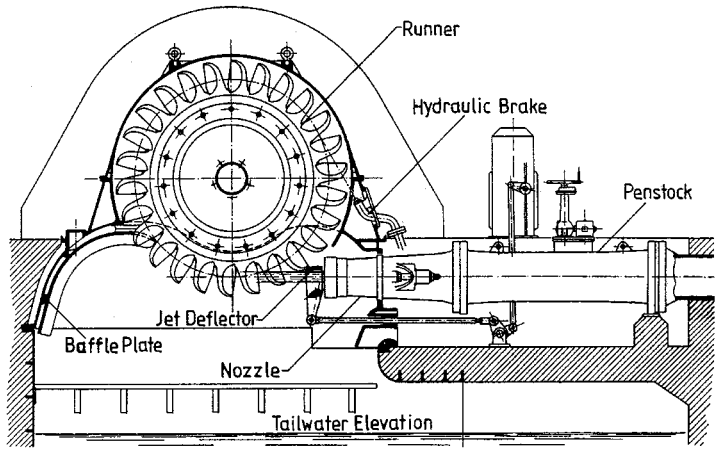


FIGURE 8.4.2 Cross section of a single wheel, single jet Pelton turbine. This is the third-highest-head pelton turbine in the world, H = 1447 m, n = 500 rpm, P = 35.2 MW, $N_s \sim 0.038$. (Courtesy of Vevey Charmilles Engineering Works, Adapted from J. Raabe, *Hydro Power: The Design, Use, and Function of Hydromechanical, Hydraulic, and Electrical Equipment*, VDI Verlag, Dusseldorf, Germany.)

Most Pelton wheels are mounted on a horizontal axis, although newer vertical-axis units have been developed. Because of physical constraints on orderly outflow from the unit, the maximum number of nozzles is generally limited to six or fewer. While the power of a reaction turbine is controlled by the **wicket gates**, the power of the Pelton wheel is controlled by varying the nozzle discharge by means of an automatically adjusted needle, as illustrated in Figure 8.4.2. Jet deflectors, or auxiliary nozzles are provided for emergency unloading of the wheel. Additional power can be obtained by connecting two wheels to a single generator or by using multiple nozzles. Since the needle valve can throttle the flow while maintaining essentially constant jet velocity, the relative velocities at entrance and exit remain unchanged, producing nearly constant efficiency over a wide range of power output.

* In principle, a draft tube could be used, which requires the runner to operate in air under reduced pressure. Attempts at operating an impulse turbine with a draft tube have not met with much success.

Reaction Turbines. Reaction turbines are classified according to the variation in flow direction through the runner. In radial- and mixed-flow runners, the flow exits at a radius different than from the radius at the inlet. If the flow enters the runner with only radial and tangential components, it is a radial-flow machine. The flow enters a mixed-flow runner with both radial and axial components. Francis turbines are of the radial- and mixed-flow types, depending on the design specific speed. A Francis turbine is illustrated in Figure 8.4.3.

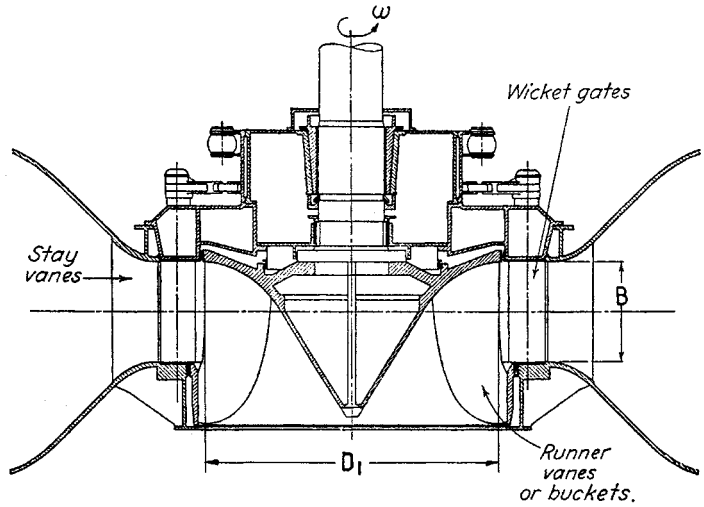


FIGURE 8.4.3 Francis turbine, $N_s \sim 0.66$. (Adapted from J.W. Daily, in *Engineering Hydraulics*, H. Rouse, Ed., New York, 1950. With permission.)

Axial-flow propeller turbines are generally either of the fixed-blade or Kaplan (adjustable-blade) variety. The "classical" propeller turbine, illustrated in Figure 8.4.4, is a vertical-axis machine with a scroll case and a radial wicket gate configuration that is very similar to the flow inlet for a Francis turbine. The flow enters radially inward and makes a right-angle turn before entering the runner in an axial direction. The Kaplan turbine has both adjustable runner blades and adjustable wicket gates. The control system is designed so that the variation in blade angle is coupled with the wicket gate setting in a manner which achieves best overall efficiency over a wide range of flow rates.

Some modern designs take full advantage of the axial-flow runner; these include the tube, bulb, and Straflo types illustrated in Figure 8.4.5. The flow enters and exits the turbine with minor changes in direction. A wide variation in civil works design is also permissible. The tubular-type can be fixed propeller, semi-Kaplan, or fully adjustable. An externally mounted generator is driven by a shaft which extends through the flow passage either upstream or downstream of the runner. The bulb turbine was originally designed as a high-output, low-head unit. In large units, the generator is housed within the bulb and is driven by a variable-pitch propeller at the trailing end of the bulb. Pit turbines are similar in principle to bulb turbines, except that the generator is not enclosed in a fully submerged compartment (the bulb). Instead, the generator is in a compartment that extends above water level. This improves access to the generator for maintenance.

Principles of Operation

Power Available, Efficiency

The power that can be developed by a turbine is a function of both the head and flow available:

$$P = \eta \rho g Q H \tag{8.4.1}$$

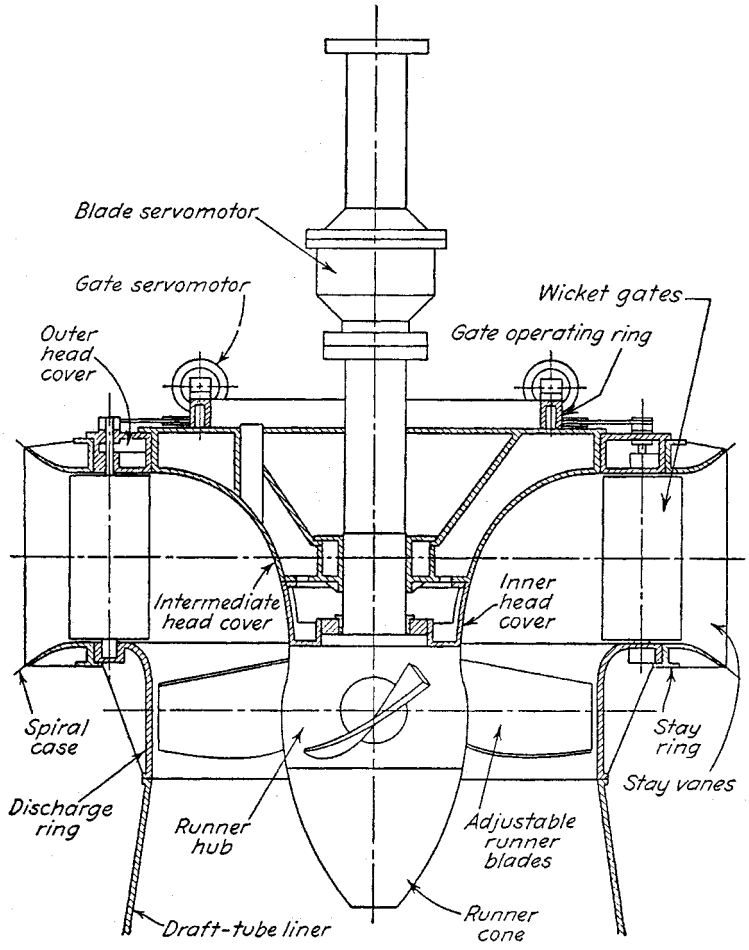


FIGURE 8.4.4 Smith-Kaplan axial flow turbine with adjustable-pitch runner blades $N_s \sim 2.0$. (From J.W. Daily, in *Engineering Hydraulics*, H. Rouse, Ed., New York, 1950. With permission.)

where η is the turbine efficiency, ρ is the density off water (kg/m^3), g is the acceleration due to gravity (m/sec^2), Q is the flow rate (m^3/sec), and H is the *net head* in meters. Net head is defined as the difference between the total head at the inlet to the turbine and the total head at the tailrace as illustrated in Figure 8.4.1. Different definitions of net head are used in practice, which depend on the value of the exit velocity head, $V_e^2/2g$, used in the calculation. The International Electrotechnical Test Code uses the velocity head at the draft tube exit.

The efficiency depends on the actual head and flow utilized by the turbine runner, flow losses in the draft tube, and the frictional resistance of mechanical components.

Similitude and Scaling Formulae

Under a given head, a turbine can operate at various combinations of speed and flow depending on the inlet settings. For reaction turbines, the flow into the turbine is controlled by the wicket gate angle, α. The flow is typically controlled by the nozzle opening in impulse units. Turbine performance can be described in terms of nondimensional variables:

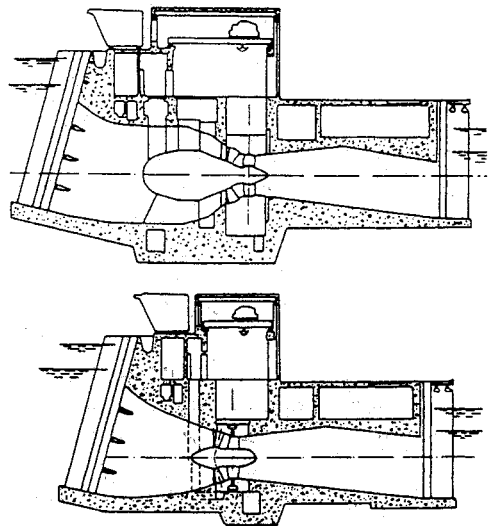


FIGURE 8.4.5 Comparison between bulb (upper) and Straflow (lower) turbines. (Courtesy of U.S. Department of Energy.)

$$\psi = \frac{2gH}{\omega^2 D^2} \tag{8.4.2}$$

$$\phi = \frac{Q}{\sqrt{2gH}D^2} \tag{8.4.3}$$

where ω is the rotational speed of the turbine in radians per second and D is the diameter of the turbine.

The hydraulic efficiency of the runner alone is given by

$$\eta_h = \frac{\phi}{\sqrt{\psi}}\left(C_1 \cos\alpha_1 - C_2 \cos\alpha_2\right) \tag{8.4.4}$$

where C_1 and C_2 are constants that depend on the specific turbine configuration and α_1 and α_2 are the inlet and outlet angles that the absolute velocity vectors make with the tangential direction. The value of $\cos\alpha_2$ is approximately zero at peak efficiency. The terms ϕ, ψ, α_1, and α_2 are interrelated. By using model test data, isocontours of efficiency can be mapped in the ϕ-ψ plane. This is typically referred to as a hill diagram as shown in Figure 8.4.6.

The specific speed is defined as

$$N_s \equiv \frac{\omega\sqrt{Q}}{(2gH)^{3/4}} = \sqrt{\frac{\phi}{\psi}} \tag{8.4.5}$$

A given specific speed describes a specific combination of operating conditions that ensures similar flow patterns and same efficiency in geometrically similar machines regardless of size and rotational speed of the machine. It is customary to define the design specific speed in terms of the value at the design head and flow where peak efficiency occurs. The value of specific speed so defined permits a classification of different turbine types.

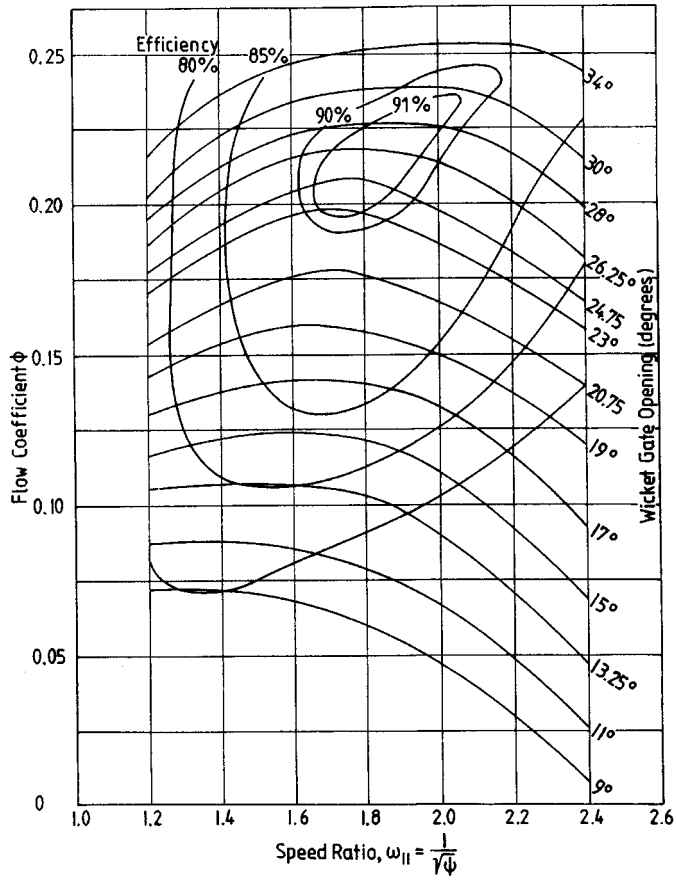


FIGURE 8.4.6 Typical hill diagram. (Adapted from T.L. Wall, Draft tube surging times two: the twin vortex problem, in *Hydro Rev.* 13(1): 60–69, 1994. With permission.)

The specific speed defined herein is dimensionless. Many other forms of specific speed exist which are dimensional and have different numerical values depending on the system of units used (Arndt, 1991).* The similarity arguments used to arrive at the concept of specific speed indicate that a given machine of diameter D operating under a head H will discharge a flow Q and produce a torque T and power P at a rotational speed ω given by

$$Q = \phi D^2 \sqrt{2gH} \tag{8.4.6}$$

$$T = T_{11} \rho D^3 2gH \tag{8.4.7}$$

$$P = P_{11} \rho D^2 (2gH)^{3/2} \tag{8.4.8}$$

$$\omega = \frac{2u_1}{D} = \omega_{11} = \frac{\sqrt{2gH}}{D} \qquad \left[\omega_{11} = \frac{1}{\sqrt{\psi}}\right] \tag{8.4.9}$$

* The literature also contains two other minor variations of the dimensionless form. One differs by a factor of $1/\pi^{1/2}$ and the other by $2^{3/4}$.

with

$$P_{11} = T_{11}\omega_{11} \tag{8.4.10}$$

where T_{11}, P_{11}, and ω_{11} are also nondimensional.* In theory, these coefficients are fixed for a machine operating at a fixed value of specific speed, independent of the size of the machine. Equations 8.4.6 through 8.4.10 can be used to predict the performance of a large machine using the measured characteristics of a smaller machine or model.

Factors Involved in Selecting a Turbine

Performance Characteristics

Impulse and reaction turbines are the two basic types of turbines. They tend to operate at peak efficiency over different ranges of specific speed. This is due to geometric and operational differences.

Impulse Turbines. Of the head available at the nozzle inlet, a small portion is lost to friction in the nozzle and to friction on the buckets. The rest is available to drive the wheel. The actual utilization of this head depends on the velocity head of the flow leaving the turbine and the setting above tailwater. Optimum conditions, corresponding to maximum utilization of the head available, dictate that the flow leaves at essentially zero velocity. Under ideal conditions, this occurs when the peripheral speed of the wheel is one half the jet velocity. In practice, optimum power occurs at a speed coefficient, ω_{11}, somewhat less than 1.0. This is illustrated in Figure 8.4.7. Since the maximum efficiency occurs at fixed speed for fixed H, V_j must remain constant under varying flow conditions. Thus, the flow rate Q is regulated with an adjustable nozzle. However, maximum efficiency occurs at slightly lower values of ω_{11} under partial power settings. Present nozzle technology is such that the discharge can be regulated over a wide range at high efficiency.

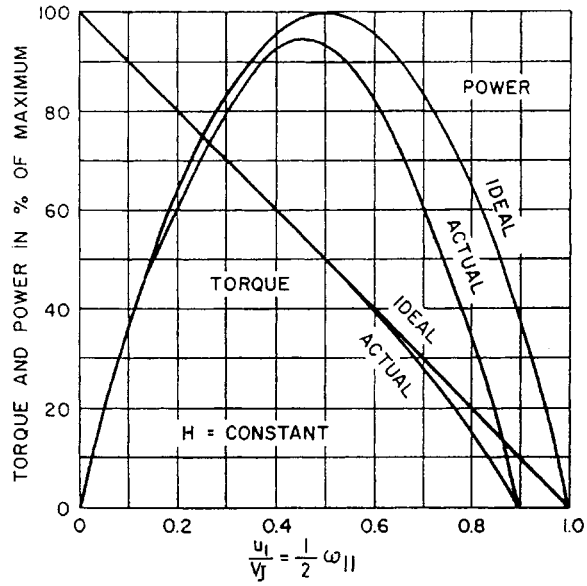


FIGURE 8.4.7 Ideal and actual variable-speed performance for an impulse turbine. (Adapted from J.W. Daily, in *Engineering Hydraulics*, H. Rouse, Ed., New York, 1950. With permission.)

* The reader is cautioned that many texts, especially in the American literature, contain dimensional forms of T_{11}, P_{11}, and ω_{11}.

A given head and penstock configuration establishes the optimum jet velocity and diameter. The size of the wheel determines the speed of the machine. The design specific speed is approximately

$$N_s = 0.77\frac{d_j}{D} \quad \text{(Pelton turbines)} \tag{8.4.11}$$

Practical values of d_j/D for Pelton wheels to ensure good efficiency are in the range 0.04 to 0.1, corresponding to N_s values in the range 0.03 to 0.08. Higher specific speeds are possible with multiple-nozzle designs. The increase is proportional to the square root of the number of nozzles. In considering an impulse unit, one must remember that efficiency is based on net head; the net head for an impulse unit is generally less than the net head for a reaction turbine at the same *gross head* because of the lack of a draft tube.

Reaction Turbines. The main difference between impulse and reaction turbines is the fact that a pressure drop takes place in the rotating passages of the reaction turbine. This implies that the entire flow passage from the turbine inlet to the discharge at the tailwater must be completely filled. A major factor in the overall design of modern reaction turbines is the draft tube. It is usually desirable to reduce the overall equipment and civil construction costs by using high-specific speed runners. Under these circumstances the draft tube is extremely critical both flow-stability and efficiency viewpoints.* At higher specific speed, a substantial percentage of the available total energy is in the form of kinetic energy leaving the runner. To recover this efficiently, considerable emphasis should be placed on the draft tube design.

The practical specific speed range for reaction turbines is much broader than for impulse wheels. This is due to the wider range of variables which control the basic operation of the turbine. The pivoted guide vanes allow for control of the magnitude and direction of the inlet flow. Because there is a fixed relationship among blade angle, inlet velocity, and peripheral speed for shock-free entry, this requirement cannot be completely satisfied at partial flow without the ability to vary blade angle. This is the distinction between the efficiency of fixed-propeller and Francis-types at partial loads and the fully adjustable Kaplan design.

Referring to Equation 8.4.4, optimum hydraulic efficiency of the runner would occur when α_2 is equal to 90°. However, the overall efficiency of the turbine is dependent on the optimum performance of the draft tube as well, which occurs with a little swirl in the flow. Thus, the best overall efficiency occurs with $\alpha_2 \approx 75°$ for high-specific speed turbines.

The determination of optimum specific speed in a reaction turbine is more complicated than for an impulse unit since there are more variables. For a radial-flow machine, an approximate expression is

$$N_s = 1.64\left[C_v \sin\alpha_1 \frac{B}{D_1}\right]^{1/2} \omega_{11} \quad \text{(Francis turbines)} \tag{8.4.12}$$

where C_v is the fraction of net head that is in the form of inlet velocity head and B is the height of the inlet flow passage (Figure 8.4.3). Value of N_s for Francis units is normally found to be in the range 0.3 to 2.5.

Standardized axial-flow machines are available in the smaller sizes. These units are made up of standard components, such as shafts and blades. For such cases,

$$N_s \sim \frac{\sqrt{\tan\beta}}{n_B^{3/4}} \quad \text{(propeller turbines)} \tag{8.4.13}$$

* This should be kept in mind when retrofitting an older, low-specific-speed turbine with a new runner of higher capacity.

where β is the blade pitch angle and n_B is the number of blades. The advantage of controllable pitch is also obvious from this formula, the best specific speed simply being a function of pitch angle.

It should be further noted that ω_{11} is approximately constant for Francis units and N_s is proportional to $(B/D_1)^{1/2}$. It can be also shown that velocity component based on the peripheral speed at the throat, ω_{11e}, is proportional to N_s. In the case of axial-flow machinery, ω_{11} is also proportional to N_s. For minimum cost, peripheral speed should be as high as possible, consistent with cavitation-free performance. Under these circumstances, N_s would vary inversely with the square root of head:

$$N_s = \frac{C}{\sqrt{H}} \quad (H \text{ is in meters}) \tag{8.4.14}$$

where the range of C is 21 to 30 for fixed propeller units, 21 to 32 for Kaplan units, and 16 to 25 for Francis units.

Performance Comparison. The physical characteristics of various runner configurations are summarized in Figure 8.4.8. It is obvious that the configuration changes with speed and head. Impulse turbines are efficient over a relatively narrow range of specific speed, whereas Francis and propeller turbines have a wider useful range. An important consideration is whether or not a turbine is required to operate over a wide range of load. Pelton wheels tend to operate efficiently over a wide range of power loading because of their nozzle design. In the case of reaction machines that have fixed geometry, such as Francis and propeller turbines, efficiency can vary widely with load. However, Kaplan and Deriaz* turbines can maintain high efficiency over a wide range of operating conditions. The decision of whether to select a simple configuration with a relatively "peaky" efficiency curve or go to the added expense of installing a more complex machine with a broad efficiency curve will depend on the expected operation of the plant and other economic factors.

Note in Figure 8.4.8 that there is an overlap in the range of application of various types of equipment. This means that either type of unit can be designed for good efficiency in this range, but other factors, such as generator speed and cavitation, may dictate the final selection.

Speed Regulation

The speed regulation of a turbine is an important and complicated problem. The magnitude of the problem varies with size, type of machine and installation, type of electrical load, and whether or not the plant is tied into an electrical grid. Note that runaway or no-load speed can be higher than design speed by factors as high as 2.6. This is an important design consideration for all rotating parts, including the generator.

The speed of a turbine has to be controlled to a value that matches the generator characteristics and the grid frequency:

$$n = \frac{120 f}{N_p} \tag{8.4.15}$$

where n is turbine speed in rpm, f is the required grid frequency in Hz, and N_p is the number of poles in the generator. Typically, N_p is in multiples of 4. There is a tendency to select higher speed generators to minimize weight and cost. However, consideration has to be given to speed regulation.

It is beyond the scope of this section to discuss the question of speed regulation in detail. Regulation of speed is normally accomplished through flow control. Adequate control requires sufficient rotational inertia of the rotating parts. When load is rejected, power is absorbed, accelerating the flywheel; and when load is applied, some additional power is available from deceleration of the flywheel. Response

* An adjustable blade mixed-flow turbine (Arndt, 1991).

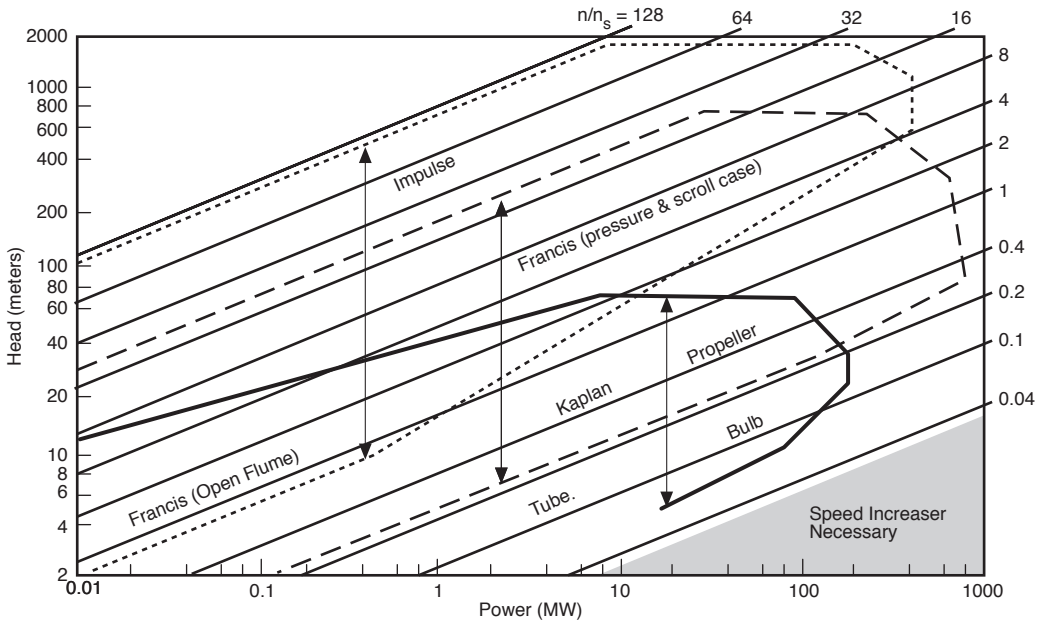


FIGURE 8.4.8 Application chart for various turbine types (n/n_S is the ratio of turbine speed in rpm, n, to specific speed defined in the metric system, $n_S = nP^{1/2}/H^{3/4}$ with P in kilowatts). (From Arndt, R.E.A., in *Hydropower Engineering Handbook*, J.S. Gulliver and R.E.A. Arndt, Eds., McGraw-Hill, New York, 1991, 4.1–4.67. With permission.)

time of the governor must be carefully selected, since rapid closing can lead to excessive pressures in the penstock.

A Francis turbine is controlled by opening and closing the wicket gates, which vary the flow of water according to the load. The actuator components of a governor are required to overcome the hydraulic and frictional forces and to maintain the wicket gates in fixed position under steady load. For this reason, most governors have hydraulic actuators. On the other hand, impulse turbines are more easily controlled, because the jet can be deflected or an auxiliary jet can by bypass flow from the power-producing jet without changing the flow rate in the penstock. This permits long delay times for adjusting the flow rate to the new flywheel; and when load is applied power conditions. The spear or needle valve controlling the flow rate can close quite slowly, say, in 30 to 60 sec, thereby minimizing pressure rise in the penstock.

Several types of governors are available, which vary with the work capacity desired and/or the degree of sophistication of control. These vary from pure mechanical to mechanical-hydraulic and electrohydraulic. Electrohydraulic units are sophisticated pieces of equipment and would not be suitable for remote regions. The precision of governing necessary will depend on whether the electrical generator is synchronous or asynchronous (induction type). There are advantages to the induction type of generator. It is less complex and therefore less expensive, but has typically slightly lower efficiency. Its frequency is controlled by the frequency of the grid it is feeding into, thereby eliminating need of an expensive the conventional governor. It cannot operate independently but can only feed into a network and does so with lagging power factor which may or may not be a disadvantage, depending on the nature of the load. Long transmission lines, for example, have a high capacitance and in this case the lagging power factor may be an advantage.

Speed regulation is a function of the flywheel effect of the rotating components and the inertia of the water column of the system. The start-up time of the rotating system is given by

$$t_s = \frac{I\omega^2}{P} \tag{8.4.16}$$

where I = moment of inertia of the generator and turbine, kg · m^2 (Bureau of Reclamation, 1966).

The start-up time of the water column is given by

$$t_p = \frac{\sum LV}{gH} \tag{8.4.17}$$

where L = length of water column
V = velocity in each component of the water column

For good speed regulation, it is desirable to keep $t_s/t_p > 4$. Lower values can also be used, although special precautions are necessary in the control equipment. Higher ratios of t_s/t_p can be obtained by increasing I or decreasing t_p. Increasing I implies a larger generator, which also results in higher costs. The start-up time of the water column can be reduced by reducing the length of the flow system, by using lower velocities, or by adding **surge tanks**, which essentially reduce the effective length of the conduit. A detailed analysis should be made for each installation, since for a given length, head, and discharge the flow area must be increased to reduce t_p, which leads to associated higher construction costs.

Cavitation and Turbine Setting

Another factor that must be considered prior to equipment selection is the evaluation of the turbine with respect to tailwater elevations. Hydraulic turbines are subject to pitting due to cavitation (Arndt, 1981, 1991). For a given head, smaller, lower-cost, high-speed runner must be set lower (i.e, closer to tailwater or even below tailwater) than a larger, higher-cost, low-speed turbine runner. Also, atmospheric pressure or elevation above sea level is a factor, as are tailwater elevation ion variations and operating requirements. This is a complex concept which can only be accurately resolved by model tests. The runner design will have different cavitation characteristics. Therefore, the anticipated turbine location or setting with respect to tailwater elevations is an important consideration in turbine selection.

Cavitation is not normally a problem with impulse wheels. However, by the very nature of their operation, cavitation is an important factor in reaction turbine installations. The susceptibility for cavitation to occur is a function of the installation and the turbine design. This can be expressed conveniently in terms of Thoma's sigma defined as

$$\sigma_T = \frac{H_a - H_v - z}{H} \tag{8.4.18}$$

where H_a is the atmospheric pressure head, H_V is the vapor pressure need (generally negligible), and z is the elevation of a turbine reference plane above the tailwater (see Figure 8.4.1). Draft tube losses and the exit velocity head have been neglected.

σ_T must be above a certain value to avoid cavitation problems. The critical value of σ_T is a function of specific speed (Arndt, 1991). The Bureau of Reclamation (1966) suggests that cavitation problems can be avoided when

$$\sigma_T > 0.26 N_s^{1.64} \tag{8.4.19}$$

Equation 8.4.19 does not guarantee total elimination of cavitation, only that cavitation is within acceptable limits. Cavitation can be totally avoided only if the value σ_T at an installation is much greater than the limiting value given in Equation 8.4.19. The value of σ_T for a given installation is known as

the plant sigmas, σ_p. Equation 8.4.19 should only be considered as a guide in selecting σ_p, which is normally determined by a model test in the manufacturer's laboratory. For a turbine operating under a given head, the only variable controlling σ_p is the turbine setting z. The required value of σ_p then controls the allowable setting above tailwater:

$$z_{\text{allow}} = H_a - H_v - \sigma_p H \tag{8.4.20}$$

It must be borne in mind that H_a varies with elevation. As a rule of thumb, H_a decreases from the sea-level value of 10.3 m by 1.1 m for every 1000 m above sea level.

Defining Terms

Draft tube: The outlet conduit from a turbine which normally acts as a diffuser. This is normally considered to be an integral part of the unit.

Forebay: The hydraulic structure used to withdraw water from a reservoir or river. This can be positioned a considerable distance upstream from the turbine inlet.

Head: The specific energy per unit weight of water. *Gross head* is the difference in water surface elevation between the forebay and tailrace. *Net head* is the difference between *total head* (the sum of velocity head $V^2/2g$, pressure head $p/\rho g$, and elevation head z at the inlet and outlet of a turbine. Some European texts use specific energy per unit mass, e.g., specific kinetic energy is $V^2/2$.

Runner: The rotating component of a turbine in which energy conversion takes place.

Specific speed: A universal number for a given machine design.

Spiral case: The inlet to a reaction turbine.

Surge tank: A hydraulic structure used to diminish overpressures in high-head facilities due to water hammer resulting from the sudden stoppage of a turbine

Wicket gates: Pivoted, streamlined guide vanes that control the flow of water to the turbine.

References

Arndt, R.E.A. 1981. Cavitation in fluid machinery and hydraulic structures. *Ann. Rev. Fluid Mech.* 13:273–328.

Arndt, R.E.A. 1991. Hydraulic turbines, in *Hydropower Engineering Handbook,* J.S. Gulliver and R.E.A. Arndt, Eds., pp. 4.1–4.67 McGraw-Hill, New York.

Bureau of Reclamation. 1966. Selecting Hydraulic Reaction Turbines, Engineering Monograph No. 20.

Daily, J.W. 1950. Hydraulic machinery, in H. Rouse, Ed., Engineering Hydraulics, New York.

International Code for the Field Acceptance Tests of Hydraulic Turbines, 1963. International Electrotechnical Commission, Publication 41.

Raabe, J. 1985. *Hydro Power: The Design, Use, and Function of Hydromechanical, Hydraulic, and Electrical Equipment*. VDI Verlag, Dusseldorf, Germany.

Wahl, T.L. 1994. Draft tube surging times two: The twin vortex problem. *Hydro Rev.* 13(1):60–69.

Further Information

J. Fluids Eng, published quarterly by the ASME.

ASME Symposia Proc. on Fluid Machinery and Cavitation, published by the Fluids Eng. Div.

Hydro Review, published eight times per year by HCI Publications, Inc., Kansas City, MO.

L.F. Moody and T. Zowski, Hydraulic machinery, in *Handbook of Applied Hydraulics,* C.V. Davis and K.E. Sorenson, Eds., McGraw-Hill, New York, 1992.

Waterpower and Dam Construction, published monthly by Reed Business Publishing, Surrey, U.K.

8.5 Stirling Engines

William B. Stine

Introduction

The Stirling engine was patented in 1816 by Rev. Robert Stirling, a Scottish minister (Figure 8.5.1). Early Stirling engines were coal-burning, low-pressure air engines built to compete with saturated steam engines for providing auxiliary power for manufacturing and mining. In 1987, John Ericsson built an enormous marine Stirling engine with four 4.2-m-diameter pistons. Beginning in the 1930s, the Stirling engine was brought to a high state of technology development by Philips Research Laboratory in Eindhoven, The Netherlands with the goal of producing small, quiet electrical generator sets to be used with high-power-consuming vacuum tube electronic devices. Recently, interest in Stirling engines has resurfaced, with solar electric power generation (Stine and Diver, 1994) and hybrid automotive applications in the forefront.

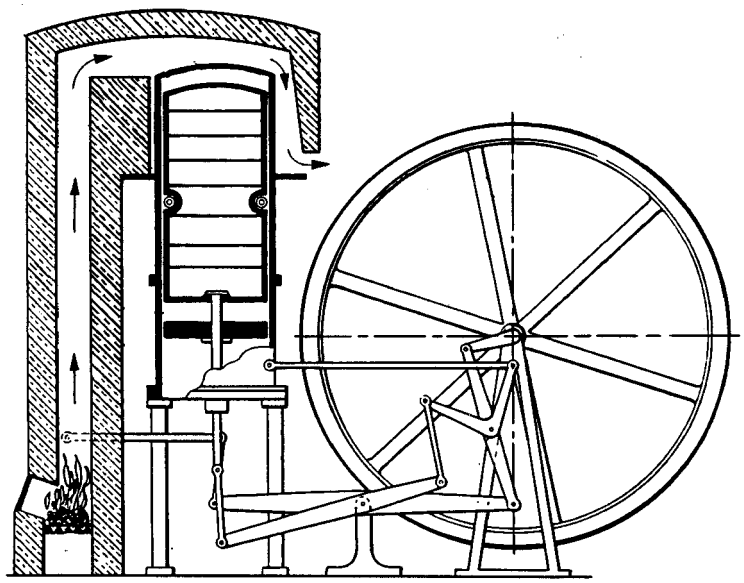


FIGURE 8.5.1 The original patent Stirling engine of Rev. Robert Stirling.

In theory, the **Stirling cycle** engine can be the most efficient device for converting heat into mechanical work with high efficiencies requiring high-temperatures. In fact, with regeneration, the efficiency of the Stirling cycle equals that of the Carnot cycle, the most efficient of all ideal thermodynamic cycles. (See West, 1986 for further discussion of the thermodynamics of Stirling cycle machines.)

Since their invention, prototype Stirling engines have been developed for automotive purposes; they have also been designed and tested for service in trucks, buses, and boats (Walker, 1973). The Stirling engine has been proposed as a propulsion engine in yachts, passenger ships, and road vehicles such as city buses (Meijer, 1992). The Stirling engine has also been developed as an underwater power unit for submarines, and the feasibility of using the Stirling for high-power space-borne systems has been explored by NASA (West, 1986). The Stirling engine is considered ideal for solar heating, and the first solar application of record was by John Ericsson, the famous British-American inventor, in 1872 (Stine and Diver, 1994).

Stirling engines are generally externally heated engines. Therefore, most sources of heat can be used to drive them, including combustion of just about anything, radioisotopes, solar energy, and exothermic

chemical reactions. High-performance Stirling engines operate at the thermal limits of the materials used for their construction. Typical temperatures range fiom 650 to 800°C (1200 to 1470°F), resulting in engine conversion efficiencies of around 30 to 40%. Engine speeds of 2000 to 4000 rpm are common

Thermodynamic Implementation of the Stirling Cycle

In the ideal Stirling cycle, a **working gas** is alternately heated and cooled as it is compressed and expanded. Gases such as helium and hydrogen, which permit rapid heat transfer and do not change phase, are typically used in the high-performance Stirling engines. The ideal Stirling cycle combines four processes, two constant-temperature heat-exchange processes and two constant-volume heat-exchange processes. Because more work is done by expanding high-temperature, high-pressure gas than is required to compress low-temperature, low-pressure gas, the Stirling cycle produces net work, which can drive an electric alternator or other mechanical devices.

Working Gases

In the Stirling cycle, the working gas is alternately heated and cooled in constant-temperature and constant-volume processes. The traditional gas for Stirling engines has been air at atmospheric pressure. At this pressure, air has a reasonably high density and therefore can be used directly in the cycle with loss of working gas through seals a minor problem. However, internal component temperatures are limited because of the oxygen in air which can degrade materials rapidly.

Because of their high heat-transfer capabilities, hydrogen and helium are used for high-speed, high-performance Stirling engines. To compensate for the low density of these gases, the mean pressure of the working gas is raised by charging the gas spaces in the engine to high pressures. Compression and expansion vary above and below this **charge pressure**. Hydrogen, thermodynamically a better choice, generally results in more-efficient engines than does helium (Walker, 1980). Helium, on the other hand, has fewer material-compatibility problems and is safer to work with. To maximize power, high-performance engines typically operate at high pressure, in the range of 5 to 20 MPa (725 to 2900 psi). Operation at these high gas pressures makes sealing difficult, and seals between the high-pressure region of the engine and those parts at ambient pressure have been problematic in some engines. New designs to reduce or eliminate this problem are currently being developed.

Heat Exchange

The working gas is heated and cooled by heat exchangers that add heat from an external source, or reject heat to the surroundings. Further, in most engines, an internal heat storage unit stores and rejects heat during each cycle.

The **heater** of a Stirling engine is usually made of many small-bore tubes that are heated externally with the working gas passing through the inside. External heat transfer by direct impingement of combustion products or direct adsorption of solar irradiation is common. A trade-off between high heat-transfer rate using many small-bore tubes with resulting pumping losses, and fewer large-bore tubes and lower pumping losses drives the design. A third criterion is that the volume of gas trapped within these heat exchangers should be minimal to enhance engine performance. In an attempt to provide more uniform and constant-temperature heat transfer to the heater tubes, **reflux** heaters are being developed (Stine and Diver, 1994). Typically, by using sodium as the heat-transfer medium, liquid is evaporated at the heat source and is then condensed on the outside surfaces of the engine heater tubes.

The **cooler** is usually a tube-and-shell heat exchanger. Working gas is passed through the tubes, and cooling water is circulated around the outside. The cooling water is then cooled in an external heat exchanger. Because all of the heat rejected from the power cycle comes from the cooler, the Stirling engine is considered ideal for cogeneration applications.

Most Stirling engines incorporate an efficiency-enhancing **regenerator** that captures heat from the working gas during constant-volume cooling and replaces it when the gas is heated at constant volume. Heating and cooling of the regenerator occurs at over 60 times a second during high-speed engine operation. In the ideal cycle, all of the heat-transferred during the constant volume heating and cooling

processes occurs in the regenerator, permitting the external heat addition and rejection to be efficient constant-temperature heat-transfer processes. Regenerators are typically chambers packed with fine-mesh screen wire or porous metal structures. There is enough thermal mass in the packing material to store all of the heat necessary to raise the temperature of the working gas from its low to its high temperature. The amount of heat stored by the regenerator is generally many times greater than the amount added by the heater.

Power Control

Rapid control of the output power of a Stirling engine is highly desirable for some applications such as automotive and solar electric applications. In most Stirling engine designs, rapid power control is implemented by varying the density (i.e., the mean pressure) of the working gas by bleeding gas from the cycle when less power is desired. To return to a higher power level, high-pressure gas must be reintroduced into the cycle. To accomplish this quickly and without loss of working gas, a complex system of valves, a temporary storage tank, and a compressor are used.

A novel method of controlling the power output is to change the length of stroke of the power piston. This can be accomplished using a variable-angle swash plate drive as described below. Stirling Thermal Motors, Inc., uses this method on their STM 4-120 Stirling engine (Figure 8.5.2).

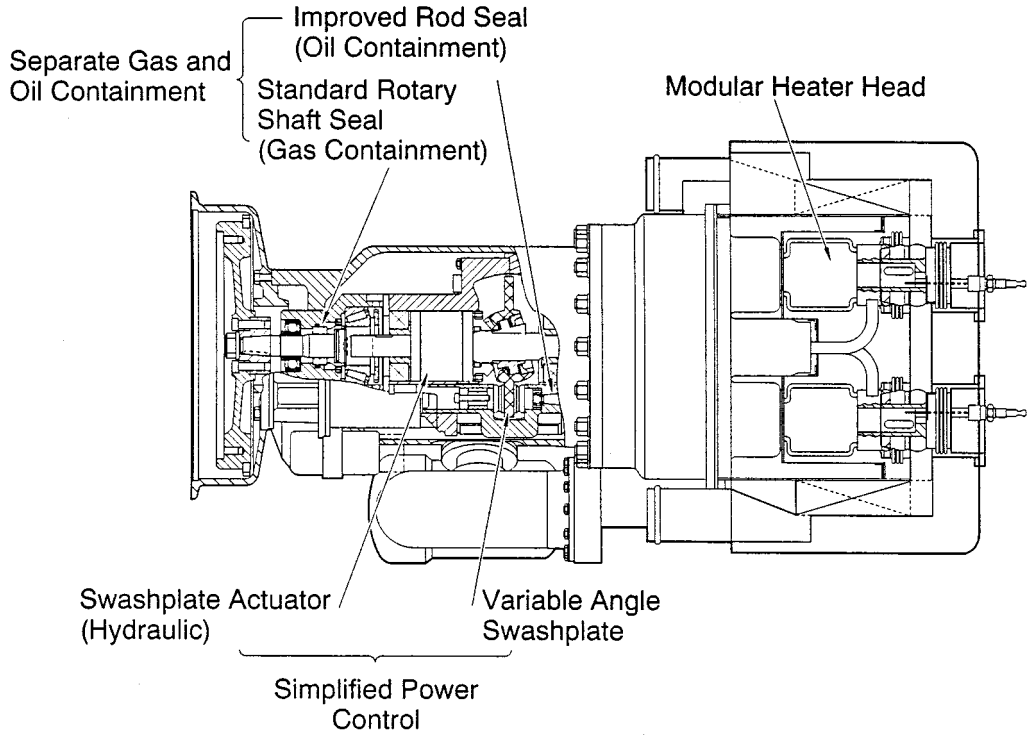


FIGURE 8.5.2 Stirling Thermal Motors 4-120 variable swash plate Rinia configuration engine. (Courtesy Stirling Thermal Motors, Ann Arbor, Michigan.)

Mechanical Implementation of the Stirling Cycle

Piston/Displacer Configurations

To implement the Stirling cycle, different combinations of machine components have been designed to provide for the constant-volume movement of the working gas between the high- and low-temperature regions of the engine, and compression and expansion during the constant-temperature heating and

cooling. The compression and expansion part of the cycle generally take place in a cylinder with a piston. Movement of the working gas back and forth through the heater, regenerator, and cooler at constant volume is often implemented by a **displacer**. A displacer in this sense is a hollow plug that, when moved to the cold region, displaces the working gas from the cold region causing it to flow to the hot region and vice versa. Only a small pressure difference exists between either end of the displacer, and, therefore, sealing requirements and the force required to move it are minimal.

Three different design configurations are generally used (Figure 8.5.3). Called the alpha-, beta-, and gamma-configurations. Each has its distinct mechanical design characteristics, but the thermodynamic cycle is the same. The **alpha-configuration** uses two pistons on either side of the heater, regenerator, and the cooler. These pistons first move uniformly in the same direction to provide constant-volume processes to heat or cool the gas. When all of the gas has been moved into one cylinder, one piston remains fixed and the other moves to compress or expand the gas. Compression work is done by the **cold piston** and expansion work done on the **hot piston** The alpha-configuration does not use a displacer. The Stirling Power Systems V-160 engine (Figure 8.5.4) is an example of this configuration.

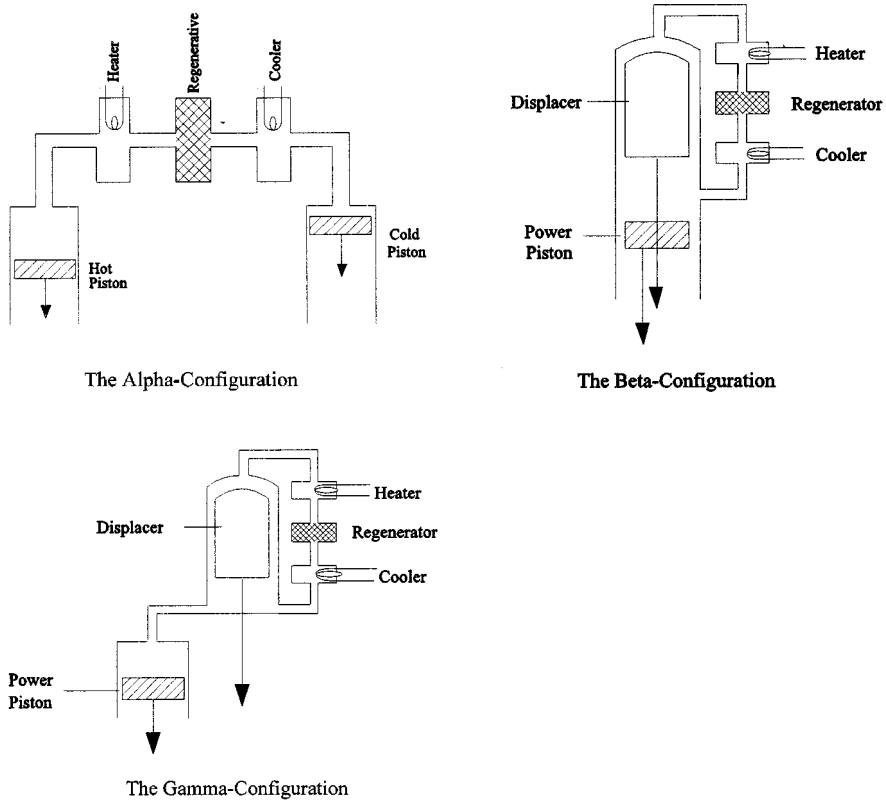


FIGURE 8.5.3 Three fundamental mechanical configurations for Stirling Engines.

A variation on using two separate pistons to implement the alpha-configuration is to use the front and back side of a single piston called a **double-acting piston**. The volume at the front side of one piston is connected, through the heater, regenerator, and cooler, to the volume at the back side of another piston. With four such double-acting pistons, each 90° out of phase with the next, the result is a four-cylinder alpha-configuration engine. This design is called the *Rinia* or *Siemens configuration* and the United Stirling 4-95 (Figure 8.5.5) and the Stirling Thermal Motors STM 4-120 (Figure 8.5.2) are current examples.

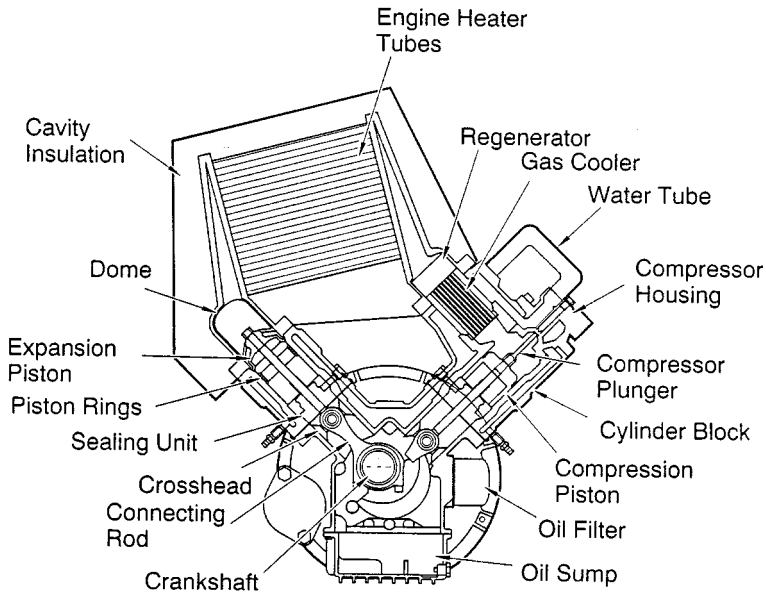


FIGURE 8.5.4 Stirling Power Systems/Solo Kleinmotoren V-160 alpha-configuration Stirling engine.

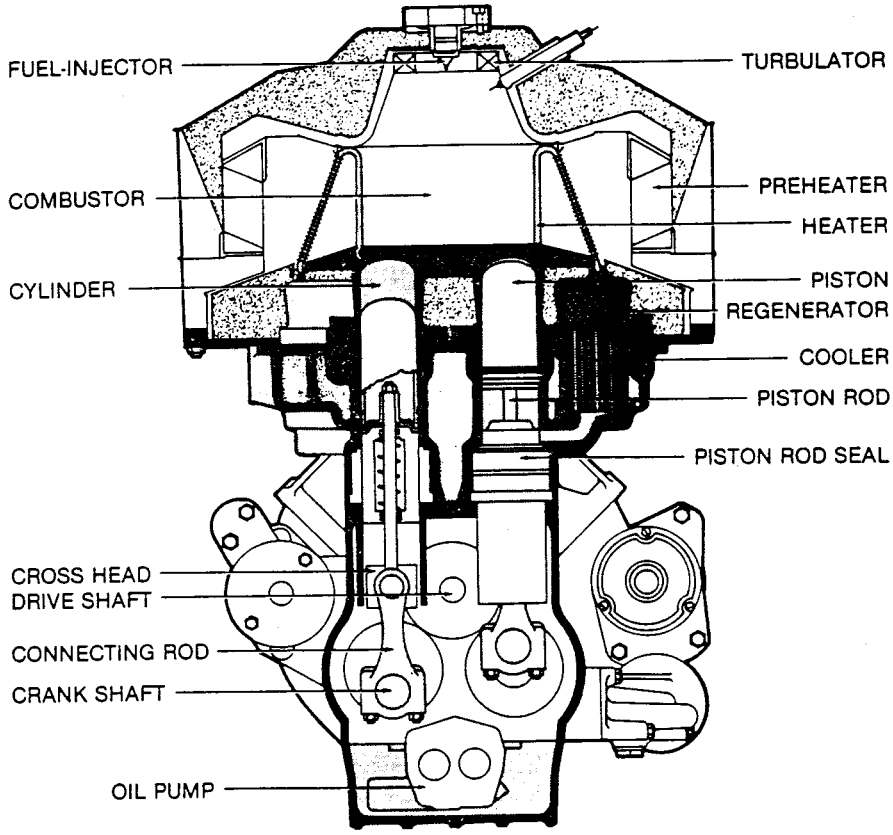


FIGURE 8.5.5 The 4-95 high-performance Siemens configuration Rinia engine by United Stirling (Malmo, Sweden).

The **beta-configuration** is a design incorporating a displacer and a power piston in the same cylinder. The displacer shuttles gas between the hot end and the cold end of the cylinder through the heater, regenerator, and cooler. The power piston, usually located at the cool end of the cylinder, compresses the working gas when the gas is in the cool end and expands the working gas when the gas has been moved to the hot end. The original patent engine by Robert Stirling and most free-piston Stirling engines discussed below are of the beta-configuration.

The third configuration, using separate cylinders for the displacer and the power piston, is called the **gamma-configuration**. Here, the displacer shuttles gas between the hot end and the cold end of a cylinder through the heater, regenerator, and cooler, just as with the beta-configuration. However, the power piston is in a separate cylinder, pneumatically connected to the displacer cylinder.

Piston/Displacer Drives

Most Stirling engine designs dynamically approximate the Stirling cycle by moving the piston and displacer with **simple harmonic motion**, either through a crankshaft, or bouncing as a spring/mass second-order mechanical system. For both, a performance penalty comes from the inability of simple harmonic motion to perfectly follow the desired thermodynamic processes. A more desirable dynamic from the cycle point of view, called overdriven or **bang-bang motion**, has been implemented in some designs, most notably the **Ringbom configuration** and engines designed by Ivo Kolin (West, 1986).

Kinematic Engines. Stirling engine designs are usually categorized as either kinematic or free-piston. The power piston of a **kinematic Stirling engine** is mechanically connected to a rotating output shaft. In typical configurations, the power piston is connected to the crankshaft with a connecting rod. In order to eliminate side forces against the cylinder wall, a **cross-head** is often incorporated, where the connecting rod connects to the cross-head, which is laterally restrained so that it can only move linearly in the same direction as the piston. The power piston is connected to the cross-head and therefore experiences no lateral forces. The critical sealing between the high-pressure and low-pressure regions of the engine can now be created using a simple **linear seal** on the shaft between the cross-head and the piston. This design also keeps lubricated bearing surfaces in the low-pressure region of the engine, reducing the possibility of fouling heat-exchange surfaces in the high-pressure region of the engine. If there is a separate displacer piston as in the beta- and gamma configurations, it is also mechanically connected to the output shaft.

A variation on crankshaft/cross-head drives is the **swash plate** or **wobble-plate drive**, used with success in some Stirling engine designs. Here, a drive surface affixed to the drive shaft at an angle, pushes fixed piston **push rods** up and down as the slanted surface rotates beneath. The length of stroke for the piston depends on the angle of the plate relative to the axis of rotation. The STM 4-120 engine (Figure 8.5.2) currently being commercialized by Stirling Thermal Motors incorporates a **variable-angle swash plate drive** that permits variation in the length of stroke of the pistons.

Free-Piston Engine/Converters. An innovative way of accomplishing the Stirling cycle is employed in the free-piston engine. In this configuration, the power piston is not mechanically connected to an output shaft. It bounces alternately between the space containing the working gas and a spring (usually a gas spring). In many designs, the displacer is also free to bounce on **gas springs** or mechanical springs (Figure 8.5.6). This configuration is called the Beale free-piston Stirling engine after its inventor, William Beale. Piston stroke, frequency, and the timing between the two pistons are established by the dynamics of the spring/mass system coupled with the variations in cycle pressure. To extract power, a magnet can be attached to the power piston and electric power generated as it moves past stationary coils. These Stirling engine/alternator units are called **free-piston Stirling converters**. Other schemes for extracting power from free-piston engines, such as driving a hydraulic pump, have also been considered.

Free-piston Stirling engines have only two moving parts, and therefore the potential advantages of simplicity, low cost, and ultra-reliability. Because electricity is generated internally, there are no dynamic seals between the high-pressure region of the engine and ambient, and no oil lubrication is required. This design promises long lifetimes with minimal maintenance. A number of companies are currently

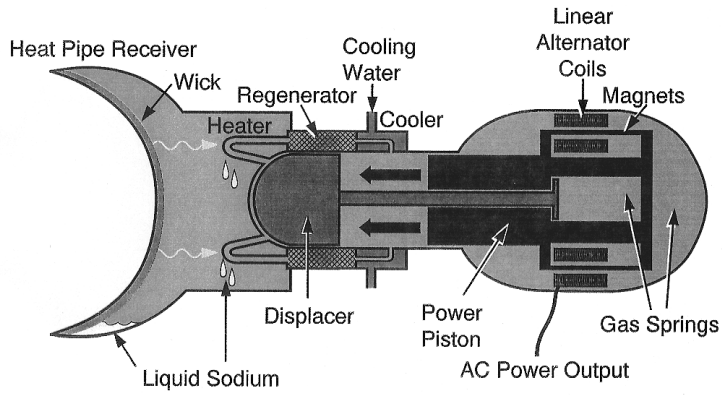


FIGURE 8.5.6 Basic components of a Beale free-piston Stirling converter incorporating a sodium heat pipe receiver for heating with concentrated solar energy.

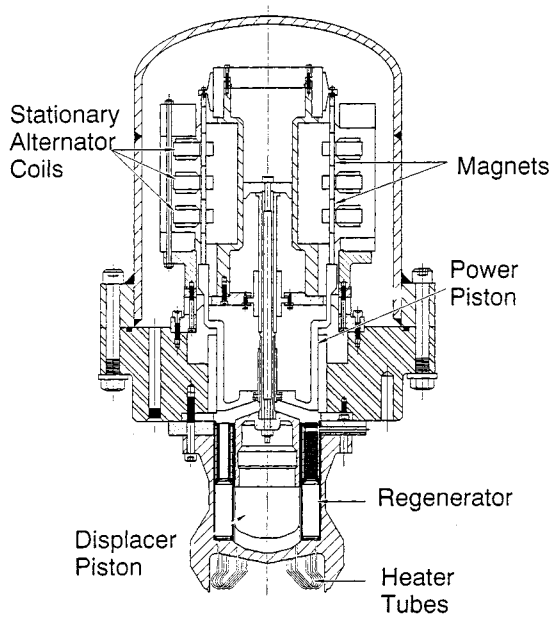


FIGURE 8.5.7 The Sunpower 9-kWe free-piston beta-configuration Stirling engine.

developing free-piston engines including Sunpower, Inc. (Figure 8.5.7), and Stirling Technology Company.

Seals and Bearings

Many proposed applications for Stirling engine systems require long-life designs. To make systems economical, a system lifetime of at least 20 years with minimum maintenance is generally required. Desired engine lifetimes for electric power production are 40,000 to 60,000 hr — approximately ten times longer than that of a typical automotive internal combustion engine. Major overhaul of engines, including replacement of seals and bearings, may be necessary within the 40,000- to 60,000-hr lifetime, which adds to the operating cost. A major challenge, therefore, in the design of Stirling engines is to reduce the potential for wear in critical components or to create novel ways for them to perform their tasks.

Piston seals differ from those used in internal combustion engines in a number of ways. Sealing of the power piston is critical since **blow-by loss** of the hydrogen or helium working gas must be captured and recompressed, or replaced from a high-pressure cylinder. Displacer sealing is less critical and only necessary to force most of the working gas through the heater, regenerator, and cooler. Oil for friction reduction or sealing cannot be used because of the danger of it getting into the working gas and fouling the heat-exchange surfaces. This leads to two choices for sealing of pistons, using **polymer sealing rings** or **gas bearings** (simply close tolerance fitting between piston and wall).

Free-piston engines with gas springs have special internal sealing problems. Small leakage can change the force-position characteristics of the "spring" and rapidly upset the phase and displacement dynamics designed into the engine. In order to prevent this, *centering ports* are used to ensure that the piston stays at a certain location; however, these represent a loss of potential work from the engine.

Materials

Materials used in Stirling engines are generally normal steels with a few exceptions. Materials that can withstand continuous operation at the cycle high temperature are required for the heater, regenerator, and the hot side of the displacement volume. Because most engines operate at high pressure, thick walls are often required. In the hot regions of the engine, this can lead to *thermal creep* due to successive heating and cooling. In the cool regions, large spaces for mechanical linkages can require heavy, thick walls to contain the gas pressure. Use of composite structure technology or reducing the size of the pressurized space can eliminate these problems.

Future of the Stirling Engine

The principal advantages of the Stirling engine, external heating and high efficiency, make this the engine of the future, replacing many applications currently using internal combustion engines and providing access to the sun as an inexpensive source of energy (Figure 8.5.6). For hybrid-electric automotive applications, the Stirling engine is not only almost twice as efficient as modern spark-ignition engines, but because of the continuous combustion process, it burns fuel more cleanly and is not sensitive to the quality or type of fuel. Because of the simplicity of its design, the Stirling engine can be manufactured as an inexpensive power source for electricity generation using biomass and other fuels available in developing nations.

Most importantly, the Stirling engine will provide access to inexpensive solar energy. Because it can receive its heat from a resource 93 million miles away using concentrating solar collectors, and because its manufacture is quite similar to the gasoline or diesel engine, and because economies of scale are certain when producing thousands of units per year, the Stirling engine is considered to be the least expensive alternative for solar energy electric generation applications in the range from 1 kWe to 100 MWe.

Defining Terms

Alpha-configuration: Design of a Stirling engine where two pistons moving out of phase, and cause the working gas between them to go through the four processes of the Stirling cycle.

Beale free-piston Stirling engine: Stirling engine configuration where the power piston and displacer in a single cylinder are free to bounce back and forth along a single axis, causing the enclosed working gas to go through the four processes of the Stirling cycle. Restoration forces are provided by the varying pressure of the working gas, springs (gas or mechanical), and the external load which can be a linear alternator or a fluid pump.

Beta-configuration: Design of a Stirling engine where the displacer and power piston are located in the same cylinder and cause the enclosed working gas to go through the four processes of the Stirling cycle.

Blow-by: The gas that leaks past a seal, especially a piston-to-cylinder seal.

Charge pressure: Initial pressure of the working gas.

Cooler: Heat exchanger that removes heat from the working fluid and transfers it out of the cycle.

Cross-head: A linear sliding bearing surface connected to a crankshaft by a connecting rod. It is purpose is to provide linear reciprocating motion along a single line of action.

Displacer: Closed volume 'plug' that forces the working fluid to move from one region of the engine to another by displacing it.

Double-acting piston: A piston in an enclosed cylinder where pressure can be varied on both sides of the piston, resulting a total amount of work being the sum of that done on or by both sides.

Dynamic seals: Seals that permit transfer of motion without permitting gas or oil leakage. These can be either *linear seals* permitting a shaft to move between two regions (i.e., the piston rod seals in a Stirling engine), or *rotating seals* that permit rotating motion to be transmitted from one region to another (i.e., the output shaft of a Stirling engine).

Free-piston Stirling converters: A name given to a hermetically sealed free-piston Stirling engine incorporating an internal alternator or pump.

Gamma-configuration: A design of a Stirling engine where the displacer and power piston are located in separate, connected cylinders and cause the enclosed working gas to go through the four processes of the Stirling cycle.

Gas bearing: A method of implementing the sliding seal between a piston and cylinder as opposed to using piston rings. Uses a precision-fitting piston that depends on the small clearance and long path for sealing and on the viscosity of the gas for lubrication.

Gas spring: A piston that compresses gas in a closed cylinder where the restoration force is linearly proportional to the piston displacement. This is a concept used in the design of free-piston Stirling engines.

Heater: A heat exchanger which adds heat to the working fluid from an external source.

Kinematic stirling engine: Stirling engine design that employ physical connections between the power piston, displacer, and a mechanical output shaft.

Linear seal: see **dynamic seals**.

Overdriven (bang-bang) motion: Linear motion varying with time as a square-wave function. An alternative to simple harmonic motion and considered a better motion for the displacer of a Stirling engine but difficult to implement.

Phase angle: The angle difference between displacer and power piston harmonic motion with a complete cycle representing 360°. In most Stirling engines, the harmonic motion of the power piston follows (lags) the motion of the displacer by approximately 90°.

Push rod: A thin rod connected to the back of the piston that transfers linear motion through a dynamic linear seal, between the low- and high-pressure regions of an engine.

Reflux: A constant-temperature heat-exchange process where a liquid is evaporated by heat addition and then condensed as a result of cooling.

Regenerator: A heat-transfer device that stores heat from the working gas during part of a thermodynamic cycle and returns it during another part of the cycle. In the Stirling cycle the regenerator stores heat from one constant-volume process and returns it in the other constant-volume process.

Ringbom configuration: A Stirling engine configuration where the power piston is kinematically connected to a power shaft, and the displacer is a free piston that is powered by the difference in pressure between the internal gas and atmospheric pressure.

Simple harmonic motion: Linear motion varying with time as a sine function. Approximated by a piston connected to a crankshaft or a bouncing of a spring mass system.

Stirling cycle: A thermodynamic power cycle with two constant-volume heat addition and rejection processes and two constant-temperature heat-addition and rejection processes.

Swash plate drive: A disk on a shaft, where the plane of the disk is tilted away from the axis of rotation of the shaft. Piston push rods that move parallel to the axis of rotation but are displaced from the axis of rotation, slide on the surface of the rotating swash plate and therefore move up and down.

Variable-angle swash plate drive: A swash plate drive where the tilt angle between the drive shaft and the plate can be varied, resulting in a change in the displacement of the push rods.

Wobble plate drive: Another name for a swash plate drive.

Working gas: Gas within the engine that exhibits pressure and temperature change as it is heated or cooled and/or compressed or expanded.

References

Meijer, R.F. 1992. Stirling engine, in *McGraw-Hill Encyclopedia of Science and Technology*, 7th ed., pp. 440–445, McGraw-Hill, New York.

Stine, W.B. and Diver, R.E. 1994. A *Compendium of Solar Dish Stirling Technology,* Report SAND94-7026, Sandia National Laboratories, Albuquerque, NM 87185.

Walker, G. 1973. The Stirling engine, *Sci. Am.,* 229(2):80–87.

Walker, G. 1980. *Stirling Engines,* Clarendon Press, Oxford.

West, C.D. 1986. *Principles and Applications of Stirling Engines,* Van Nostrand Reinhold, New York.

Further Information

Books

Hargraves, C.M. *The Philips Stirling Engine,* Elsevier Press, London, 1991.

Organ, A.J. *Thermodynamics and Gas Dynamics of the Stirling Cycle Machine,* Cambridge University Press, Cambridge, 1992.

Senft, J.R. *Ringbom Stirling Engines,* Oxford University Press, Oxford, 1993.

Stine, W.B. and R.E. Diver, A *Compendium of Solar Dish/Stirling Technology,* SAND93-7026, Sandia National Laboratory, Albuquerque, 1994.

Urieli, I. and D.M. Berchowitz, *Stirling Cycle Engine Analysis,* Adam Hilger, Bristol, 1984

Walker, G. *Stirling Engines,* Clarendon Press, Oxford, 1980.

Walker, G. and J.R. Senft, *Free-Piston Stirling Engines,* Springer-Verlag, New York, 1985.

Walker, G., G. Reader, O.R. Fauvel, E.R. Bingham, *The Stirling Alternative,* Bordon & Breach, New York, 1994.

West, C.D. *Principles and Applications of Stirling Engines,* Van Nostrand Reinhold, New York, 1986.

Periodicals

Proceedings of the Intersociety Energy Conversion Engineering Conference (IECEC), published annually.

Stirling Machine World, a quarterly newsletter devoted to advancements in Stirling engines, edited by Brad Ross, 1823 Hummingbird Court, West Richland, WA 99353-9542.

Stirling Engine Developers

Stirling Technology Company, 4208B W. Clearwater Ave., Kennewick, WA 99336.

Stirling Thermal Motors, 275 Metty Drive, Ann Arbor, MI 48103.

Sunpower Incorporated, 6 Byard Street, Athens, OH 45701.

Clever Fellows Innovation Consortium, 302 Tenth St., Troy, NY 12180.

Mechanical Technologies Inc., 968 Albany-Shaker Rd., Latham, NY 12110.

Solo Kleinmotoren GmbH, Postfach 60 0152; D-71050 Sindelfingen; Germany.

Aisin-Seiki Ltd., 1, Asahi-Mach: 2-chome; Kariya City Aich: Pref 448; Japan.

8.6 Advanced Fossil Fuel Power Systems

Anthony F. Armor

Introduction

The generation of electric power from fossil fuels has undergone significant and, in some cases, dramatic changes over the last 20 years or so. Technology improvements in fossil fuel combustion have been largely driven by environmental issues, by the need to conserve fossil fuel resources, and by the economics of the competitive marketplace. The importance of fossil fuel-fired electric generation to the world is undeniable — more than 70% of all power in the U.S. is fossil fuel based, and worldwide the percentage is higher and growing. Today, most power plants worldwide burn coal, but increasingly generating companies are turning to natural gas, as the cost of gas-fired generation and the long-term supply of gas appear favorable. This section reviews the current status and likely future deployment of competing generation technologies based on fossil fuels.

It is likely, particularly in the developed world, that gas turbine-based plants will dominate the new generation market in the immediate future. The most advanced combustion turbines (CTs) now achieve more than 40% **lower heating value** (LHV) efficiency in simple cycle mode and greater than 50% efficiency in **combined cycle** (CC) mode. In addition, combustion turbine/combined cycle (CT/CC) plants offer siting flexibility, swift construction schedules, and capital costs between $400/kW and $800/kW. These advantages, coupled with adequate natural gas supplies and the assurance, in the longer term, of **coal gasification** backup, make this technology currently the prime choice for green field and repowered plants in the United States and in Europe.

But for the developing world, particularly China and India, there is good reason why the direct coal-fired power plant may still be the primary choice for many generation companies. Fuel is plentiful and inexpensive, and sulfur dioxide scrubbers have proved to be more reliable and effective than early plants indicated. In fact, as high as 99% SO_2 removal efficiency is now possible.

Combustion of coal can occur in three basic forms, direct combustion of pulverized coals (PC), combustion of coal in a suspended bed of coal and inert matter, and coal gasification. The pulverized coal (PC) plant, the most common form of coal combustion, has the capability for much improved efficiency even with full **flue gas desulfurization** (FGD), because materials technology has now advanced to the point where higher steam pressures and temperatures are possible. In the United States, Europe, Japan, and Russia, the advanced superficial PC plant is moving ahead commercially.

Worldwide, the application of atmospheric and pressurized fluidized bed combustion (FBC) plants has increased, and such plants offer reductions in both SO_2 and NO_x while permitting the efficient combustion of vast deposits of low-rank fuels such as lignites. In the United States, there are now over 150 large operating units for power generation, and throughout Europe, China, and the former Soviet Union countries small FBC units have been extensively deployed.

Coal gasification power plants exist at the 100 and 160 MW levels and are planned up to 450 MW. Much of the impetus is now coming from the U.S. Department of Energy (DOE) clean coal program where three gasification projects are in progress and four more are planned, and gasification plants are under construction in Europe and Japan. Gasification not only leads to minimum atmospheric and solid emissions, but also provides an opportunity to take advantage of gas turbine advances. With the rapid a, advances now being introduced in CT technology, the coal gasification option is a leading turn-of-the-century candidate for new plant construction.

Clean Coal Technology Development

At an increasing rate in the last few years, innovations have been developed and tested that are aimed at reducing emissions through improved combustion and environmental control, in the near term, and, in the longer term, through fundamental changes in the way coal is preprocessed before converting its

chemical energy to electricity. Such technologies are referred to as "clean coal technologies" described as a family of precombustion, combustion/conversion, and postcombustion technologies (Torrens, 1990). They are designed to provide the coal user with added technical capabilities and flexibility and the world with an opportunity to exploit our most abundant fossil source. They can be categorized as

- *Precombustion*, where sulfur and other impurities are removed from the fuel before it is burned;
- *Combustion*, where techniques to prevent pollutant emissions are applied in the boiler while the coal burns;
- *Postcombustion*, where the flue gas released from the boiler is treated to reduce its content of pollutants;
- *Conversion*, where coal, rather than being burned, is changed into a gas or liquid that can be cleaned and used as a fuel.

Coal Cleaning

Cleaning of coal to remove sulfur and ash is well established in the United States with more than 400 operating plants, mostly at the mine. Coal cleaning removes primarily pyritic sulfur (up to 70% SO_2 reduction is possible) and in the process increases the heating value of the coal, typically about 10% but occasionally 30% or higher. Additionally, if slagging is a limiting item, increased megawatts may be possible, as occurred at one station which increased generation from 730 to 779 MW. The removal of organic sulfur, chemically part of the coal matrix, is more difficult, but may be possible using microorganisms or through chemical methods; research is underway (Couch, 1991). Finally, heavy metal trace elements can be removed also, conventional cleaning removing (typically) 30 to 80% of arsenic, mercury, lead, nickel, antimony, selenium, and chromium.

Cleaning of Low-Rank Coal

With large deposits of high-moisture, and sometimes high-ash, low-rank coals and lignites, there is interest, but as yet no large-scale activity, in cleaning these coals. Improvement in heating value and reduction of boiler **slagging and fouling** problems are outcomes. Economics will decide whether precleaning or direct combustion (perhaps in a fluidized bed) will be the future choice. Large subbituminous and lignite fields exist in the United States, the former U.S.S.R., Germany, and Eastern Europe, as Table 8.6.1 indicates.

TABLE 8.6.1 World Coal Production

Country	Total Coal Production, Mt	Bituminous, Mt	Subbituminous and Lignite, Mt	Approximate Amount Bituminous Cleaned, %
China	1018	985	33	20
United States	833	581	252	55
Former U.S.S.R.	760	550	210	60
Germany	487	77	410	95
Poland	266	193	73	40
Australia	224	179	45	75
South Africa	214	214	0	—
India	191	180	11	20
Czechoslovakia	126	26	100	—
U.K.	100	100	0	75
Canada	61	33	28	—

Source: World Energy Conference, 1989.

Pulverized Coal Plants

There has been a perception that the PC power plant has come to the end of the road, that advanced coal technologies will quite soon make obsolete the PC plant with a scrubber, whose efficiency hovers around 35%. This perception may be premature. In fact, the PC plant has the capability for much improved heat rate (about 8500 Btu/kWh) even with full FGD. Beyond these units, the PC-fired CC with topping turbine (Figure 8.6.1) has a projected heat rate of 7200 Btu/kWhr, which includes full scrubbing capability. The PC plant is a proven, reliable power source with unit capabilities demonstrated at more than to 1000 MW using a single-shaft steam turbine. One plant, commercially available now, uses steam at 4500 psig and 1100°F, all ferritic materials for major boiler and turbine components, leading-edge technology in environmental controls, and the latest techniques in waste heat utilization (Poe et al., 1991). It is modular, fuel flexible, and designed for **on/off cycling capability**.

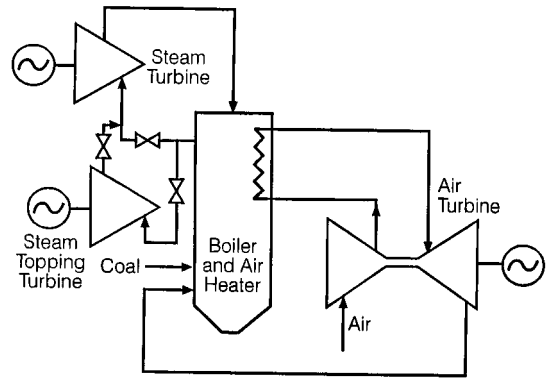


FIGURE 8.6.1 A PC CC with topping steam turbine has a projected heat rate of 7200 Btu/kWhr. The air turbine uses 1800°F air, or 2300°F air with supplemental firing. The topping turbine uses steam at 1300°F.

Higher steam temperatures (to 1150°F) and **supercritical steam pressures** are an important aspect of the modern advanced PC plant. They are possible now because of advances in ferritic materials technology that will extend life, provide greater creep and fatigue strength, and be resistant to **temper embrittlement** and, in the boiler, to coal ash corrosion (Armor et al., 1988).

Of particular note are

- **Coextruded tubing** or monotubing for superheaters and reheaters, resistant to coal ash corrosion.
- Super-9-chrome steel (P91), for steam piping, valves, headers, casings.
- Improved creep-resistant 12-chrome forgings for high-pressure/intermediate-pressure (HP/IP) turbines.
- "Superclean" 3.5 NiCrMoV rotors for low-pressure (LP) turbines, resistant to temper embrittlement.

Built in 1959, Eddystone 1 at PECO Energy was, and still is, the supercritical power unit with the highest steam conditions in the world. When constructed in the early 1960s, Eddystone 1 had a main steam pressure of 5000 psi, and main steam temperature of 1200°F. **Double reheat** of the steam was employed. PECO Energy continues to operate Eddystone 1, an impressive achievement for a prototype unit. More-recent advanced plants include the Chuba Electric Kawagoe unit in Japan, a 700-MW double-reheat supercritical with steam conditions of 4750 psi, 1050°F, and the Esbjerg unit of Elsam in Denmark, a 400-MW supercritical with steam conditions of 3700 psi, 1040°F. Both plants use advanced ferritic steels for turbine and boiler thick wall components. There are over 170 supercritical units in the United States, more than 220 in Russia (Oliker and Armor 1992), and over 60 in Japan. A number are installed in Germany, Denmark, Holland, and other European countries, and increasingly in the Far East.

Emissions Controls for Pulverized Coal Plants

Today, worldwide, about 40% of electricity is generated from coal and the total installed coal-fired generating capacity is more than 1000 GW, largely made up of 340 GW in North America; 220 GW in Western Europe, Japan, and Australia; 250 GW in Eastern Europe and the former U.S.S.R., and 200 GW in China and India. In the decade 1990 to 2000, 190 GW of new coal-fired capacity will likely be added. So the control of particulates, sulfur dioxides, and nitrogen oxides from those plants is one of the most pressing needs of today and of the future, together with the potential impact of carbon dioxide emissions, with their contribution to global warming. To combat these concerns, a worldwide move toward environmental retrofitting of older fossil-fired power plants is underway, focused largely on sulfur dioxide scrubbers and combustion or postcombustion optimization for nitrogen oxides.

Conventional Lime/Limestone Wet Scrubber

The dominant SO_2 scrubbing system is the wet limestone design, limestone being one quarter the cost of lime as a reagent. In this system (Figure 8.6.2) the limestone is ground and mixed with water in a reagent preparation area. It is then conveyed to a spray tower called an absorber, as a slurry of 90% water and 10% solids, and sprayed into the flue gas stream. The SO_2 in the flue gas is absorbed in the slurry and collected in a reaction tank where it combines with the limestone to produce water and calcium sulfate or calcium sulfate crystals. A portion of the slurry is then pumped to a thickener where these solids/crystals settle out before going to a filter for final dewatering. Mist eliminators installed in the system ductwork at the spray tower outlet collect slurry/moisture entrained in the flue gas stream. Calcium sulfate is typically mixed with fly ash (1:1) and lime (5%) and disposed of in a landfill.

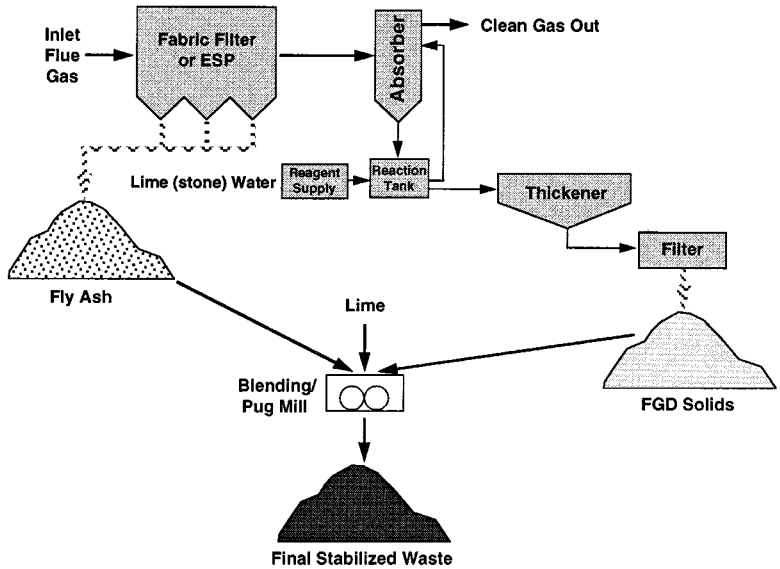


FIGURE 8.6.2 The conventional lime/limestone wet scrubber is the dominant system in operation in the United States. With recent refinements this system can be 98 to 99% effective in removing SO_2.

Various improvements can be made to this basic process, including the use of additives for performance enhancement and the use of a hydrocyclone for dewatering, replacing the thickener, and leading to a salable gypsum by-product. The Chiyoda-121 process (Figure 8.6.3) reverses the classic spray scrubber and bubbles the gas through the liquid. This eliminates the need for spray pumps, nozzle headers, separate oxidation towers, and thickeners. The Chiyoda process is being demonstrated in a DOE Clean Coal Technology (CCT) project on a 100-MW unit at the Yates plant of the Georgia Power Company.

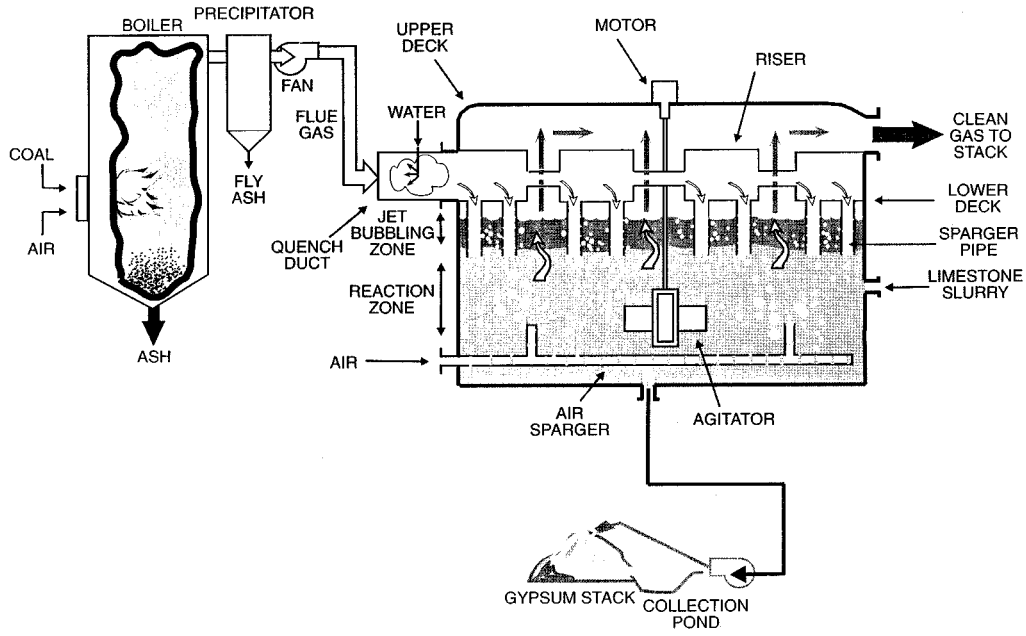


FIGURE 8.6.3 The Chioda-121 scrubber simplifies the process by bubbling the flue gas through the liquid, eliminating some equipment needs.

Spray Drying

Spray drying (Figure 8.6.4) is the most advanced form of dry SO_2 control technology. Such systems tend to be less expensive than wet FGD but remove typically a smaller percentage of the sulfur (90% compared with 98%). They are usually used when burning low-sulfur coals, and utilize fabric filters for particle collection, although recent tests have shown applicability to high-sulfur coals also.

Spray driers use a calcium oxide reagent (quicklime) which when mixed with water produces a calcium hydroxide slurry. This slurry is injected into the spray drier where it is dried by the hot flue gas. As the drying occurs, the slurry reacts to collect SO_2. The dry product is collected at the bottom of the spray tower and in the downstream particulate removal device where further SO_2 removal may take place. It may then be recycled to the spray drier to improve SO_2 removal and alkali utilization.

For small, older power plants with existing **electrostatic precipitators** (ESPs), the most cost-effective retrofit spray-dry configuration locates the spray drier and fabric filter downstream of the ESP, separating in this manner the spray drier and fly ash waste streams. The fly ash can then be sold commercially.

Control of Nitrogen Oxides

Nitrogen oxides can be removed either during or after coal combustion. The least-expensive option and the one generating the most attention in the United States is combustion control, first through adjustment of the fuel/air mixture and second through combustion hardware modifications. Postcombustion processes seek to convert NO_x to nitrogen and water vapor through reactions with amines such as ammonia and urea. Selective catalytic reduction (SCR) injects ammonia in the presence of a catalyst for greater effectiveness. So the options (Figure 8.6.5) can be summarized as

Operational changes. Reduced excess air and biased firing, including taking burners out of service;
Hardware combustion modifications. Low-NO_x burners, air staging, and fuel staging (reburning);
Postcombustion modifications. Injection of ammonia or urea into the convection pass, SCR, and wet or dry NO_x scrubbing (usually together with SO_2 scrubbing).

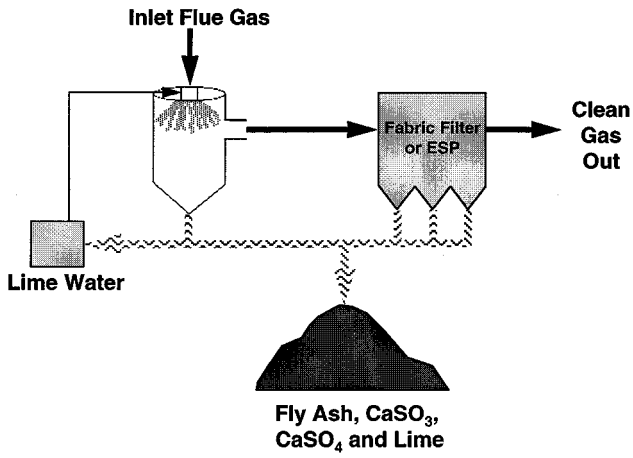


FIGURE 8.6.4 Spray driers use a calcium oxide reagent mixed with water, which is dried by the flue gas. A dry product is collected in a fabric filter.

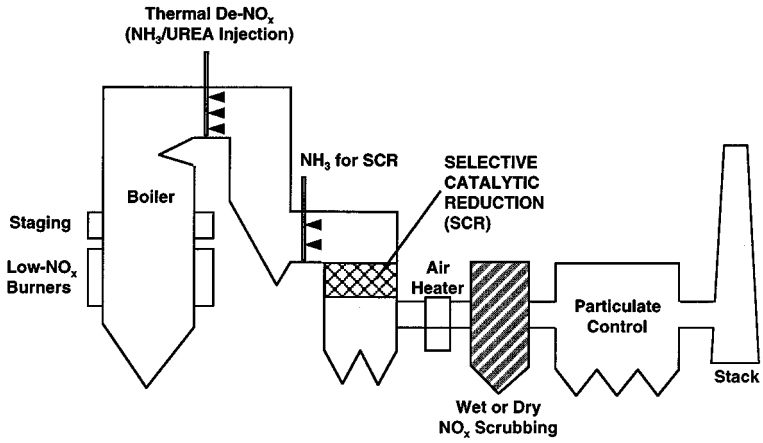


FIGURE 8.6.5 Control options for NO_x include operational, hardware, and postcombustion modifications.

Low-NO_x burners can reduce NO_x by 50% and SCR by 80%, but the low-NO_x burner option is much more cost-effective in terms of cost per ton of NO_x removed. Reburning is intermediate in cost per removed ton and can reduce NO_x 50 or 75% in conjunction with low-NO_x burners.

Fluidized Bed Plants

Introduced nearly 30 years ago, the **fluidized bed** combustion boiler has found growing application for power generation. From the first FBC boiler, generating 5000 lb/hr of steam in 1967, the technology has matured to the 250-MW-size units available today. In North America more than 170 units now

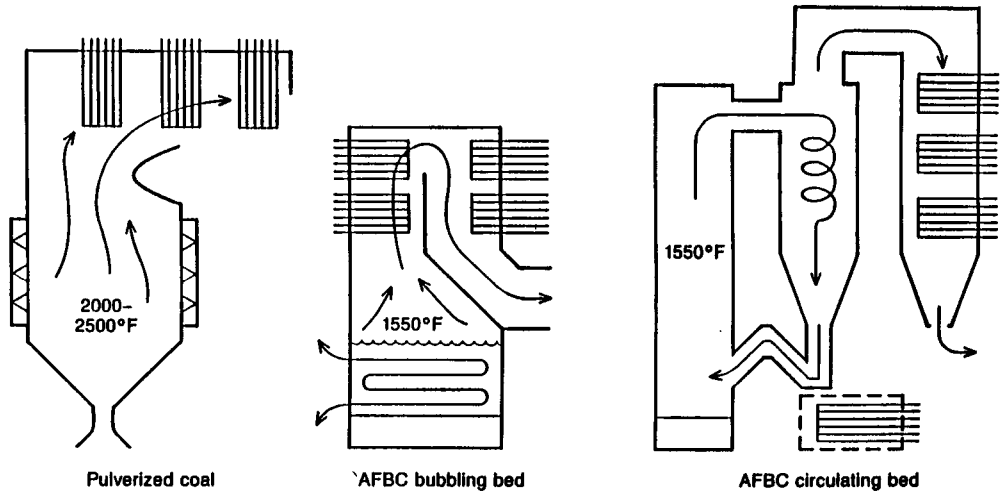


FIGURE 8.6.6 An illustration of the distinguishing features of PC and fluidized bed boilers. Noticeable in this diagram are the in-bed tubes characteristic of bubbling beds and the cyclone separator of the circulating bed.

generate in excess of 6000 MW. Burning coal in a suspended bed with limestone or dolomite permits effective capture of sulfur, and fuel flexibility allows a broad range of opportunity fuels. These fuels might include coal wastes (culm from anthracite, gob from bituminous coal), peat, petroleum coke, and a wide range of coals from bituminous to lignite. A low (1500°F) combustion temperature leads to low NO_x formation. The salient features of atmospheric fluidized bed boilers, compared with a PC boiler, are shown in Figure 8.6.6.

Utility-size demonstration projects at the Tennessee Valley Authority in 1989 (Shawnee, 160 MW) (Manaker, 1992) and Northern States Power in 1986 (Black Dog, 133 MW) (Hinrichsen, 1989) are examples of successful atmospheric bubbling bed units. The Black Dog unit has been dispatched in a daily cycling mode and has successfully fired a blend of coal and petroleum coke. But the focus of atmospheric FBC (AFBC) in the United States is now on the circulating fluid bed (CFB). In fact, more than 70% of operating fluid bed boilers in the United States are of the circulating type. The CFB unit at Nucla (Tri-State G&T Association) (Blunder, 1989) has been successful in demonstrating the technology at the 110-MW level, and commercial CFB plants have now reached 250 MW in size. Most fluidized bed units for electricity generation are being installed by independent power producers in the 50- to 100-MW-size range, where the inherent SO_2 and NO_x advantages over the unscrubbed PC plant have encouraged installations even in such traditional non-coal arenas as California (Melvin and Friedman, 1994). Worldwide, the AFBC boiler is employed largely for steam heat, with hundreds of them in operation in Russia and India, and thousands in China. The extension of the concept of fluidized beds to units where the fuel mixture is burned under several atmospheres pressure (PFBC) has now opened the way to smaller modular units with opportunities to move to efficient PFBC CCs.

Atmospheric Fluidized Bed Combustion

In the bubbling bed version of the AFBC, the fuel and inert matter, together with limestone or dolomite for SO_2 capture, is suspended through the action of fluidizing air, which flows at a velocity of 3 to 8 ft/sec in essentially a one-pass system. CFBs differ from bubbling beds in that much of the bed material passes through a cyclone separator before being circulated back to the boiler (Figure 8.6.7). In-bed tubes are generally not used for CFB units, permitting a much higher fluidizing velocity of 16 to 26 ft/sec. Since the early AFBC designs, attention has been directed toward increasing unit efficiency, and reheat designs are now usual in large units. When SO_2 capture is important, a key parameter is the ratio of calcium in the limestone to sulfur in coal. Typical calcium-to-sulfur ratios for 90% SO_2 reduction are

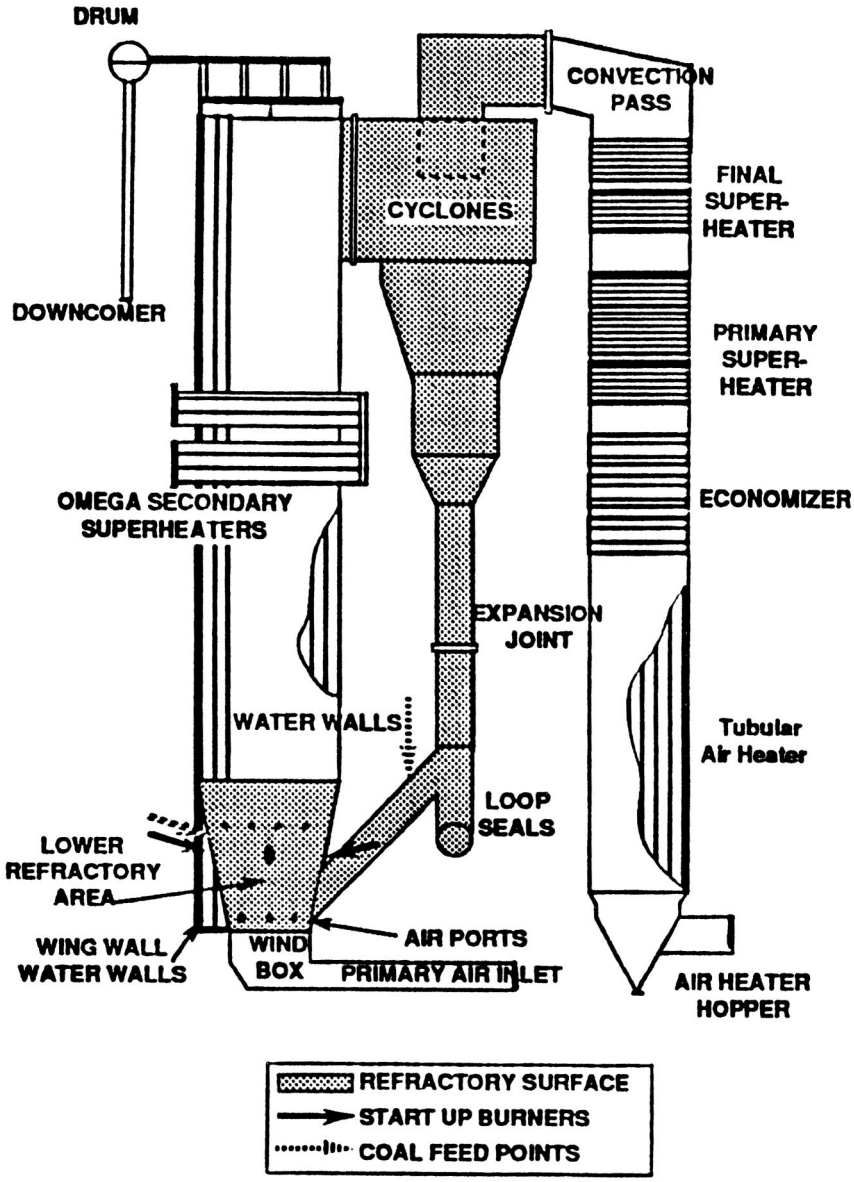


FIGURE 8.6.7 A Pyropower circulating fluid bed boiler installed at the ACE Cogeneration Company at Trona, California. This 108-MW unit burns low sulfur, western bituminous coal with limestone in a bed which circulates back to the boiler after passing through a cyclone separator.

in the range of 3.0 to 3.5 for bubbling beds and 2.0 to 2.5 for circulating beds. This depends on the fuel, however, and the 200-MW CFB units at the Conoco/Entergy plant in Lake Charles, Louisiana burning 100% **petroleum coke** (4.5% S), have a Ca/S ratio of below 1.7 for more than 90% sulfur capture. NO_x levels in AFBCs are inherently low and nominally less than 0.2 lb/MM Btu.

It is important to note that for CFBs, boiler efficiencies can be as high as a PC unit (Table 8.6.2). In fact, designs now exist for AFBCs with supercritical steam conditions, with prospects for cycles up to 4500 psia, 1100°F with **double reheat** (Skowyra et al., 1995).

TABLE 8.6.2 Typical Boiler Efficiencies, PC and Fluidized Beds

		Calculated Heat Loss %		
Loss/Gain Parameter	**PC**	**Highest-Efficiency CFB**	**Lowest-Efficiency CFB**	**Bubbling Bed**
Moisture in limestone	NA	0.06	0.10	0.10
Calcination	NA	1.02	1.69	2.70
Sulfation credit	NA	–1.60	–1.60	–1.60
Unburned carbon	0.25	0.50	2.0	4.0
Heat in dry flue gas	5.28	5.57	5.60	5.75
Moisture in fuel	1.03	1.03	1.03	1.03
Moisture from burning H_2	41.9	4.19	4.19	4.19
Radiation and convection	0.30	0.30	0.80	0.30
Moisture in air	0.13	0.14	0.14	0.14
Sensible heat in boiler ash	0.03	0.09	0.76	0.50
Bottom ash	0.05	NA	NA	NA
Fan-power credit	–0.25	–0.75	–0.40	–0.50
Pulverizer/crusher power gain	–0.20	NA	NA	NA
Total losses/gains	10.81	10.55	14.31	16.51
Overall boiler efficiency, %	89.19	89.45	85.69	83.49

Source: POWER, January 1987.

Pressurized Fluidized Bed Combustion

In a PFBC CC unit (Figure 8.6.8), coal in a fluid bed is burned with dolomite or limestone in a pressurized steel chamber, raising steam for a steam turbine generator. The pressurized flue gases are expanded through a gas turbine. Commercial plants at about the 80-MW level in Sweden, the United States, and Spain have demonstrated that bubbling bed PFBC plants with a calcium-to-sulfur molar ratio of about 1.5 offer sulfur capture up to 95%, together with inherently low NO_x emissions due to low combustion temperatures. Cleanup of the flue gas before entry to the gas turbines is a key technical objective, and first-generation units have used cyclones together with gas turbines ruggedized with special blade coatings. For more-advanced, higher-efficiency PFBC systems, hot-gas cleanup technology, where the gas is directed through large ceramic filter units, will likely be needed.

To date, the 80-MW units at Vaertan (Sweden) and Escatron (Spain) and the 70-MW unit at Tidd (AEP) have operated satisfactorily, and larger units up to 350 MW are now under development. The modular aspect of the PFBC unit is a particularly attractive feature leading to short construction cycles and low-cost power. This was particularly evident in the construction of the Tidd plant, which first generated power from the CC on November 29, 1990. The heat rate and capital cost of the PFBC plant are forecast to reach very competitive levels which, when combined with shortened construction schedules, will position the technology for a role in future generation plans, particularly where modular additions have advantages.

One promising use for PFBC units is for small in-city cogeneration plants where the inherent size advantages, high efficiencies, and effective coal gas cleanup approach permit compact plants to be retrofitted in place of heating boilers, while the small steam turbines can be easily adapted to both electricity and hot water supply (Olesen, 1985).

Advanced PFBCs

In conventional PFBC plants, the overall cycle efficiency is limited to less than 42%. That is because the operating temperature of the combustor — which must be held to 1650°F (900°C) or less to avoid sintering the ash and releasing alkali metals that could foul or corrode the gas turbine — in effect sets the gas turbine inlet temperature far below the 2350°F (1290°C) or so featured in the most-efficient heavy-frame machines currently in use firing natural gas or oil.

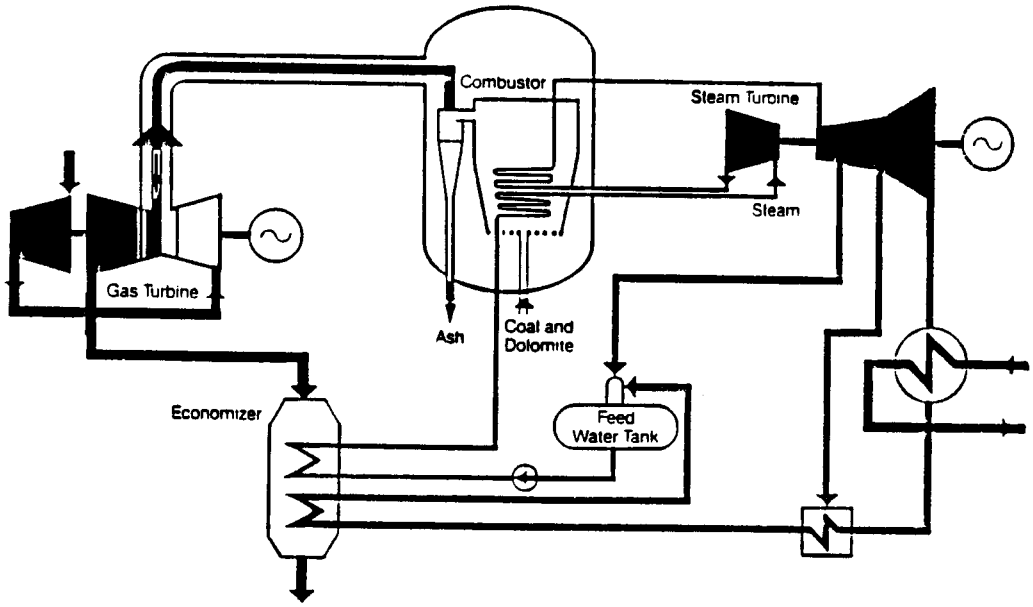


FIGURE 8.6.8 Pressurized fluidized bed with combined cycle. This 70-MW system has operated at the Tidd plant of American Electric Power.

Increasing the inlet temperature of the gas turbine, which typically provides 20 to 25% of the power in a PFBC plant, could raise the overall cycle efficiency to more than 45%. A simple means of boosting the flue gas temperature would be to fire a topping fuel, such as natural gas, ahead of the gas turbine. But 2350°F would be well above the **ash-softening temperature**, so high temperature, HP filtration systems would be essential to remove all particulate matter before firing the natural gas.

A further advance would be to use gas from coal rather than natural gas as the topping fuel. The coal would be pyrolyzed in a low-oxygen environment under pressure to produce both a low-Btu fuel gas and a residual char. The fuel gas would be passed through its own hot-gas cleanup filters before being fired in a topping combustor ahead of the gas turbine; the char would be burned in a circulating bed PFBC, from which the flue gas would also be filtered and then combined with the topping cycle gas stream. (Or the char could be simply fired in an AFBC.) With net heat rates below 7600 Btu/kWhr (45% efficiency), carbon dioxide emissions would be correspondingly low.

A U.S. DOE CCT project known as the Four Rivers Energy Modernization Project, involves building a Foster-Wheeler 95-MW advanced circulating-type PFBC unit at the Air Products chemical manufacturing facility in Calvert City, Kentucky. Steam from the unit, which will feature a coal-gas-fired topping combustor and a hot-gas cleanup system, will be used in chemical production, and the power will be sold to the Tennessee Valley Authority. The project is scheduled to begin commercial operation in October, 1998 (Carpenter and Dellefield, 1994).

Gasification Plants

One option of growing interest to coal-burning utilities is that of coal gasification. After the EPRI Cool Water demonstration in 1984 at the 100-MW level, the technology has moved ahead in the United States largely through demonstrations under the CCT program (U.S. DOE, 1994). Overseas, the 250-MW Buggenham plant in Holland is now operational, and the PSI/Destec 262-MW and TECO Energy 250-MW gasification plant demonstrations are also on-line. Beyond this, there is a 300-MW gasification unit scheduled for Endesa, Spain and a 330-MW unit for RWE in Germany (Figure 8.6.9).

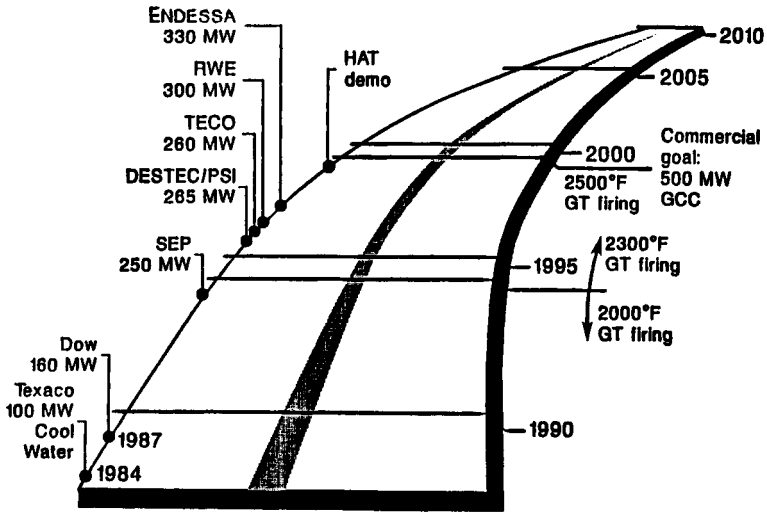


FIGURE 8.6.9 By building on the early success of the 100-MW Cool Water gasification-CC plant in California, demonstrations in the 250 to 350 MW range will be carried out in the 1995 to 2000 time frame.

Gasification-based plants have among the lowest emissions of pollutants of any central station fossil technology. Through use of the efficiency advantages of CCs, CO_2 emissions are also low. Fuel flexibility is an additional benefit since the gasifier can accommodate a wide range of coals, plus petroleum coke. Integrated gasification CC (IGCC) plants permit a hedge against long-term increases in natural gas prices since natural gas-fired CTs can be installed initially, and gasifiers at a later time when a switch to coal becomes prudent (Douglas, 1986).

The pioneering Cool Water plant, the first of its kind in the world, operated for more than 4 years, gasifying 1.1 million tons of coal and producing 2.8 million MWhr of electricity. The project was a collaborative effort of the industry involving the utility (Southern California Edison), equipment manufacturers (Texaco, General Electric), and consultants/research consortia (Bechtel, EPRI, and others). Particularly notable was the achievement of exceptionally low levels of emissions of SO_2, NO_x, and particulates, as shown in Figure 8.6.10.

Basically, IGCC plants replace the traditional coal combustor with a gasifier and gas turbine. Ultra-low emissions are realized, over 99% of the sulfur in the coal being removed before the gas is burned in the gas turbine. A gasification cycle can take advantage of all the technology advances being made in CTs and steam turbines, so as to enhance overall cycle efficiency. Net system efficiencies of 45% are expected to be demonstrated by the turn of the century, and when, in the next decade, the **fuel cell** begins to replace the gas turbine, plant efficiencies will climb to the 60% level. Major demonstrations are underway, as part of the CCT program, at Sierra Pacific Power, Pinon Pine (a KRW 99-MW air-blown, pressurized, fluidized bed coal gasifier); Tampa Electric, Polk County (a Texaco, 250-MW oxygen-blown, entrained-flow gasifier); at Tamco Power, Toms Creek (a 190-MVV joint gasification/PC plant with a Tampella Power fluidized bed gasifier); and at PSI Energy/Destec Energy, Wabash River (a 262-MW plant based on the Destec two-stage, entrained-flow, oxygen-blown gasifier). More-detailed descriptions of the Pinon Pine and Polk County gasification systems follow.

Pinon Pine IGCC

At the Sierra Pacific Power Pinon Pine plant (Figure 8.6.11), dried and crushed coal is introduced into a pressurized, air-blown, fluidized bed gasifier. Crushed limestone is added to the gasifier to capture a portion of the sulfur and to inhibit conversion of fuel nitrogen to ammonia. The sulfur reacts with the limestone to form calcium sulfide, which, after oxidation, exits along with the coal ash in the form of agglomerated particles suitable for landfill.

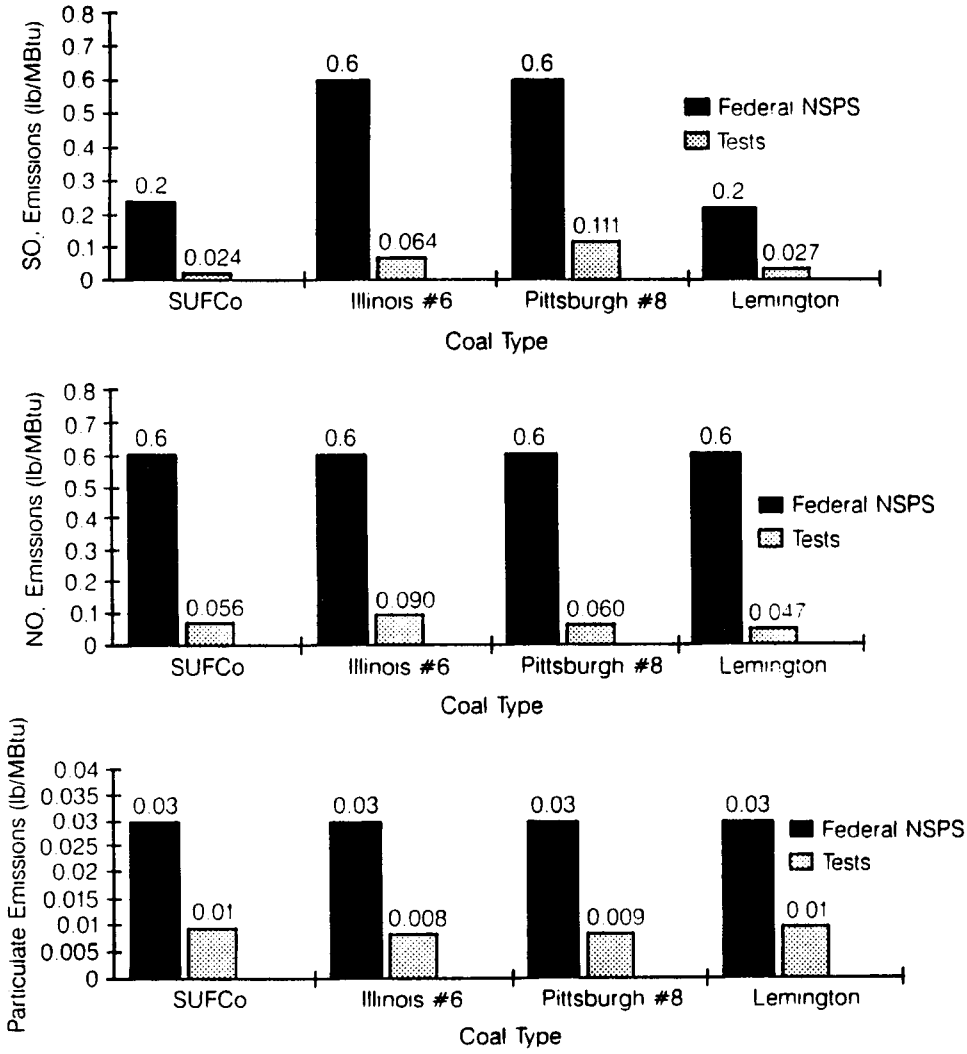


FIGURE 8.6.10 Tests at Cool Water on four coals show emissions of SO_2, NO_x, and particulates substantially below the Federal New Source Performance Standards.

Hot, low-Btu coal gas leaving the gasifier passes through cyclones, which return most of the entrained particulate matter to the gasifier. The gas, which leaves the gasifier at about 1700°F, is cooled to about 1100°F before entering the hot-gas cleanup system. During cleanup, virtually all of the remaining particulates are removed by ceramic candle filters, and final traces of sulfur are removed by reaction with metal oxide sorbent.

The hot, cleaned gas then enters the CT, which is coupled to a generator designed to produce 61 MW. Exhaust gas is used to produce steam in a heat-recovery steam generator. Superheated HP steam drives a condensing steam turbine/generator designed to produce about 46 MW.

Owing to the relatively low operating temperature of the gasifier and the injection of steam into the combustion fuel stream, the NO_x emissions are likely to be below 0.053 lb/M Btu. Because of the combination of in-bed sulfur capture and **hot-gas cleanup**, SO_2 emissions will be below 0.045 lb/M Btu (98% reduction).

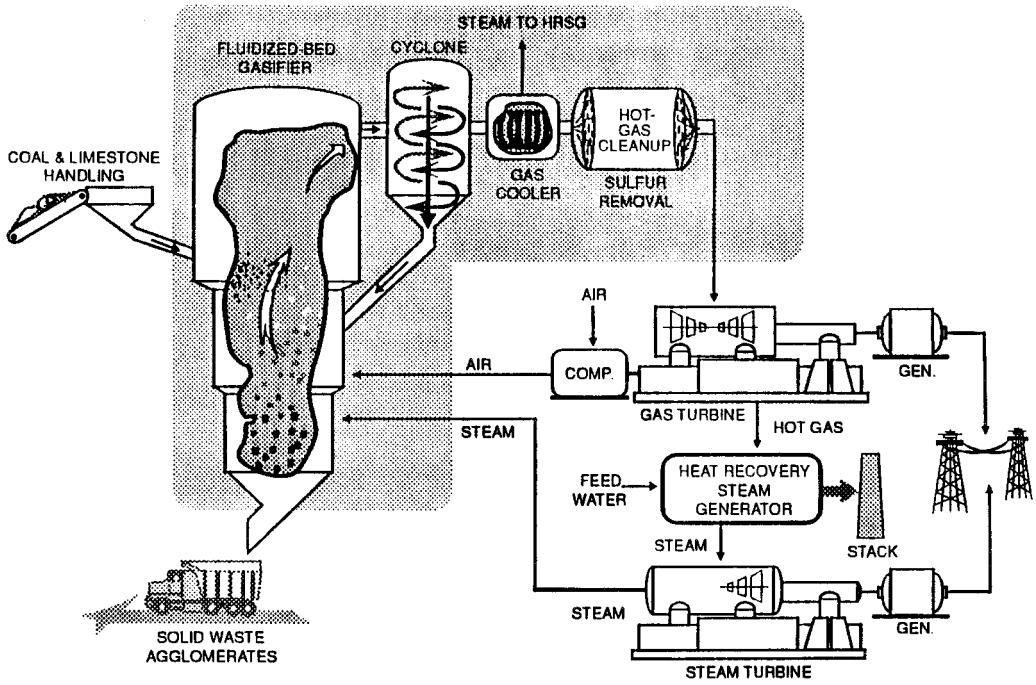


FIGURE 8.6.11 Integrated gasification CC at Sierra Pacific Power, Pinon Pine plant. This uses a KRW air-blown, pressurized fluidized bed coal gasifier, and produces 99 MW (net).

At Pinon Pine, 880 ton/day of coal are converted into 107 MW (gross), or 99 MW (net), for export to the grid. Western bituminous coal (0.5 to 0.9% sulfur) from Utah is the design coal although tests will be done using West Virginia or Pennsylvania bituminous coal containing 2 to 3% sulfur.

Polk County IGCC

The Texaco pressurized, oxygen-blown, entrained-flow gasifier will be used at the Tampa Electric Polk County plant to produce a medium-Btu fuel gas (Figure 8.6.12). Coal/water slurry and oxygen are combined at high temperature and pressure to produce a high-temperature syngas. Molten coal ash flows out of the bottom of the vessel and into a water-filled quench tank, where it is turned into a solid slag. The syngas from the gasifier moves to a high-temperature heat-recovery unit which cools the gases.

The cooled gases flow to a particulate-removal section before entering gas cleanup trains. A portion of the syngas is passed through a moving bed of metal oxide absorbent to remove sulfur. The remaining syngas is further cooled through a series of heat exchangers before entering a conventional gas cleanup train where sulfur is removed by an acid-gas removal system. These cleanup systems combined are expected to maintain sulfur levels below 0.21 lb/M Btu (96% capture). The cleaned gases are then routed to a CC system for power generation. A gas turbine generates about 192 MW. Thermally generated NO_x is controlled to below 0.27 lb/M Btu by injecting nitrogen as a diluent in the combustion section of the turbine. A heat-recovery steam generator uses heat from the gas turbine exhaust to reduce HP steam. This steam, along with the steam generated in the gasification process, is routed to the steam turbine to generate an additional 120 MW. The IGCC heat rate for this demonstration is expected to be approximately 8600 Btu/kWhr (40% efficient).

The demonstration project involves only the first 250-MW portion of the planned 1150-MW Polk County power station. Coals being used in the demonstration are Illinois 6 and Pittsburgh 8 bituminous coals having sulfur contents ranging from 2.5 to 3.5%.

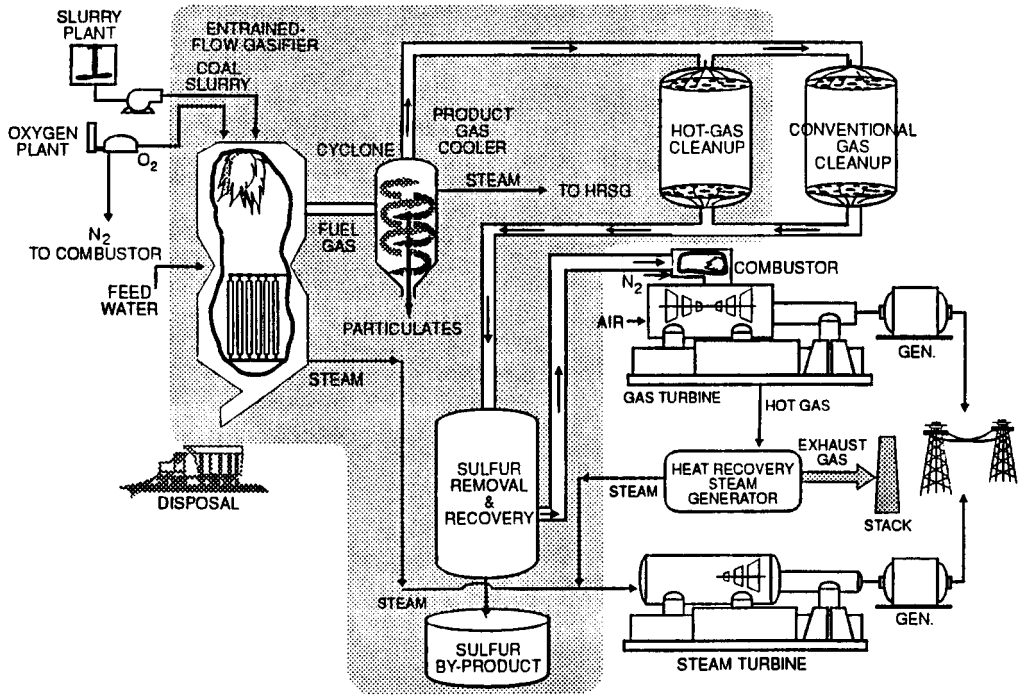


FIGURE 8.6.12 Integrated gasification CC at Tampa Electric, Polk County plant. A Texaco oxygen-blown gasifier is used. Total net generation is 250 MW.

By-products from the process — sulfuric acid and slag — can be sold commercially, sulfuric acid by-products as a raw material to make agricultural fertilizer and the nonleachable slag for use in roofing shingles and asphalt roads and as a structural fill in construction projects.

Buggenum IGCC

Meanwhile, tests are in progress on the 250-MW IGCC plant in Buggenum, Netherlands. After successful operations running on natural gas, a switch was made to coal gas using Columbia coals. Buggenum comprises a 2000-ton/day single reactor coal gasification unit and an air separation plant able to produce 1700 ton/day of 95% pure oxygen. Syngas drives a Siemens CC power unit, including a 156-MW V94.2 gas turbine and a 128-MW steam turbine. The gasifier, operating at 28 bar pressure and 2700°F is designed to produce syngas containing 42% nitrogen, 25% carbon monoxide, and 12% hydrogen, with a combustion value of 4.3 MJ/kg.

The environmental constraints are defined by permit requirements fixing upper limits of SO_2 at 0.22 g/kWhr, NO_x at 0.62g/kWhr, and particulates at 0.007 g/kWhr.

Key steps for limiting emissions include

- Removing fly ash with cyclone and candle filters after gas cooling;
- Removing halogens and other soluble pollutants with water scrubbing;
- Desulfurizing gas by catalytic and chemical processes. Sulfur is fixed in sulfinol-M solvent, which is further treated to produce elemental sulfur; and
- Desulfurized gas is mixed with nitrogen from the air separation units and saturated with water vapor to reduce its lower heating value from about 11,000 to 4300 kJ/kg greatly reducing NO_x production.

Combustion Turbine Plants

CT-based plants are the fastest growing technology in power generation. Through the turn of the century natural gas-fired CTs and CCs burning gas will account for 50 to 70% of the 900 to 1000 GW of new generation to be ordered worldwide. Almost all of these CT and CC plants will be gas fired, leading to a major expansion of gas for electricity generation (Armor et al., 1992).

It is likely that CTs and CCs will grow steadily more important in all generation regimes, peaking, midrange, and base load. If the present 2300°F firing temperature machines operate reliably and durably, CT and CC plants will begin to replace older steam plants and uneconomic nuclear plants. So until the emergence of large-scale fuel cells, CT plants will be a competitive choice for new fossil generation, and advanced CT cycles, with intercooling, reheat, possibly chemical recuperation, and most likely with humidification, will spearhead the drive to higher efficiencies and lower capital costs. Gasification, which guarantees a secure bridge to coal in the near term, will come into its own if natural gas prices rise under demand pressure, and by 2015 coal through gasification may be the economic fuel for a significant fraction of new base-load generation. The rate at which these trends develop depends in large measure on the speed of deregulation and advent of competition in the electricity industry.

Modern gas turbines for power generation are mostly heavy-frame machines, with ratings in a simple cycle configuration around 150 to 170 MW for the high firing temperatures (~2300°F) of the **"F-class" machines**. Efficiencies (LHV) are 36 to 38% in simple cycles. In CCs, the units are 220 to 350 MW in size and 53 to 55% efficient. The next generation of CTs, with efficiencies from 57 to 60% has recently been announced (Table 8.6.3). Smaller-scale aeroderivative machines have benefited from turbofan engines designed for wide-body aircraft and today are available in ratings of 35 to 65 MW and with efficiencies of 40% or more for turbine inlet temperatures around 2250°F. Beyond this, work is underway to design and build an advanced **aeroderivative turbine**, and three teams are involved: United Technologies/Fluor Daniel is looking at a humid air turbine at 200 MW and 55% efficiency; Rolls-Royce/Bechtel is studying an intercooled, regenerative design based on the Aero-Trent machine; and General Electric/Bechtel is melding the GE LM 6000 and the GE heavy-frame designs to achieve close to 60% efficiency in a 100-MW combined cycle.

TABLE 8.6.3 Modern Gas Turbine Specifications

		Large Heavy-Frame Machines			
		Simple Cycle		Combined Cycle	
Turbine		MW	Efficiency % (LHV)	MW	Efficiency % (LHV)
		Current			
GE	GE7FA/9FA	168/227	36.0	253/351	55.0
W/MHI	*W*501F/701F	164/237	38.1	236/337	53.7
Siemens	V84.3/94.3	154/222	36.2	227/328	54.5
		New			
ABB	GT24/26	165/240	37.6	250/365	57.2
Siemens	V84.3A/94.3A	170/240	38.0	254/359	57.0
W/MHI	*W*501G/701G	230/255	38.5	345/380	58.0
GE	GE7G/9G	240/282	39.5	350/420	58.0
GE	GE7H/9H	—	—	400/480	60.0

Recently, General Electric announced a new concept which uses steam from the steam turbine to cool the gas turbine blades permitting a firing temperature of 2600°F. This avoids the losses due to the normal method of diverting compressor air for cooling. This "H" class turbine, it is said, will break the 60% barrier for a 400-MW CC. Comparison of the F- and H-class machines for General Electric is shown in Table 8.6.4.

TABLE 8.6.4 Comparison of F- and H-Class Machines

	General Electric Advanced Machines			
	7FA	**7H**	**9FA**	**9H**
Characteristics				
Firing temperature °F(°C)	2350 (1300)	2600/1430	2350 (1300)	2600 (1430)
Air flow, lb/sec (kg/sec)	974 (442)	1230/558	1327 (602)	1510 (685)
Pressure ratio	15	23	15	23
Specific work, MW/lb/sec (MW/kg/sec)	0.26 (0.57)	0.33 (0.72)	0.26 (58)	0.32 (70)
Performance				
Simple cycle output, MW	168	—	227	—
Simple cycle efficiency, %	36	—	36	—
CC net output, MW	253	400	349	480
CC net efficiency, %	55	60	55	60
NOX (ppmvd at 15% O_2)	9	9	25	25

Source: GE Power Systems, *Power System for the 21st Century: H Gas Turbine Combined Cycle*, 1995. With permission.

Humidified Air Power Plants (Cohn, 1994)

A new class of CTs has been designed based on humidifying the combustion air. In these combustion turbine cycles the compressor exit air is highly humidified prior to combustion. This reduces the mass of dry air needed and the energy required to compress it, raising plant efficiency.

The continuous plant cycle for this concept is termed the **humid air turbine** (HAT). This cycle has been calculated to have a heat rate for natural gas about 5% better than current high-technology CCs. The HAT cycle is adaptable to coal gasification leading to the low emissions and high-efficiency characteristics of gasification CC plants but at a low capital cost since the steam turbine bottoming cycle is eliminated. A simple humidified air turbine cycle is shown in Figure 8.6.13. The addition of moisture means that perhaps 25% more mass flow goes through the turbine, than through the compressor. This suggests the use of separate spools for the turbine and compressor. By using present-day 2350°F firing temperatures it is reasonable to expect a HAT heat rate of about 6100 Btu/kWhr from this cycle.

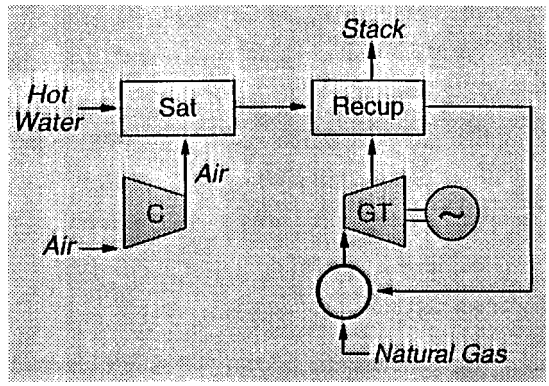


FIGURE 8.6.13 The HAT cycle adds moisture to the compressor exit air, reducing the air mass flow needed and increasing cycle efficiency.

As noted above, the ideal natural gas-fired HAT plant has been calculated to have higher efficiency (about 2 points higher) than a CC for the same turbine cooling technology. Thus, it would provide the lowest heat rate for a natural gas-fired thermal plant and would be utilized in base-load or long intermediate dispatch. The capital cost of this power plant has been calculated to be only slightly higher

than that of a CC. However, the anticipated development cost for the ideal turbomachinery has been estimated to be very high, in excess of $250 million.

In contrast, the CHAT (cascaded humid air turbine) plant utilizes turbine components, which are now available, with few exceptions, in a cascade arrangement that allows them to match together. The development cost of the CHAT equipment is currently estimated to be only in the $5 to 10 million range, making its development much more practical.

The HAT and CHAT cycles can be integrated with gasification. Because these cycles directly incorporate humidification, they can make direct use of hot water generated in the gasification plant, but cannot readily utilize steam. Thus, they match well with the lower-capital-cost, but lower-efficiency, quench types of gasifier. This provides an overall power plant with efficiency about the same as an IGCC. Moreover, the capital cost of the IGHAT plant has been calculated to be about $150/kW less than an IGCC plant. These humidification cycles have yet to be offered commercially. The main obstacle is the need to demonstrate experimentally low-emission, high-efficiency, full-scale combustors utilizing very humid air.

Other Combustion Turbine Cycle Enhancements

There are several variants of the CT-based Brayton cycle which increase plant efficiency and capacity (Lukas, 1986). **Regenerative cycles** use storage-type heat exchangers, where porous or honeycomb wall structures store energy from the hot gases. This is released later to the cold gases. A **recuperative cycle** uses a heat exchanger where the hot and cold streams are separated by walls through which heat transfer occurs. This is the approach commonly used in CTs allowing gains in efficiency and reduced fuel consumption, but no specific output increase.

Intercooling between compressor stages increases useful output by about 30% for a given air mass flow, by reducing the volume flow and increasing available energy to the power turbine. It has minimal effect on efficiency, since heat removed must be added back in the combustion chamber, but is commonly used in conjunction with recuperation.

In a reheat cycle the fuel is introduced at two locations, increasing the total energy available to produce work. A combination of intercooling, reheat, and recuperation is shown in Figure 8.6.14.

Steam injection, where the steam is injected directly into the combustion chamber, increases the mass flow through the turbine and results in increased output power. Steam-injected gas turbine (SIGT) cycles have been compared from the viewpoints of efficiency, power generation, capital and operating costs, and environmental impacts with CC systems (Esposito, 1989). Above 50-MW size, it was found that CC plants were more economical and achieved significantly better heat rates, although cooling tower fog, visible plumes, and drift deposition favored SIGT plants for a flat site.

Capital and Operating Costs of Power Plants

Table 8.6.5 lists typical costs for constructing and operating fossil fuel plants as of 1993. These costs are variable depending on plant design and location; the table assumes a plant in the Northeast United States burning Pittsburgh bituminous coal. Fuel costs, not listed explicitly, will be affected by the plant efficiency, listed in the table.

Costs will vary according to type of coal burned, the size of the unit, the plant location in the United States, and the extent of environmental control employed. In this table, Pittsburgh bituminous coal is assumed, which has 7% ash, 2% sulfur, 5% moisture, and a higher heating value of 13,395 Btu/lb. Wet scrubbers for PC plants use limestone-forced oxidation. Fixed O&M costs include labor, maintenance, and overhead. Variable O&M costs are largely consumables: water, chemicals, other materials.

Summary

The preceding section has described how the future for electric power generation will increasingly be dominated by environmental control needs, putting an emphasis on the base efficiency of new generation and on heat rate recovery for existing units. The PC-fired power plant with FGD will remain a focus of

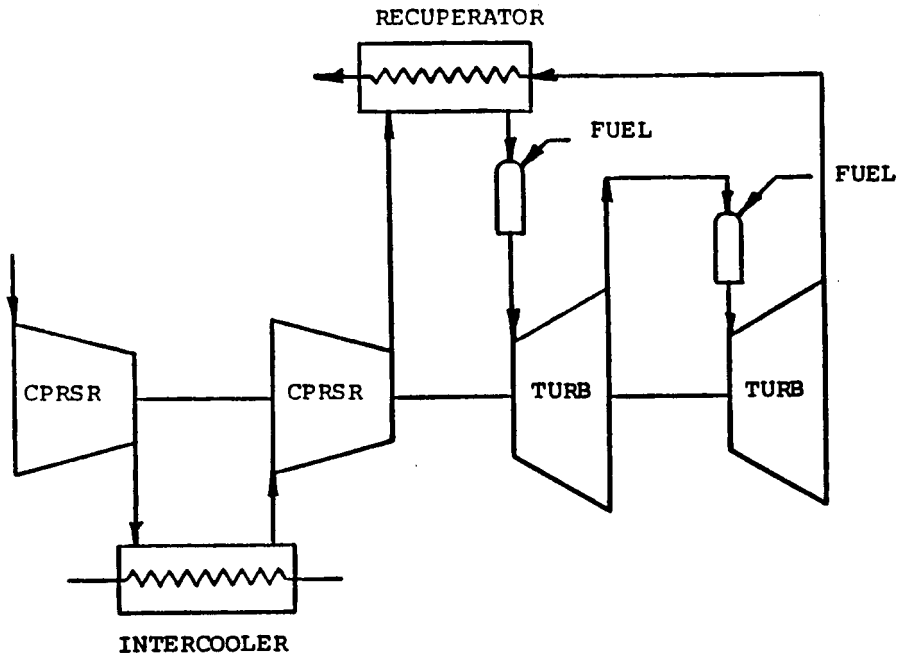


FIGURE 8.6.14 Improvement in CT performance is illustrated in this schematic, which combines an intercooler for the compressor, with a recuperator using CT exhaust heat, and a reheat cycle for the turbine to improve efficiency.

TABLE 8.6.5 Cost Projections for Representative Generation Technologies for a Plant in the Northeast U.S.

	PC Plants			
Plants	**Subcritical Wet Scrubber, 300 MW**	**Supercritical Wet Scrubber, 400 MW**	**AFBC Circulating Bed, 200 MW**	**PFBC CC, 340 MW**
Capital cost (1993),[a] $/kW	1607	1600	1805	1318
Nonfuel O&M costs				
Variable, MILLS/kWhr	3.0	2.8	5.4	1.8
Fixed, $/kW·year	46.6	43.1	37.0	37.6
Efficiency, % (HHV)	36	39	35	41

	Coal Gasification CC, 500 MW	**Coal Gasification Humid Air Turbine, 500 MW**	**Coal Gasification Molten Carbonate Fuel Cell, 400 MW**	**Gas-Fired CC, 225 MW**
Capital cost (1993),[a] $/kW	1648	1447	2082	595
Nonfuel O&M costs				
Variable, MILLS/kWhr	0.5	1.3	1.1	0.4
Fixed, $/kW·year	49.9	40.4	57.2	26.5
Efficiency, % (HHV)	42	42	50	46

[a] Costs of new plants are likely to reduce, in real terms, over the next 10 years due to technology developments and increased worldwide competition for markets in the developing countries. New technologies (PFBC, IGCC, fuel cells) will lower capital costs as production capacity grows.

Source: Technical Assessment Guide, EPRI TR-102275-V1R7, Electrical Power Research Institute, Palo Alto, CA, June, 1993. With permission.

most near-term activity related to upgrades and retrofits. But new technology, based on coal gasification, is under development and being tested in a growing number of demonstration plants which promise extremely low emissions.

The future for many nations will be based on exploiting the opportunities offered by clean and efficient use of coal. This implies access to the range of new technologies now being tested at large scale in the United States and other developed nations. This strategy is both timely and prudent on a global basis, as the world increasingly voices concerns related to carbon combustion.

Base-load, central generation plants will largely be focused in the immediate future on the rapidly developing areas of the world — Asia (particularly China and India) and Latin America. In these areas the fuel of choice will likely be coal, particularly in Asia, and the generating unit most often will be a conventional PC unit, but also could be an atmospheric or pressurized fluidized bed unit. In North America, Europe, and Japan, gas-fired central plants using CTs, often in a CC, will dominate through the next decade. As the cost of natural gas, relative to coal, increases, this will then encourage the installation of gasification units enabling the enormous world coal reserves to be utilized. Then, sometime after the turn of the century, smaller distributed generating sources will begin to emerge, based on gas-fired fuel cells, small CTs, or possibly photovoltaics. As the economics for the distributed option become favorable, these smaller generating units could encourage broad electrification of the developing countries of the world.

Defining Terms

Aeroderivative turbine: In the 1960s gas turbines derived from military jet engines formed a source of utility peaking capacity. Now, modern airline fan-jets are being converted to utility service. These lighter CTs are highly efficient and can have low NO_x emissions, high pressure ratios, and low capital cost.

Ash-softening temperature: The tendency for fly ash to adhere to tube banks is increased as the ash softens and melts. The point at which the ash begins to soften is dependent on the type of coal and is difficult to predict, depending on the many coal constituents. Slagging and fouling of tubes can lead to severe tube corrosion.

Coal gasification: Coal can be converted into a mixture of carbon monoxide and hydrogen by burning it with a controlled deficiency of oxygen. Use of pure oxygen produces a medium-calorific-value gas, and air a low calorific value gas. This "syngas" can then be used to power a CT.

Coextruded tubing: Tubing for superheaters and reheaters must be strong enough to withstand the pressures and temperatures expected, and also corrosion resistant to depositions of fly ash. By making tubing with a strong inner layer and corrosion-resistant outer layer through an extrusion process, both concerns can be dealt with.

Cogeneration: Cogeneration refers to the production of multiple products from a power plant. Typically, process steam, or hot water for heating, is produced in addition to electricity. This approach leads to high plant utilization, the "effective" heat rate being 70% or more.

Combined cycle: Power stations which employ both CTs (Brayton cycle) and condensing steam turbines (Rankine cycle) where the waste heat from the CTs generates steam for the steam turbines, are called *combined* cycle plants. Overall plant efficiency improves.

Electrostatic precipitators: Flue gas particles, when electrically charged in an ionized gas flow, collect on electrodes in the presence of a strong electrostatic field. Collected dust is discharged by rapping into hoppers. A collection efficiency above 99% is possible.

Double reheat: Modern designs of fossil steam-generating units remove a portion of the steam before full expansion through the turbine and reheat it in the boiler before returning it to the turbine. This enhances the thermal efficiency of the cycle by up to 5%. For supercritical cycles two stages of reheat can be justified — double reheat.

F-class machines: Recent designs of CTs have increased efficiencies resulting from increased firing temperatures. The first generation of these machines has firing temperatures of about 2300°F. They have been termed F-class machines (for example the GE 7F). Even higher temperatures have now been incorporated into "G-class" turbines.

Flue gas desulfurization: Removal of sulfur dioxide, SO_2, from combustion gases is accomplished in a number of FGD methods. Most of these involve wet "scrubbing" of the gas using lime or limestone and result in a calcium sulfate waste product. A 95% removal efficiency, or higher, is possible.

Fluidized bed: A process of burning solid fuels, particularly coal, by suspending the fuel within a column of air supplied from below the furnace. This method permits effective combustion of poor-quality fuels, lowers NO_x emissions due to low combustion temperatures, and captures sulfur in the bed by mixing limestone or dolomite in with the fuel.

Fuel cell: Fuel cells convert gaseous or liquid fuels directly to electricity without any combustion process. Like a continuous battery, the fuel cell has electrodes in an electrolyte medium. Typically, hydrogen and air are supplied and DC electricity, water, and carbon dioxide are produced. They are currently high-cost, low-size devices, but with minimum environmental emissions.

Hot-gas cleanup: Cycles which use gas from the combustion of coal, typically pressurized fluidized bed or gasification cycles, need to clean up the ash particles before passing them through a gas turbine. This prevents severe erosion of the turbine blades and other components. Hot-gas cleanup can involve the application of hanging particulate traps, using ceramic filters.

Humid air turbine: A new type of CT uses humidified compressor exit air for the combustor. The mass of dry air needed is thus lessened for a given mass flow, and turbine efficiency increases. Several applications of this HAT appear attractive in gasification and compressed air storage cycles.

Intercooling: Increased output from a CT can be obtained by cooling the air between compressor stages. This reduces volume flow and increases energy to the power turbine.

Lower heating value: Fuels containing hydrogen produce water vapor as a product of combustion. The fuel heating value is said to be "lower" if the combustion process leaves all products in the gaseous state, but "higher" if the fuel heating value includes the latent heat of vaporization. Practice in the United States is to use the higher value.

On/off cycling capability: Generating units are often not required on a 24-hr basis. Some are shut down during low-demand times and started up perhaps hours later. This form of on/off cycling imposes thermal stresses on the equipment, leading to premature equipment failure unless special measures are taken to deal with this.

Petroleum coke: Petroleum coke is a residual product of the oil-refining process, and in its fuel-grade form is an almost pure carbon by-product. About 19 million tons of fuel-grade pet coke is produced each year in the United States. It is inexpensive although it may have high sulfur and vanadium content.

Recuperative cycle: Recuperative cycles for CTs use walls between the hot and cold streams through which heat is transferred. This improves efficiency and reduces fuel consumption.

Regenerative cycles: Combustion turbine cycles using heat exchangers to store and transfer heat from the hot gases to the cold gases are termed regenerative cycles.

Slagging and fouling: The mineral matter in coal can attach itself following combustion to the boiler walls and heat-exchanger surfaces. Oxides of silicon, aluminum, iron, calcium, and magnesium can foul all boiler surfaces, requiring soot blowers for cleaning. Hot ash can melt, becoming sticky and sometimes coalescing in the furnace to cause slagging problems.

Spray drying: Spray driers, for desulfurization, used typically when burning lower-sulfur coals, use a spray of quicklime which is dried by the hot flue gas and results in a dry solid product. A 90% removal efficiency is typical.

Steam injection: Injecting stream directly into the combustion chamber of a CT increases turbine mass flow and thus increases the output power.

Temper embrittlement: Tempering of steel in the manufacturing process removes some of the brittleness and is carried out by a heating and cooling process. During operation, though, it is possible that ductility can worsen close to specific tempering temperatures. The material is then said to be temper embrittled, and premature cracking may follow.

References

Armor, A.F., Generation technologies through the year 2005, in *The Electric Industry in Transition*, Public Utilities Reports, Inc., December 1994.

Armor, A.F. et al., Improved materials for life extension of coal-fired power plants, in *Proceedings, International Conference on Life Extension and Assessment*, Nederlands Instituut Voor Lastechniek, The Hague, Holland, June 13–15, 1988.

Armor, A.F., Touchton, G.L., and Cohn, A., Powering the future: advanced combustion turbines and EPRI's program, paper presented at EPRI Coal Gasification Conference, San Francisco, October 1992.

Blunden, W.E., Colorado-UTE's Nucla Circulating AFBC Demonstration Project, EPRI Report CS-5831, February 1989.

Carpenter, L.K. and Dellefield, R.J., The U.S. Department of Energy PFBC perspective, paper presented at EPRI Fluidized Bed Combustion for Power Generation Conference, Atlanta, Georgia, May 17–19, 1994.

Cohn, A., Humidified power plant options, in *AFPS Developments*, Electric Power Research Institute, Spring 1994.

Couch, G., Advanced coal cleaning technology, IEACR/44, London, IEA Coal Research, December 1991.

Douglas, J., IGCC: Phased construction for flexible growth," *EPRI Journal,* September 1986.

Esposito, N.T., A Comparison of Steam-lnjected Gas Turbine and Combined Cycle Power Plants, EPRI Report GS-6415, June 1989.

Hinrichsen, D., AFBC Conversion at Northern States Power Company, EPRI Report CS-5501, April, 1989.

Lucas, H., Survey of Alternative Gas Turbine and Cycle Designs, EPRI Report AP-4450, February 1986.

Manaker, A.M., TVA 160-MWe Atmospheric Fluidized-Bed Combustion Demonstration Project, EPRI Report TR-100544, December 1992.

Melvin, R.H. and Friedman, M.A., Successful Coal-Fired AFBC Cogeneration in California: 108 MW ACE Cogeneration Facility, paper presented at EPRI Fluidized Bed Combustion Conference, Atlanta, May 17–19, 1994.

Olesen, C., Pressurized fluidized bed combustion for power generation, in EPRI CS-4028, *Proceedings: Pressurized Fluidized-Bed Combustion Power Plants,* May 1985.

Oliker, I. and Armor, A.F., *Supercritical Power Plants in the U.S.S.R.,* EPRI Report TR-100364, February 1992.

Poe, G.G. et al., EPRI's state-of-the-art power plant, in *Proceedings, Third International Conference on Improved Coal-Fired Power Plants,* San Francisco, April 2–4, 1991.

Skowyra, et al., Design of a supercritical sliding pressure circulating fluidized bed boiler with vertical waterwalls, in *Proceedings of 13th International Conference on Fluidized Bed Combustion,* ASME, New York, 1995.

Torrens, I.M., Developing clean coal technologies, *Environment*, 32(6):11–33, July/August 1990.

U.S. Department of Energy, Clean Coal Technology Demonstration Program, DOE/FE-0299P, March 1994.

Further Information

Steam, Its Generation and Use, Babcock and Wilcox, New York.

Combustion: Fossil Power Systems, Combustion Engineering, Inc., Windsor, CT.

Tapping global expertise in coal technology, *EPRI J.*, Jan/Feb., 1986.

IGCC: new fuels, new players, *EPRI J.*, July/Aug., 1994.

A brighter future for PFBC, *EPRI J.*, Dec. 1993.

Fuel cells for urban power, *EPRI J.*, Sept. 1991.

Distributed generation, *EPRI J.*, April/May, 1993.

8.7 Energy Storage

Chand K. Jotshi and D. Yogi Goswami

Introduction

Energy storage is very important for utility load leveling, electrical vehicles, solar energy systems, uninterrupted power supply, and energy systems at remote locations. Two important parameters for energy storage are duration of storage and **specific energy** or **energy density**. Duration of energy storage may vary from many years to a fraction of a second. In a nuclear power plant, nuclear fuel is stored within a reactor for a year. Coal piles, gas and oil storage tanks, or pumped hydro are kept by power utilities for several days use, depending upon the need. Similarly for a solar energy system, requirement of energy storage may be on an hourly, daily, or weekly basis. Specific energy or energy density is a critical factor for the size of a storage system.

Energy can be stored as mechanical, thermal, chemical, electrical, or magnetic energy. In this section, storage of thermal, mechanical, and electrical energy are described.

Thermal Energy Storage (TES)

Thermal energy can be stored as sensible heat, latent heat, or as the heat of chemical reaction (thermochemical).

Sensible heat, Q, is stored in a material of mass m and specific heat C_p by raising the temperature of the storage material and is expressed by Equation (8.7.1):

$$Q = mc_p\Delta T \tag{8.7.1}$$

Most common sensible heat storage materials are water, organic oils, rocks, ceramics, and molten salts. Some of these materials along with their physical properties are listed in Table 8.7.1. Water has the highest specific heat value of 4190 J/kg · C.

TABLE 8.7.1 Physical Properties of Some Sensible Heat Storage Materials

Storage Medium	Temperature Range, °C	Density (ρ), kg/m³	Specific Heat (C), J/kg K	Energy Density(ρC) kWhr/m³ K	Thermal Conductivity (W/m K)
Water	0–100	1000	4190	1.16	0.63 at 38°C
Water (10 bar)	0–180	881	4190	1.03	—
50-ethylene glycol–50 water	0–100	1075	3480	0.98	—
Dowtherm A® (Dow Chemical, Co.)	12–260	867	2200	0.53	0.112 at 260°C
Therminol 66® (Monsanto Co.)	–9–343	750	2100	0.44	0.106 at 343°C
Draw salt ($50NaNO_3$-$50KNO_3$)[a]	220–540	1733	1550	0.75	0.57
Molten Salt ($53KNO_3/40NaNO_2/7NaNO_3$)[a]	142–540	1680	1560	0.72	0.61
Liquid sodium	100–760	750	1260	0.26	67.5
Cast iron	m.p. (1150–1300)	7200	540	1.08	42.0
Taconite	—	3200	800	0.71	—
Aluminum	m.p. 660	2700	920	0.69	200
Fireclay	—	2100–2600	1000	0.65	1.0–1.5

[a] Composition in percent by weight.
Note: m.p. = melting point.

Thermal energy, Q, can be stored as latent heat in a material of mass, *m*, that undergoes phase transformation as given by Equation (8.7.2):

$$Q = m\lambda \tag{8.7.2}$$

where λ = heat of phase transformation.

Four types of phase transformations useful for latent heat storage are: solid $\rightleftarrows$ liquid, liquid $\rightleftarrows$ vapor, solid $\rightleftarrows$ vapor, and solid $\rightleftarrows$ solid. Since phase transformation is an isothermal process, thermal energy is stored and retrieved at a fixed temperature known as the transition temperature. Some common phase change materials (PCMs) used for thermal storage are paraffin waxes, nonparaffins, inorganic salts (both anhydrous and hydrated), and eutectics of organic and/or inorganic compounds. Table 8.7.2 lists some PCMs with their physical properties.

TABLE 8.7.2 Physical Properties of Latent Heat Storage Materials or PCMs

Storage Medium	Melting Point °C	Latent Heat, kJ/kg	Specific Heat (kJ/kg °C)		Density (kg/m³)		Energy Density (kWhr/m³ K)	Thermal Conductivity (W/m K)
			Solid	Liquid	Solid	Liquid		
$LiClO_3 \cdot 3H_2O$	8.1	253	—	—	1720	1530	108	—
$Na_2SO_4 \cdot 10H_2O$	32.4	251	1.76	3.32	1460	1330	92.7	2.25
$Na_2S_2O_3 \cdot 5H_2O$	48	200	1.47	2.39	1730	1665	92.5	0.57
$NaCH_3COO \cdot 3H_2O$	58	180	1.90	2.50	1450	1280	64	0.5
$Ba(OH)_2 \cdot 8H_2O$	78	301	0.67	1.26	2070	1937	162	0.653ℓ
$Mg(NO_3) \cdot 6H_2O$	90	163	1.56	3.68	1636	1550	70	0.611
$LiNO_3$	252	530	2.02	2.041	2310	1776	261	1.35
$LiCO_3/K_2CO_3$, (35:65)[a]	505	345	1.34	1.76	2265	1960	188	—
$LiCO_3/K_2CO_3/Na_2CO_3$ (32:35:33)[a]	397	277	1.68	1 63	2300	2140	165	—
n-Tetradecane	5.5	228	—	—	825	771	48	0.150
n-Octadecane	28	244	2.16	—	814	774	52.5	0.150
HDPE (cross-linked)	126	180	2.88	2.51	960	900	45	0.361
Steric acid	70	203	—	2.35	941	847	48	0.172ℓ

[a] Composition in percent by weight.

Note: ℓ = liquid.

Thermochemical energy can be stored as heat of reaction in reversible chemical reactions. In this mode of storage, the reaction in the forward direction is endothermic (storage of heat), while the reverse reaction is exothermic (release of heat). For example,

$$A + \Delta H \rightleftarrows B + C \tag{8.7.3}$$

The amount of heat *Q* stored in a chemical reaction depends on the heat of reaction and the extent of conversion as given by Equation (8.7.4):

$$Q = a_r m \Delta H \tag{8.7.4}$$

where a_r = fraction reacted, ΔH = heat of reaction per unit mass, and m = mass.

Chemical reaction is generally a highly energetic process. Therefore, a large amount of heat can be stored in a small quantity of a material. Another advantage of thermochemical storage is that the products of reaction can be stored at room temperature and need not be insulated. For sensible and latent heat storage materials, insulation is very important. Examples of reactions include decomposition of metal hydrides, oxides, peroxides, ammoniated salts, carbonates, sulfur trioxide, etc. Some useful chemical reactions are reported in Table 8.7.3.

TABLE 8.7.3 Properties of thermochemical storage media

Reaction	Condition of Reaction: Pressure, kPa	Condition of Reaction: Temperature, °C	Component (Phase)	Pressure, kPa	Temperature, °C	Density, kg/m³	Volumetric Storage Density, kWhr/m³
$MgCO_3$ (s) + 1200 kJ/kg =	100	427–327	$MgCO_3$(s)	100	20	1500	
MgO (s) + CO_2(g)			$CO_2(\ell)$	7400	31	465	
							187
$Ca(OH)_2$(s) + 1415 kJ/kg =	100	572–402	$Ca(OH)_2$(s)	100	20	1115	
CaO(s) + H_2O(g)			$H_2O(\ell)$	100	20	1000	345
SO_3(g) + 1235 kJ/kg =	100	520–960	$SO_3(\ell)$	100	45	1900	
SO_2(g) + $^1/_2O_2$(g)			$SO_2(\ell)$	630	40	1320	
			O_2(g)	10000	20	130	280

Note: s = solid, ℓ = liquid, g =gas.

Applications and Examples

Cool Storage has major applications in space cooling of buildings, food and medicine preservation, and transportation of items that need to be stored at low temperatures. A major application of cool storage is in the use of off-peak electricity for air-conditioning during peak hours. During off-peak hours electricity can be used to make ice or chilled water, which can be used later for air-conditioning of buildings during the peak hours. The advantage of using ice as a storage medium over chilled water is that a much larger amount of coolness can be stored in ice; 1 kg of ice stores 335 kJ, whereas 1 kg of water stores only 42 kJ for a temperature swing of 10°C. The disadvantage of ice is its lower thermal conductivity, which is responsible for lower heat-transfer rates.

Cool storage systems have been used in several buildings in the United States and Canada. The Merchandise Mart of Chicago boasts the largest ice storage system in the world: each day more than 2 million lb of ice are made and melted. For long-term cool storage, aquifers have been used for chilled water storage. Examples include cooling of buildings at the University of Alabama and the United States Postal Service in Long Island, NY, using chilled water stored in aquifers (Tomlinson and Kannberg, 1990). Other materials which have been found to have cool storage potential are PCMs like $LiClO_3 \cdot 3H_2O$, a eutectic of Glauber's salt, paraffins and their mixtures, and some gas hydrates or clatherates.

Heat Storage has major applications in space heating, crop drying, cooking, electric power generation, industrial process heat (air and steam), waste heat utilization, and solar energy utilization, etc. Heat storage in water is the most economical and well-developed technology. Epoxy-lined steel, fiberglass-reinforced polymer, concrete with plastic liner, and wood tanks are suitable containment materials for systems using water as the storage material. The storage tanks may be located above or below ground. In North America and China, aquifers have been used for long-term storage of hot water and chilled water. Pressurized water tanks are used to store heat from off-peak electricity (ASHRAE, 1995). For example, water is heated to maximum temperatures of about 138°C in a tank at a pressure of 50 psig.

Molten nitrate salt (50 wt% $NaNO_3$/50 wt% KNO_3) also known as Draw salt, which has a melting point of 222°C, has been used as a storage and a heat-transfer fluid in an experiment in Albuquerque, NM. It was the first commercial demonstration of generating power from storage (Delameter and Bergen, 1986). Solar Two, a 10-MW solar thermal power demonstration project in Barstow, CA, is also designed to use this molten salt to store solar energy (Chavez et al., 1995). Another molten nitrate salt is 40 wt% $NaNO_2$/7 wt% $NaNO_3$/53 wt% KNO_3, known as HTS (heat-transfer salt) with a melting point of 142°C. This salt has been widely used in the chemical industry.

For applications in heating and cooling of buildings the containment of PCM can become an integral part of the building. It may be part of the ceiling, wall, or floor of the building and may serve a structural or a nonstructural function. Tubes, trays, rods, panels, balls, canisters, and tiles containing PCMs have been studied in the 1970s and 1980s for space-heating applications (Moses and Lane, 1983). The PCMs used were mostly salt hydrates such as Glauber's salt ($Na_2SO_4 \cdot 10H_2O$), Hypo ($Na_2S_2O_3 \cdot 5H_2O$),

$NaCH_3COO \cdot 3H_2O$, $Na_2HPO_4 \cdot 12H_2O$, $Ba(OH)_2 \cdot 8H_2O$, $MgCl_2 \cdot 6H_2O$, and $Mg(NO_3)_2 \cdot 6H_2O$. Paraffin mixtures have been used for thermal storage in wall boards. Some PCMs, such as salt hydrates, exhibit supercooling and phase segregation problems during heat removal. Low thermal conductivity and complex mechanism of heat transfer during melting and freezing introduce complexities in the design of their containment systems.

Mechanical Energy Storage

Mechanical energy may be stored as potential or kinetic energy.

Kinetic Energy

Kinetic energy can be stored in the rotating mass of a wheel, commonly known as a flywheel. Kinetic energy of a rotating body is given by Equation (8.7.5):

$$KE = \frac{1}{2} I\omega^2 \tag{8.7.5}$$

where I = moment of inertia, and ω = angular velocity.

The maximum specific energy of a flywheel is expressed by the following equation (Jenser, 1980):

$$\frac{KE_{max}}{m} = A\frac{\rho_{max}}{\rho} \tag{8.7.6}$$

where A = shape factor, and its value depends on the geometry of flywheel; $A = 1.0$ for a constant stress disk and 0.5 for a thin-rimmed flywheel. This equation shows that high tensile strength and low density are the key parameters to store maximum energy. Tensile strength, density, and specific energy of some materials are given in Table 8.7.4.

TABLE 8.7.4 Flywheel Rotor Materials

Material	Design Stress, MN/m²	Density, kg/m³	Specific Energy, Whr/kg
Composite fiber[a]/epoxy	750	1550	51.5
E-glass fiber[a]/epoxy	250	1990	14.0
S-glass fiber[a]/epoxy	350	1900	19.6
Kevlar fiber[a]/epoxy	1000	1400	76.2
Maraging steel	900	8000	24.2
Titanium alloy	650	4500	30.8

[a] 60% fiber.

Storing energy in a flywheel is one of the oldest techniques used in ancient potteries. Present-day flywheels are much more advanced as a result of superstrong/ultralight composite materials and frictionless high-performance magnetic bearings.

Potential Energy

If a body of mass m is elevated against the gravitational force g to a height Δh, the potential energy stored is given by

$$PE = mg\Delta h \tag{8.7.7}$$

From Equation (8.7.7), 1 Wh of energy can be stored in 1 kg mass of a body by raising it to a height of 367 m.

Potential energy is also stored in a spring, either by compressing or expanding. Here, energy stored is given by

$$PE = (1/2)kx^2 \tag{8.7.8}$$

where k = spring constant and x is the distance to which the spring is compressed or expanded. Springs have been widely used to power toys and watches mainly because of very low values of energy density.

Pumped Hydro

Water may be pumped from a lower reservoir to a higher reservoir using electricity during off-peak hours, which may be used to generate electricity using hydraulic turbines during peak hours. Figure 8.7.1a shows a schematic diagram of an above-ground pumped hydro system. Advantages of pumped hydro units include simple operation, high reliability, low maintenance, long life, quick start from standstill, and economic generation of peaking electrical energy. In the United States a large number of such systems are in operation. Power-generating capacities of these systems vary between 5 and 2000 MW (Makansi, 1994). The overall efficiencies of these power plants vary between 65 and 90%, which includes the efficiencies of pumps, hydraulic turbines, generators, and losses from the upper reservoir. In spite of the technical and economic viability of pumped hydro, the requirement of a specific type of topography and some environmental concerns limit its application. To overcome these problems, a concept of underground pumped hydro storage as shown in Figure 8.7.1b can be used. In this case, large caverns or aquifers can be used as the lower reservoir.

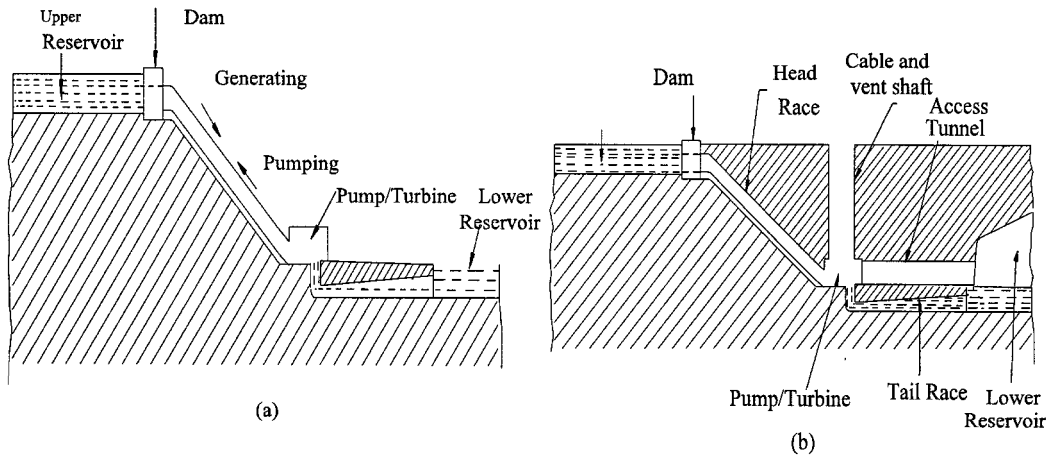


FIGURE 8.7.1 Pumped hydrostorage: (a) above ground; (b) underground.

Compressed Air Storage

In a fossil fuel power plant, approximately half of the output of a conventional gas turbine is used to drive the compressors. If the air is compressed during off-peak hours and stored, it can be used later during peak hours. Compressed air can be stored underground in abandoned mines, oil or gas fields, sealed aquifers, or natural caverns. A compressed air storage plant in Huntorf, Germany that has been in operation since 1978 is the oldest unit. A 110-MW demonstration facility at the Alabama Electric Cooperative (AEC), Inc., McIntosh, Alabama , site has been in operation since 1991. The Hunterof unit is a 4-hr facility and the AEC site is a 26-hr facility (Makansi, 1994).

Electrical Energy Storage

Electrochemical energy storage, more commonly known as battery storage, stores electrical energy as chemical energy. Batteries are classified as primary and secondary batteries. Only secondary batteries are rechargeable and are therefore suitable for energy storage applications. Lead-acid and nickel-cadmium are well-known rechargeable batteries that are most commonly used. Lead-acid batteries have been used for over a century and are still the most popular batteries.

Electrochemical operation of a cell during discharge and charge is shown in Figures 8.7.2a and b. During discharge when a cell is connected to a load, electrons flow from the anode to the cathode. In this operation oxidation, or loss of electrons, takes place at the anode, and reduction, or gain of electrons, occurs at the cathode. The cell chemistry of a lead-acid battery is as follows: the anode is lead (Pb) and the cathode is lead oxide (PbO_2); the electrolyte is H_2SO_4. The cell reaction is

$$2H_2SO_4 \rightarrow 4H^+ + 2SO_4^-$$

$$\text{Anodic reaction:} \quad Pb + SO_4^{2-} = PbSO_4 + 2e^- \tag{8.7.9}$$

$$\text{Cathodic reaction:} \quad PbO_2 + SO_4^{2-} + 4H^+ + 2e^- = PbSO_4 + 2H_2O \tag{8.7.10}$$

Theoretical voltages and capacities of some known batteries are reported in Table 8.7.5.

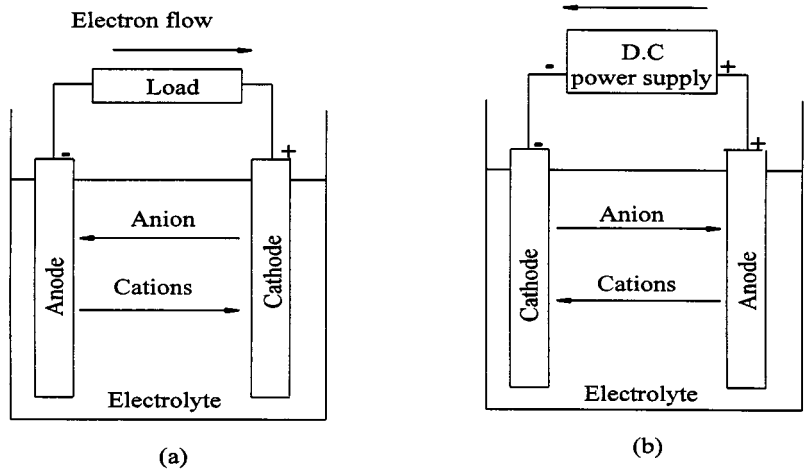


FIGURE 8.7.2 Electrochemical operation of a cell: (a) discharge; (b) charge.

TABLE 8.7.5 Properties of Some Rechargeable Batteries

No.	System	Electrolyte	Temp. °C	O.C.V. (V)	Energy Density (Theoretical), Whr/kg	Energy Density (Achievable), Whr/kg	In/Out Electrical Effciency	Cycle Life
1	Lead-Acid	H_2SO_4	20–30	2.05	171	50	75	1000
2	Nickel-Iron	KOH	20–30	1.37	267	60	55	2000
3	Zinc-Iron	KOH	50–60	1.65	1084	90	45	600
4	Sodium-Sulfur	β-Al_2O_3	300–375	1.76–2.08	664	120	75	2000
5	Lithium-Iron Sulfide	LiCl-KCl (eutectic)	400–450	1.6	869	150	75	1000

Note: O.C.V. = open cell voltage.

Applications

Battery storage is used in a wide range of applications. Currently, the main emphasis of research is on applications in vehicles and load leveling.

Electric Vehicles. For electric vehicles, the specific energy and specific power are two important parameters. The greater the specific energy, the farther a vehicle can travel. If specific power is high, a vehicle can accelerate more quickly and have a higher top speed. Other important requirements are the ability to charge and discharge a large number of times, to retain charge over an extended period of time, and to charge and discharge over a wide range of temperatures. Table 8.7.5 provides information about some rechargeable batteries.

Power Plants. Recent start-up of a commercial unit for the Puerto Rico Electric Power Authority (PREPA) is the latest development in large-scale application of lead-acid batteries. The facility stores 20 MW for 20 min, both for peaking requirement and voltage and frequency control. Maximum discharge is limited to 10 MW (Makanasi, 1994).

Defining Terms

Cool storage: The storage of thermal energy at temperatures below the nominal temperature required by the space or process.

Heat storage: The storage of thermal energy at temperatures above the nominal temperature required by the space or process.

Energy density: Amount of energy stored per unit volume, kJ/m^3 or kWhr/m^3.

Specific energy: Amount of energy stored per unit mass, kJ/kg or kWhr/kg.

References

ASHRAE, 1995. Thermal storage, in *ASHRAE Handbook, HVAC Application*, p. 40.15. American Society of Heating, Refrigerating and Air-Conditioning Engineers, 1791 Tullie Circle, N.E., Atlanta.

Beckman, G. and Gilli, P.V. 1984. *Topics in Energy — Thermal Energy Storage*. Springer-Verlag, New York.

Chavez, J.M. et al. 1995. The Solar Two Power Tower Project: a 10MWe power plant, in *Proceedings of the 1995 IECEC*, Vol. 2, pp. 469–475, ASME, New York.

Delameter, W.R. and Bergen, N.E. 1986. Review of Molten Salt Electric Experiment: Solar Central Receiver Project. SAND 86-8249, Sandia National Laboratory, Albuquerque.

Garg, H.P., Mullick, S.C., and Bhargava, A.K. 1985. *Solar Thermal Energy Storage*. D. Reidel, Boston.

Glendenning, I. 1981. Advanced mechanical energy storage, in *Energy Storage and Transportation,* G. Beghe, Ed., pp. 50–52. D. Reidel, Boston.

Jensen, J. 1980. *Energy Storage*. Newnes-Butterworth, Boston.

Makansi, J. 1994. Energy storage reinforces competitive business practices. *Power.* 138(9):63.

Moses, P.J. and Lane, G.A. 1983. Encapsulation of PCMs, in *Solar Heat Storage: Latent Heat Materials*, Vol. II, pp. 93–152. CRC Press, Boca Raton, FL.

O'Connor, L. 1993. Energizing the batteries for electric cars. *Mech. Eng*. 7:73–75.

Sharma, S.K. and Jotshi, C.K. 1979. Discussion on storage subsystems, in *Proceedings of the First National Workshop on Solar Energy Storage*, pp. 301–308. Panjab University, Chandigarh, India.

Tomlinson, J.J. and Kannberg, L.D. 1990. Thermal energy storage. *Mech. Eng*. 9:68–72.

8.8 Nuclear Power

Roberto Pagano and James S. Tulenko

Nuclear power refers to power generated through reactions involving atomic nuclei (i.e., nuclear reactions). These reactions fall into three broad categories — fusion reactions, fission reactions, and radioisotopes. In fusion, two light nuclei (most commonly isotopes of hydrogen) combine to form a heavier nucleus (usually helium), with energy being released in the process. Nuclear fusion is the source of energy generated in the stars (our sun). In artificial applications, the technology to induce fusion reactions has been available for several decades, but such reactions have been essentially uncontrolled (the hydrogen bomb). Once initiated, fusion reactions generate huge amounts of energy, which is subsequently released explosively. Means to produce and release energy from fusion in a sustained, controlled manner are still being developed. Extensive research is ongoing, both in the United States and abroad, on the development of nuclear fusion as a controlled source of power.

Nuclear fission, in contrast, is the basis of a mature technology applied to the generation of power. Fission is the fragmentation of a heavy nucleus into two, sometimes three, lighter nuclei. Certain nuclides found in nature fission spontaneously, that is, with no external intervention. However, spontaneous fission in naturally occurring nuclides takes place at a very slow rate. Fission can be induced through a nuclear reaction. Of primary interest here is the fissioning of several specific nuclei through interactions with neutrons. Again, the fissioning of a nucleus is accompanied by the release of energy.

At present, the element of primary importance with respect to nuclear fission power is uranium. Naturally occurring uranium consists of three isotopes — ^{238}U, ^{235}U, and ^{234}U. In a mixture of isotopes of an element, the abundance of any one is usually expressed as the number of atoms of that isotope present per 100 atoms of the mixture, abbreviated as atom percent a/o or weight percent w/o. Natural uranium consists of 99.2745 a/o ^{238}U, 0.7200 a/o ^{235}U, and 0.0055 a/o ^{234}U.

Radioisotope power is the third form of nuclear energy. When radioisotopes decay, high-energy electrons (beta particles), helium atoms (alpha particles), and gamma rays (photons) are emitted. When the energy of these radiations is stopped and converted to heat, a power source is created. Radioisotopes decay energies range from 0.01 to 10 MeV. Radioisotopes are generally separated from radioactive wastes produced from nuclear power plants. The most common radioisotopes are polonium-210 (alpha emitter of 5.3 MeV), plutonium-238 (alpha emitter of 5.46 MeV), cesium-144 (beta emitter of 1.25 MeV), and strontium-90 (beta emitter of 1.10 MeV).

The Fission Process

Consider first the fission of a nucleus of ^{235}U caused by an interaction with a neutron. A compound nucleus of ^{236}U is initially formed. If fission occurs, it does so within a very short time. Normally with thermal reaction, 85% of the interaction leads to fission. Alternatively, the compound nucleus dissipates energy by emitting a gamma photon, no fission occurs and the nucleus remains as ^{236}U. This latter process occurs in 20% of the interaction of ^{235}U with a neutron.

The fissioning of a nucleus produces fission fragments called fission products, a number of neutrons and gamma photons. Most frequently, the number of neutrons is 2 or 3, ranging in extreme cases from 0 to 8. In the fissioning of ^{235}U, the average number of neutrons released, designated by υ, has a value of 2.42. This value applies strictly if the fission is induced by a neutron of relatively low kinetic energy, called a thermal neutron.

In a nuclear reactor, which is a special material medium in which ^{235}U is dispersed for the reactor to work, one of the neutrons liberated in the fissioning of a nucleus must go on to induce a fission in another nucleus. This leads to the idea of a self-sustaining chain reaction or, more specifically, a **critical configuration** in which a self-sustaining chain reaction can be maintained indefinitely. To this end, an adequate supply of ^{235}U must be on hand and replenished as needed. Further, the configuration must be

such that the likelihood that any particular neutron ultimately induces a fission is adequately high to ensure that on average one neutron will induce a fission.

The energy liberated in fission results from Einstein's equation $E = Mc^2$, which says that mass and energy are equivalent. A summation of the masses of the fission fragments and the neutrons resulting from fission shows that the combined mass of the products of a fission is less than the mass of the compound nucleus before fission occurs. It is found that the energy released is equal to $E = \Delta Mc^2$ when ΔM is the difference in the masses. This nuclear energy is traditionally expressed in units of electron volts (eV), with the equivalence of 1 eV = 1.602×10^{-19} J. The single fission of a single uranium atom releases approximately 200×10^6 eV, or 200 MeV, of energy. When one realizes that the combusion of a single carbon atom ($C + O_2 \rightarrow CO_2$) releases 4 eV, a quantity 50 million times smaller, one gets a true appreciation of the concentrated power of nuclear energy.

Most of the energy liberated in a fission appears as the kinetic energy of the fission fragments (168 MeV, or 84%) and the kinetic energy of the neutrons (5 MeV, or 2.5%). The remainder is distributed among the gamma photons appearing instantaneously with the fission and the energy associated with the radioactive decay of the fission fragments. These fragments are readily stopped in the reactor. Their range is of the order of 0.01 to 0.001 mm. Thus, the major portion of the energy from a fission is deposited within a very short distance from the site of the fission.

In the largest power reactors in operation in the United States today, heat is generated at the rate of 3800 MW. At 200 MeV per fission, fissioning in such reactors must occur at the rate of 1.2×10^{20} fissions/sec to produce the power. In terms of ^{235}U, this requires the fissioning of all the nuclei contained in 0.047 g of ^{235}U, or the fissioning of approximately 4 kg per day. Thus, as a rule of thumb, the fissioning of all the nuclei present in 1 g of ^{235}U is sufficient to generate 1 MW day of thermal energy. In contrast, the generation of the same amount of energy from coal requires the combustion of 4 tons of coal of typical heating value. With the combustion of coal there is the associated release of a large quantity of carbon dioxide (~14 tons) to the atmosphere with its effects on global warming.

Cross Sections

A measure of the probability of a particular nuclide to interact with a neutron is provided by a quantity known as a **cross section**. Numerical values of cross sections are determined experimentally and are expressed in units of barns (b), with 1 b defined as 10^{-24} cm^2, or 10^{-28} m^2.

As a quantity, a cross section may be interpreted as a target area — the larger the cross section, the more likely the interaction of the nucleus with a neutron in its vicinity. For example, the cross section for fission of ^{235}U, denoted by σ_f^{235}, has a value of 582 b if the interacting neutron is traveling at the velocity associated with thermal energy (2200 m/sec). If the neutron is traveling at high energy, the cross section may drop to a value of approximately 2 b. With respect to the radiative capture of a neutron in ^{235}U leading to the formation of ^{236}U, the cross section is given by $\sigma_c^{235} = 99$ b, if the neutron is thermal. In summary, a cross section is a property specific to a given nuclide, but it is a property whose value depends on the energy of the interacting neutron.

Cross sections are additive. Thus, the cross section for the absorption of a thermal neutron in ^{235}U — whether the absorption gives rise to a fission or a radiative capture — is given by $\alpha_a = \alpha_f + \alpha_c = 582$ b + 99 b = 681 b. Further, the probability of a fission occurring as a result of a thermal neutron being absorbed in ^{235}U is given by $\alpha_f/\alpha_a = 582$ b/681 b = 0.85. Fission is, therefore, the more likely outcome of an interaction between a ^{235}U nuclide and a thermal neutron.

On average, the neutrons arising from the fissioning of ^{235}U have a kinetic energy of 2 MeV, corresponding to a speed of 2×10^7 m/sec. These neutrons are four orders of magnitude greater than the speed at which a neutron is considered to be thermal (2200 m/sec).

Categories of Nuclear Reactors

A prerequisite for a self-sustaining chain reaction is that sufficient ^{235}U be present in the medium to ensure that the absorption of a neutron in a nucleus of ^{235}U is a likely occurrence. If the population of neutrons present in the medium at any instant consists predominantly of slow neutrons, a far lesser amount of ^{235}U is needed to ensure criticality than would be the case if the population were to consist of fast neutrons. This comes about because of the difference in the values of the cross sections mentioned previously.

There is, from this particular standpoint, an incentive to slow down the neutrons originating in fission in order to reduce the inventory of ^{235}U needed to maintain criticality. In the power reactors operating today, means are provided to slow down the neutrons. The slowing down is effected through multiple elastic scatterings of the neutrons with the nuclei of light elements deliberately present in the medium acting as so-called **moderators**. Notable among such elements are hydrogen present in water, deuterium in heavy water, and carbon in the form of graphite.

All of the reactors in which substantive moderation of the neutrons occurs are categorized as **thermal reactors**. This term stems from the distinguishing feature that the neutron population is in, or near, thermal equilibrium with the nuclei of the moderator. As a consequence, there is no net exchange of energy between the neutron population and its surroundings. The neutrons are then referred to as thermal neturons.

In a population of **thermal neutrons**, the distribution of the speeds of the neutrons is characterized, adequately in many cases, by the Maxwell-Boltzmann distribution, originally formulated to apply to the molecules of an ideal gas. According to this distribution, the most probable speed of the neutrons at the reference temperature of 20°C is 2200 m/sec and the kinetic energy corresponding to the most probable speed is 0.025 eV.

Reactors that are not thermal reactors fall in the category of **fast reactors**. In these reactors, the moderation of neutrons is much reduced for reasons discussed later.

Nuclear Fuel

In light-water reactors (LWR), the type of power reactors most commonly in service today, the nuclear fuel is uranium with a content of 2 to 4% of ^{235}U. This fuel is produced by enriching natural uranium in ^{235}U by one of several technologies, principally gaseous diffusion and gaseous centrifugation (Benedict et al., 1981). Neutrons of all energies, down to and including thermal energies, can induce fission in ^{235}U. For this reason, ^{235}U is said to be fissile.

In contrast, ^{238}U, present to the extent of 96 to 98% in the fuel, can be fissioned to a significant extent by neutrons with energies in excess of a threshold of roughly 2 MeV. Fissions of ^{238}U, referred to as fast fissions, play only a slight role in the chain reaction in an LWR. However, ^{238}U absorbs neutrons radiatively to yield ^{239}U. This nuclide is radioactive and decays to ^{239}Np which, in turn, decays to ^{239}Pu, a **fissile nuclide**. Thus, in LWR fuel ^{239}Pu, produced from ^{238}U, is available for fissioning by neutrons of all energies and contributes to the chain reaction. Because of its ability to form fissile ^{239}Pu, ^{238}U is termed a **fertile nuclide**.

Conversion and Breeding

To characterize the unique capability of nuclear fuel simultaneously to produce and consume fissile material, a figure of merit known as the **conversion ratio** (CR) is informative. It is defined by the relation:

$$\text{CR} = \frac{\text{number of fissile nuclei produced from fertile nuclei}}{\text{number of fissile nuclei consumed}}$$

In fuel irradiated in an LWR, the conversion ratio typically has a value of 0.5. If the appropriate combination of materials, design, and operating parameters could be found to raise the conversion ratio

to a value greater than unity, a reactor would become a breeder reactor, that is, one that produces more fissile material than it consumes in its operation.

To illustrate the possibility of breeding, consider a parameter known as the reproduction factor and given by

$$\eta = \upsilon \frac{\text{number of neutrons causing fission in fuel}}{\text{total number of neutrons absorbed in fuel}} = \nu \frac{\sigma_f^{\text{fuel}}}{\sigma_a^{\text{fuel}}}$$

In the case of ^{235}U and thermal neutrons, the value of η is given by $\eta_{235} = 2.42 \times 582/(589 + 99) = 2.42 \times 0.85 = 2.07$.

If a chain reaction is to be self-sustaining in a reactor, the condition $\eta > 1$ must apply. To achieve breeding requires that $\eta > 2$. In other words, one neutron from fission would be available to sustain the chain reaction and another to produce a fissile nucleus from a fertile nucleus. Practical considerations indicate that the value of η must be substantially greater than 2, since neutrons will be lost by absorption in structural materials, heat removal medium, fission fragments, the moderator, if present, and by escaping from the physical confines of the reactor.

As shown, ^{235}U has a value of η slightly above 2 with neutrons at thermal energies. Breeding or near-breeding conditions could arise, in principle, if very judicious choices of materials and parameters prevail. A more attractive fuel from the standpoint of breeding is ^{233}U. This isotope of uranium is produced artificially by placing the naturally occurring nuclide ^{232}Th in a reactor. A radiative capture of a neutron in ^{232}Th leads to the formation of the radioactive nuclide ^{233}Th. Two successive radioactive decays yield ^{233}U. In a thermal reactor, the value of η with ^{233}U is 2.29, approaching the level where breeding might be contemplated. Intermediate between the two fissile isotopes of uranium is ^{239}Pu, which at thermal energies yields a value of η of 2.15.

Figure 8.8.1 shows the behavior of η as the energy of the neutrons inducing fission increases. A little above thermal energies the value of η for ^{239}Pu and ^{235}U drops below 2, indicating that breeding is impossible at such energies. As energy increases, the values of η for both reach the threshold of 2 and continue to increase steadily, with ^{239}Pu clearly the more attractive fuel from the standpoint of breeding. The value of ^{239}Pu for ^{235}U is relatively insensitive to increases in energy and remains continuously above 2. Again, at higher energies, η is the more attractive fuel.

Research on the development of **breeder reactors** has focused on the ^{239}Pu fuel cycle, both in the U.S. and abroad. Representative of these reactors is the liquid metal fast breeder reactor (LMFBR) in which liquid sodium is the heat-removal fluid and provides the small amount of moderation needed. Breeder reactors with ^{233}U as fuel, represented by an adaptation of the high-temperature gas-cooled reactor (HTGCR), in which the heat-removal fluid is helium and the moderator is graphite, although less attractive, in principle, as breeders, might with further research and development prove to be viable alternatives to the LMFBR.

LWR Fuel

Nuclear fuel in light water is in the form of small cylindrical pellets of the ceramic UO_2, with the uranium enriched to 2 to 4% in ^{235}U, as mentioned previously. These pellets are stacked vertically in tubes and the ends of the tubes are sealed off. The dimensions and further details given here apply strictly to the more common type of LWR, known as the pressurized water reactor (PWR), but may be taken as generally representative of LWRs.

The tubes containing the fuel pellets are referred to as fuel rods. They are 4.3 to 4.7 m in length and 0.0095m in outside diameter. An array typically of 17 × 17 rods constitutes a fuel assembly, as shown in Figure 8.8.2. Certain fuel rods within the assembly are replaced by guide sheaths in which absorber rods can be moved vertically. These rods absorb neutrons, thus providing one of the means of controlling the chain reaction as the rods are inserted or withdrawn. Within the fuel assembly the fuel rods are placed on a pitch of 0.0127 m, leaving vertical passages through which water can flow. A total of

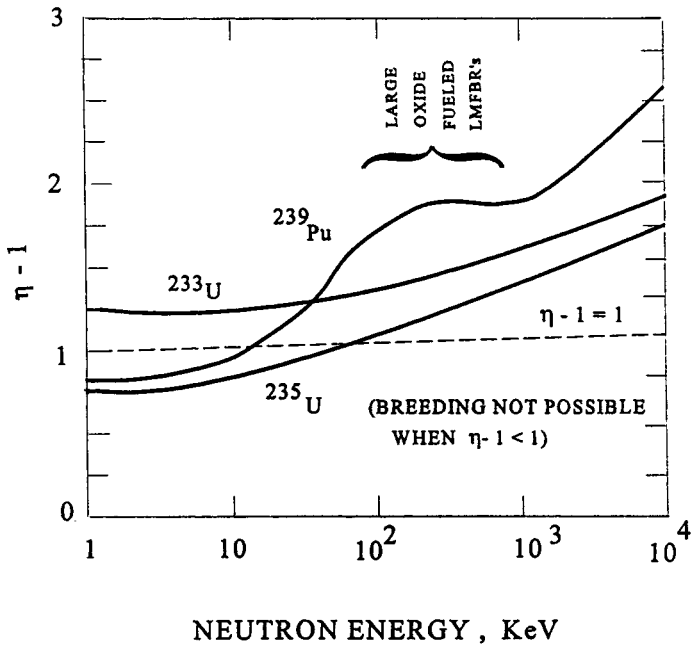


FIGURE 8.8.1 Dependence of reproduction factor of neutron energy.

approximately 200 fuel assemblies, juxtaposed to form a roughly cylindrical configuration constitute the core of the reactor. Water is circulated through the core where the water serves both to moderate the neutrons and to remove the heat generated by fissions in the nuclear fuel.

Light-Water Reactors

LWRs currently make up the largest portion of the installed nuclear generating capacity throughout the world. Among these, the PWRs are more numerous. By operating at a sufficiently high pressure, bulk boiling of the reactor coolant is suppressed in a PWR. In contrast, the coolant is allowed to boil in a boiling water reactor (BWR) and a portion of the coolant is converted to steam as it circulates through the core.

Pressurized Water Reactors

A schematic of a PWR system is shown in Figure 8.8.3. At the heart of the system is the core made up of fuel assemblies and associated control rods. The core is contained in a reactor vessel, or pressure vessel, designed to operate at a pressure typically of 15.5 mPa. Water is circulated through the core where it acts as a moderator and also removes the heat generated through fission. Typical operating temperatures at full power are 295°C at the inlet and 330°C at the outlet, an increase of 35°C as a result of the water passing through the core.

From the reactor vessel, the coolant is circulated to steam generators and returned to the reactor vessel to complete the so-called primary loop. This loop constitutes the nuclear steam supply (NSS). Steam emerging from the steam generators is directed toward the secondary loop, or balance of plant, consisting of turbine generator, condenser, and feedwater pumps. Extremely small quantities of radioactive contaminants may be present in the steam generated in PWRs, and all releases to the environment from the secondary side of the plant are carefully monitored and controlled. Otherwise, steam from the NSS differs from steam generated in fossil fueled plants only inasmuch as its grade is inferior.

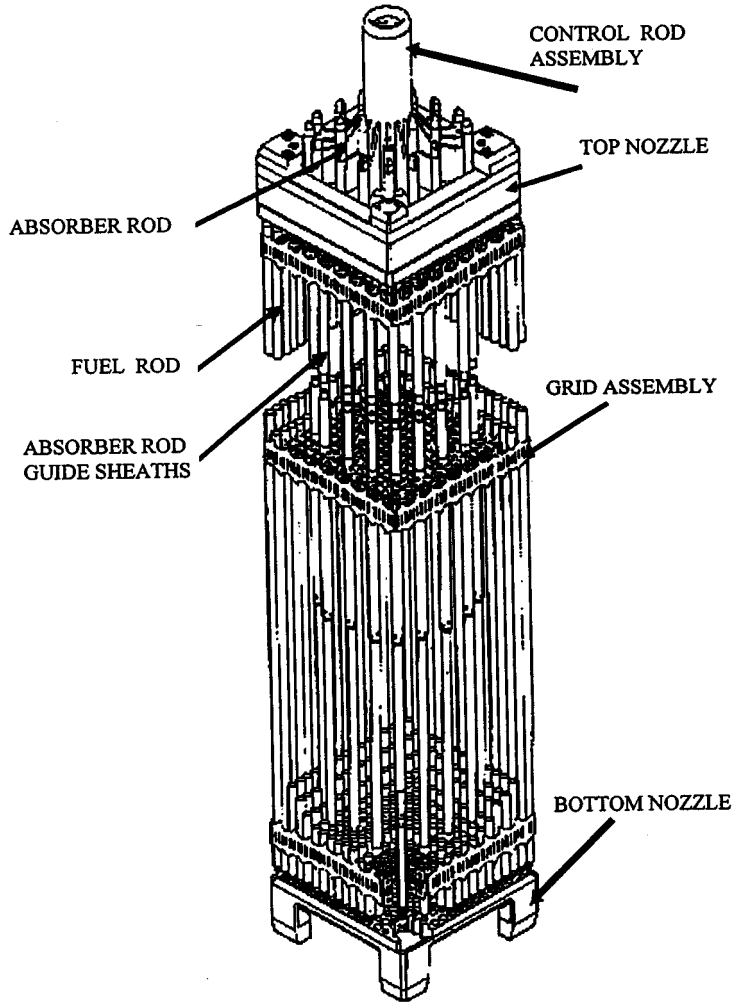


FIGURE 8.8.2 PWR fuel assembly.

Boiling Water Reactors

Figure 8.8.4 is a schematic of a BWR system. A reduced operating pressure of 7.2 MPa causes a portion of the coolant to flash to steam. Steam separators and driers allow dry steam to emerge from the reactor vessel, thus eliminating the need for separate steam generators and a secondary loop. Steam from the reactor vessel flows through a turbine generator and condenser, from which circulating pumps return the condensate to the reactor. Large quantities of radioactive contamination may be present in the steam produced by a BWR because of the direct cycle, so releases must be carefully monitored. A distinguishing feature of the BWR are the jet pumps, typically 24 in number, placed along the periphery of the core. These pumps augment the flow of coolant through the core.

Defining Terms

Breeder reactors: Reactors in which the conversion ratio is greater than unity.

Conversion ratio: The ratio of the number of fissile nuclei produced in a reactor to the number of fissile of nuclei consumed.

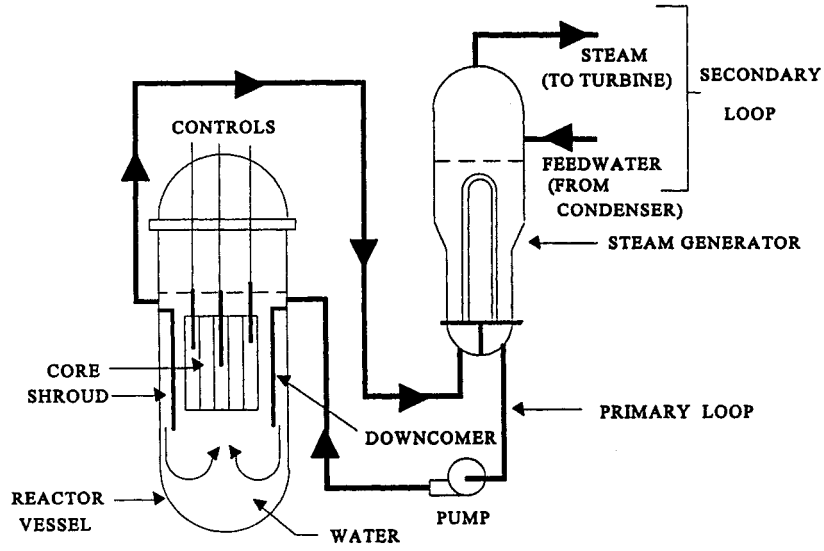


FIGURE 8.8.3 Schematic of a pressurized water reactor.

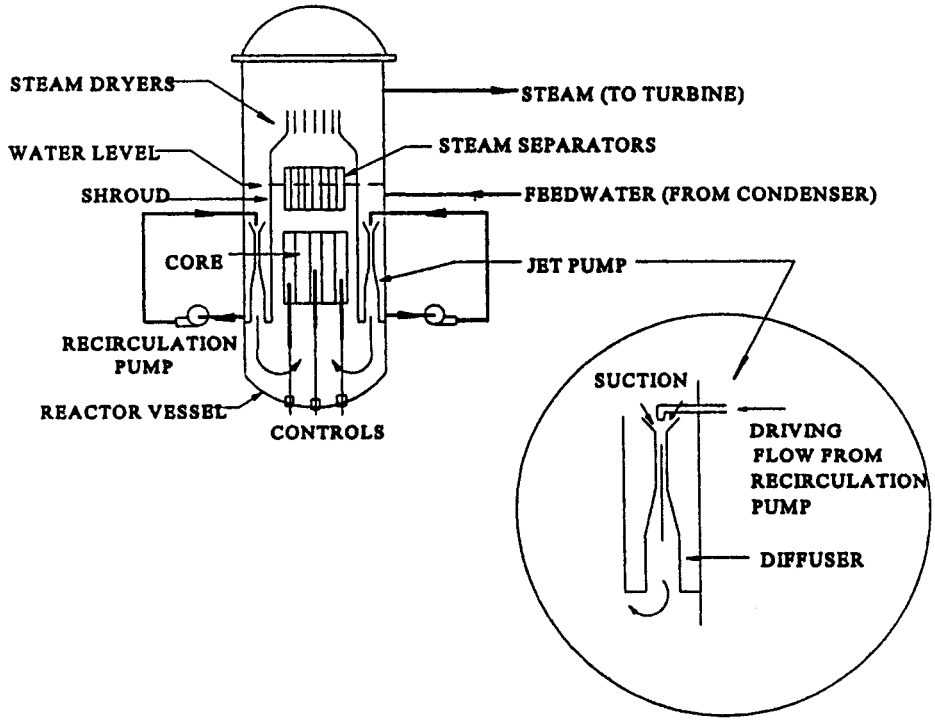


FIGURE 8.8.4 Schematic of a boiling water reactor.

Critical configuration: A medium containing nuclear fuel in which a self-sustaining chain reaction can be maintained.

Cross section: A numerical quantity, determined experimentally, related to the probability that a specific nuclide will undergo a given nuclear reaction.

Fast reactors: Reactors in which little moderation of the neutrons occurs and the neutron populations consist of neutrons of relatively high speeds.

Fertile nuclide: A nuclide that, through the absorption of a neutron and subsequent radioactive decays, can produce a fissile nuclide.

Fissile nuclide: A nuclide that can be fissioned by neutrons of all energies, down to and including thermal energies.

Moderator: A component of a reactor present expressly to slow down neutrons and produce a population of thermal neutrons.

Thermal neutrons: A population of neutrons in, or near, thermal equilibrium with the nuclei of a medium in which the populations exists.

Thermal reactors: Reactors in which the neutron populations consist predominantly of thermal neutrons.

References

Benedict, M., Pigford, T.H., and Levi, H.W. *Nuclear Engineering,* 2nd ed., McGraw-Hill, New York, 1981.

Further Information

Glasstone, S. and Sesonske, A. *Nuclear Reactor Engineering,* 3rd ed., Van Nostrand Reinhold, New York, 1981.

8.9 Nuclear Fusion

Thomas E. Shannon

Introduction

Nuclear fusion holds the promise of providing almost unlimited power for future generations. If the process can be commercialized as envisioned by reactor design studies (Najmabadi et al., 1994), many of the problems associated with central electric power stations could be eliminated. Fusion power plants would not produce the pollution caused by the burning of fossil fuel and would eliminate the concern for meltdown associated with nuclear fission. The amount of radioactive waste material produced by a fusion reactor will be much less than that of a fission reactor since there is essentially no radioactive ash from the fusion reaction. If **low activation advanced materials** such as silicon carbide composites can be developed for the reactor structural material, the problem of disposal of activated components can also be eliminated.

Fusion Fuel

Although a number of different atomic nuclei can combine to release net energy from fusion, the reaction of **deuterium and tritium** (D-T) is the basis of planning for the first generation of fusion reactors. This choice is based on considerations of reactor economy. The D-T reaction occurs at the lowest temperature, has the highest probability for reaction, and provides the greatest output of power per unit of cost (Shannon, 1989). The disadvantages of D-T as a fusion fuel are twofold. Tritium does not occur naturally in nature and must be bred in the fusion reactor or elsewhere. Second, tritium is a radioactive isotope of hydrogen with a relatively long **half-life** of 12.3 years. Since tritium can readily combine with air and water, special safety procedures will be required to handle the inventory necessary for a fusion reactor. There is hope that a less reactive fuel, such as deuterium alone (D-D) will eventually prove to be an economically acceptable alternative (Shannon, 1989).

Confinement Concepts

Magnetic fusion, based on the tokamak concept, has received the majority of research funding for fusion energy development. However, other magnetic fusion concepts, such as the stellarator, the spherical torus, reversed-field pinch, and field-reversed configurations, are being developed as possible alternatives to the tokamak (Sheffield, 1994). It may also be possible to develop fusion power reactors by inertial confinement concepts (Waganer et al., 1992). Research on these concepts has been done primarily in support of weapons development; therefore, the level of scientific understanding for power reactor applications is significantly less than that of magnetic fusion. The remainder of this discussion on reactor development, fusion energy conversion, and transport will consider only the tokamak magnetic fusion concept.

Tokamak Reactor Development

The tokamak device has proved to be the most effective means of producing the conditions necessary for magnetic fusion energy production. In 1994, researchers at the Princeton Plasma Physics Laboratory achieved in excess of 10 MW of D-T fusion power in a research tokamak, the Tokamak Fusion Test Reactor (TFTR). This accomplishment, coupled with worldwide progress in 40 years of magnetic fusion research, has established the scientific feasibility of the tokamak concept. The next major step, the International Thermonuclear Experimental Reactor (ITER) is being carried out under an international agreement among Europe, Japan, Russia, and the United States (Conn et al., 1992). A drawing of the ITER tokamak is shown in Figure 8.9.1. If the project is approved for construction, it will be in operation

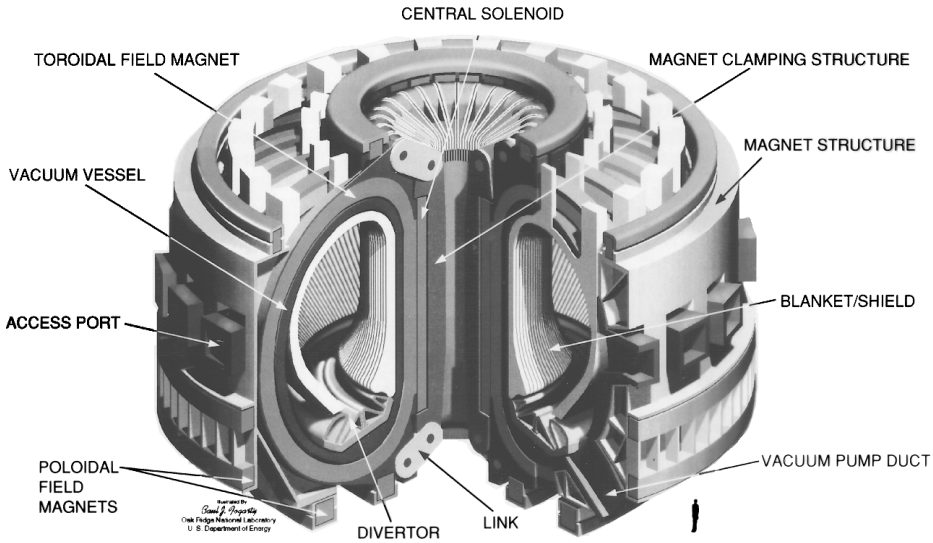


FIGURE 8.9.1 The International Thermonuclear Experimental Reactor (ITER).

around 2005. The ITER is being designed to produce a fusion power in excess of 1000 MW. This will be a significant step on the path to commercial fusion power.

The U.S. Department of Energy has proposed a strategy, shown in Figure 8.9.2, which will lead to a demonstration power reactor by the year 2025. Supporting research and development programs necessary to achieve this goal are shown in this figure.

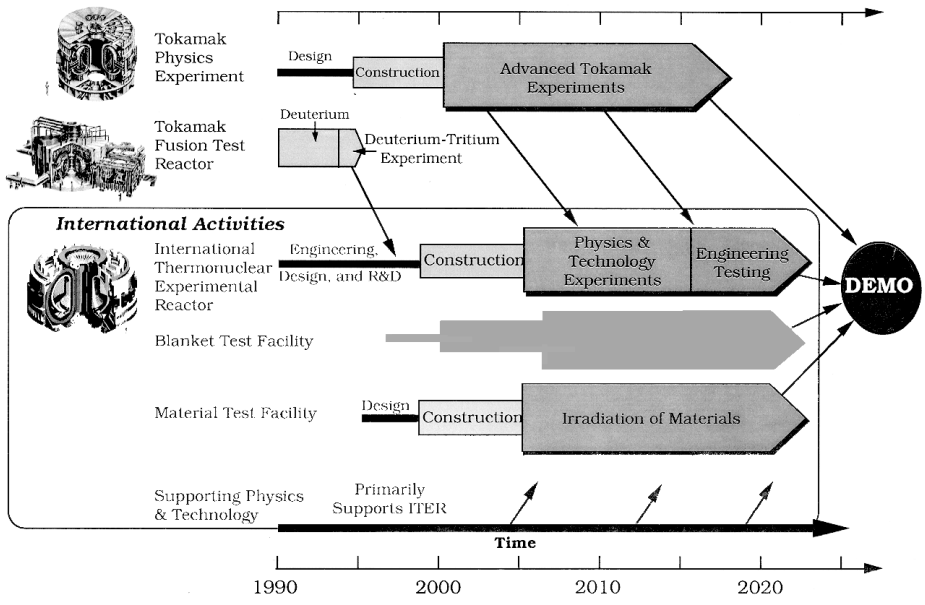


FIGURE 8.9.2 The U.S. Department of Energy magnetic fusion energy strategy.

Fusion Energy Conversion and Transport

The energy from fusion is created in the form of charged particles and neutrons. The D-T reaction produces a neutron with an energy of 14.1 MeV and an alpha particle (helium) with an energy of 3.5 MeV in the reaction

$$D + T \rightarrow He^4\ (3.5\ MeV) + n(14.1\ MeV)$$

In the tokamak device, the reaction will take place in the toroidal vacuum vessel as previously shown in the ITER drawing, Figure 8.9.1. The D-T fuel, in the form of a **plasma**, will absorb energy from the positively charged alpha particles to sustain the temperature necessary for the reaction to continue. The neutron, having no charge, will escape from the plasma and pass through the wall of the vessel and penetrate into the surrounding blanket/shield structure. The kinetic energy of the alpha particles from the fusion reaction is eventually deposited on the wall of the vacuum vessel by radiation and conduction heat transfer from the plasma while the neutron deposits most of its energy within the cross section of the blanket/shield. The resulting thermal energy is transferred by a coolant such as water to a steam generator where a conventional steam to electric generator system may be used to produce electricity. An overall schematic diagram of the energy conversion and heat-transport system is shown in Figure 8.9.3.

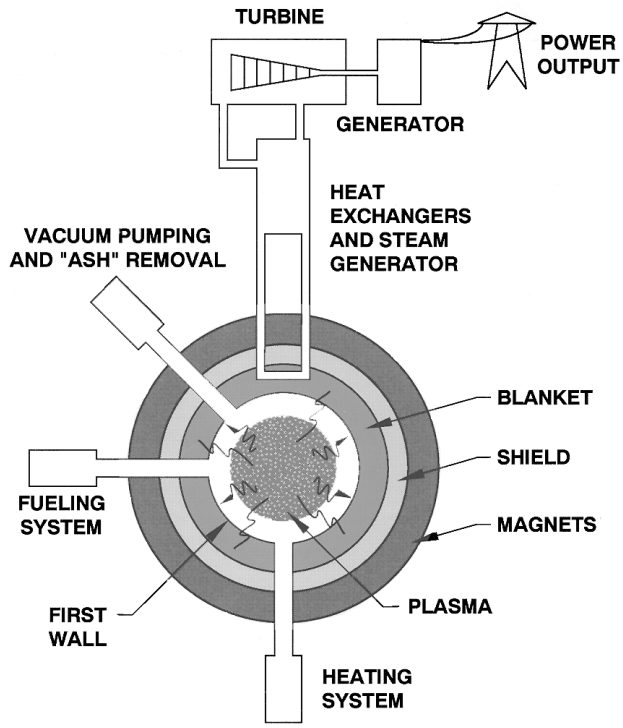


FIGURE 8.9.3 Schematic Diagram of a Magnetic Fusion Reactor Power Plant.

Defining Terms

Deuterium and tritium: Isotopes of hydrogen as the fuel for fusion reactors.
Half-life: The time required for half of the radioactive material to disintegrate.
Low activation advanced materials: Structural materials that significantly reduce the radioactivity induced by exposure to fusion neutrons.

Plasma: A gas such as a mixture of deuterium and tritium raised to a very high temperature at which the electrons and the nuclei of the atoms separate. The plasma, consisting of electrons and ions, can conduct electricity and react to magnetic fields.

References

Conn, R.W., Chuyanov, V.A., Inoue, N., and Sweetman, D.R. 1992. The International Thermonuclear Experimental Reactor, *Sci. Am.* 266(4).

Najmabadi, F. et al. 1994. The ARIES-II and -IV Second Stability Tokamak Reactors, University of California, Los Angeles, report UCLA-PPG-1461.

Shannon, T.E. 1989. Design and cost evaluation of a generic magnetic fusion reactor using the D–D fuel cycle. *Fusion Technol.* 15(2), Part 2B, 1245–1253

Sheffield, J. 1994. The physics of magnetic fusion reactors. *Rev. Mod. Phy.* 66(3).

Waganer, L. et al. 1992. Inertial Fusion Energy Reactor Design Studies. U.S. Department of Energy Report. Vol. I, II, and III, DOE/ER-54101 MDC 92E0008.

Further Information

The U.S. Department of Energy, Office of Fusion Energy maintains a home page on the World Wide Web. The address http://wwwofe.er.doe.gov provides an excellent source of up-to-date information and access to information from most institutions involved in fusion research.

8.10 Solar Thermal Energy Conversion

D. Yogi Goswami

Introduction

Solar thermal energy applications such as space and water heating have been known for a long time. Researchers over the past few decades have developed a number of additional solar thermal applications, such as industrial process heat, refrigeration and air-conditioning, drying and curing of agricultural products, and electric power production by solar thermal conversion. This section will cover solar thermal energy conversion including solar thermal collectors and conversion systems.

Solar Thermal Collectors

A simple **solar thermal collector** consists of (1) an absorber surface (usually a dark, thermally conducting surface), (2) some insulation behind the surface to reduce heat loss, (3) a trap for thermal reradiation from the surface such as glass, which transmits the shorter-wavelength solar radiation but blocks the longer-wavelength radiation from the absorber, and (4) a heat-transfer medium such as air, water, etc. High-temperature collectors require reflectors of sunlight that concentrate solar radiation on the absorber. The technology of solar collectors is developed to achieve temperatures as high as 1000°C or even higher. The design of a solar collector and the choice of working fluids depend on the desired temperature and the economics of the application. Table 8.10.1 lists the types of solar thermal collectors based on their temperature range.

TABLE 8.10.1 Types of Solar Collectors and Their Typical Temperature Range

Type of Collector	Concentration Ratio	Typical Working Temperature Range (°C)
Flat plate collector	1	≤70
High-efficiency flat plate collector	1	60–120
Fixed concentrator	2–5	100–150
Parabolic trough collector	10–50	150–350
Parabolic dish collector	200–2000	250–700
Central receiver tower	200–2000	400–1000

Source: Compiled from Goswami, D.Y., *Alternative Energy in Agriculture,* Vol. 1, CRC Press, Boca Raton, FL, 1980.

Flat Plate Collectors

Flat plate collectors may be designed to use liquids (water, oil, etc.) or air as the heat-transfer fluid. Figure 8.10.1 shows a typical liquid-type flat plate collector. The choice of materials for glazing and absorber needs special attention.

Glazing. The purpose of **glazing** is to (1) transmit the shorter-wavelength solar radiation, but block the longer-wavelength reradiation from the absorber plate, and (2) reduce the heat loss by convection from the top of the absorber plate. Glass is the most widely used glazing material. Transmittance of low iron glass in the visible and near infrared wavelength range can be as much as 91%, while for the longer-wavelength radiation (>3 μm) its transmittance is almost zero. Other materials than can be used as glazings include certain plastic sheets such as polycarbonates (Lexan® and Tuffac® — transmittance ~75%), acrylics (Plexiglass® and Lucite® — transmittance ~92%), and thin plastic films such as polyethylene. A major advantage of the plastics is that they are shatterproof; however, they scratch easily and lose transparency over time.

Absorbers. Copper is the most common material used for absorber plates and tubes because of its high thermal conductivity and high corrosion resistance. For low-temperature applications such as swimming pool heating, a plastic material called ethylene propylene polymer (trade names EPDM, HCP,

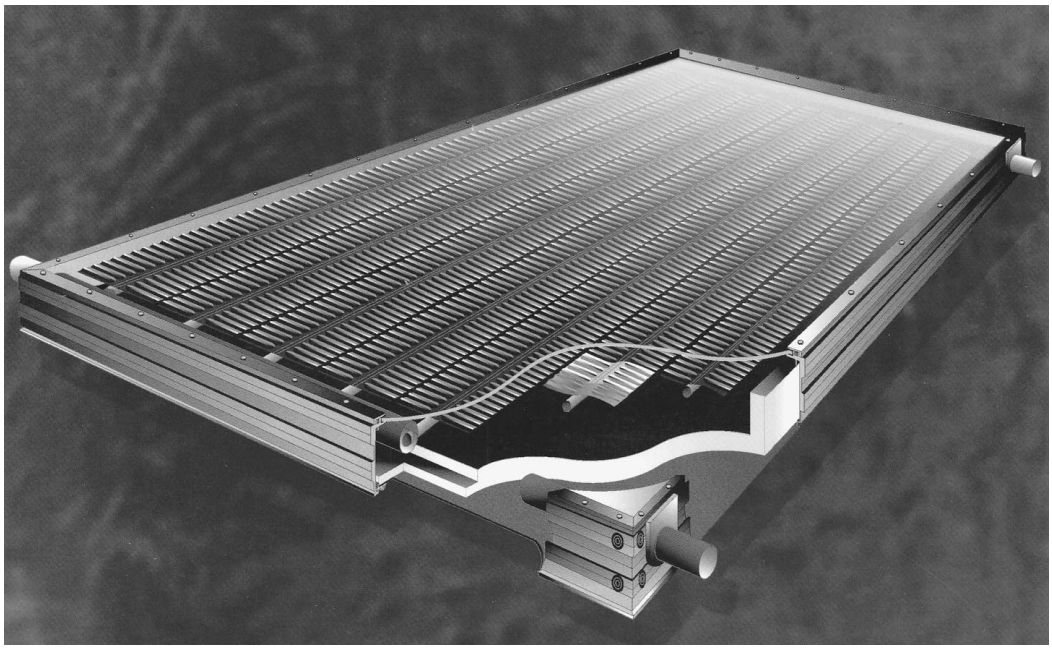


FIGURE 8.10.1 A typical liquid flat plate collector. (Courtesy of American Energy Technologies, Green Cove Springs, FL.)

etc.) can be used to provide inexpensive absorber material. To compensate for low thermal conductivity of these materials, a large surface area is provided for heat transfer. In order to increase the absorption of solar radiation and to reduce the emission from the absorber, the metallic absorber surfaces are painted or coated with flat black paint or some selective coating. Absorptivities and emissivities of some common **selective surfaces** are given in Table 8.10.2.

TABLE 8.10.2 Absorptivity and Emissivity of Common Selective Surfaces

Surface	Absorbtivity	Emissivity
Black chrome	0.95	0.1
Black nickel	0.9	0.08
Copper oxide	0.9	0.17
Lead sulfide	0.89	0.2
Flat black paint	0.98	0.98

Source: Compiled from Duffie, J.A. and Beckman, W.A., *Solar Engineering of Thermal Processes*, John Wiley and Sons, New York, 1980.

Evacuated Tube Collectors. Evacuated tube collectors have essentially a vacuum between the absorber and the glazing tube. This eliminates most of the heat loss by conduction and convection. Therefore, these collectors give a very high efficiency at higher temperatures. Evacuated tube collectors are typically used in the temperature range of 80 to 140°C

Concentrating Collectors. Concentrating collectors use reflectors or lenses to focus solar radiation from a large area onto a small area, thereby creating higher temperatures. Such collectors are usually used for temperatures higher than 100°C. Figure 8.10.2 shows schematics of some of the concentrating collectors.

Nontracking Concentrators. The simplest concentrating collector can be made by using flat wall reflectors to concentrate the solar radiation on a flat plate collector. Concentration ratios of two to three can be

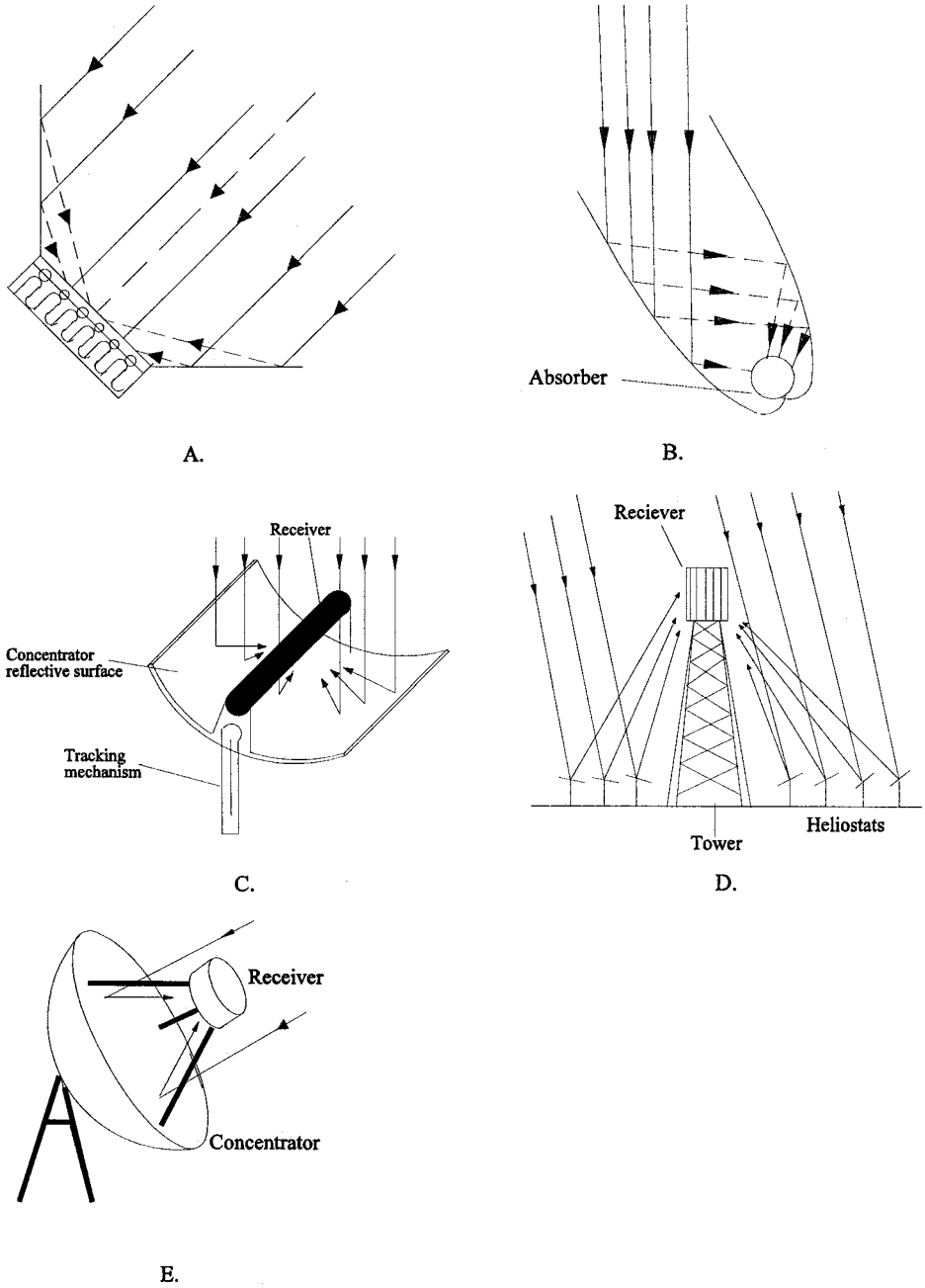


FIGURE 8.10.2 Types of concentrating collectors. (A) Flat plate collector with reflective wings; (B) Compound parabolic concentrator; (C) parabolic trough; (D) central receiver; (E) parabolic dish.

achieved this way. For slightly higher concentration ratios, a novel design, developed by Roland Winston, called a "compound parabolic concentrator" (CPC) can be used (Winston, 1974).

Tracking Concentrators. For temperatures up to 350°C, cylindrical parabolic trough collectors are used. These collectors focus solar radiation on a line focus where the absorber is located. These collectors usually require tracking on one axis only with seasonal adjustment on the other axis. A reflecting spherical or paraboloidal bowl is used when temperatures of the order of 250 to 500°C are needed. These collectors

require two-axis tracking. In some cases, the dish is kept stationary while the receiver is moved to track the focus of the reflected solar radiation. Finally, for extremely high temperatures (500 to 1000°C) needed for large-scale thermal power generation, a large field of tracking flat mirrors (called heliostats) is used to concentrate solar radiation on a receiver that is located on top of a central tower.

Collector Thermal Performance

The instantaneous efficiency of a collector is given by

$$\eta = \frac{\text{Useful energy collected}}{\text{Incident solar energy}} = \frac{Q_u / A}{I} \tag{8.10.1}$$

where

$$Q_u = mC_p\left(T_o - T_i\right) \tag{8.10.2}$$

and A = area of the collector, I = incident solar energy per unit area, m, C_p, T_i, and T_o are the mass flow rate, specific heat, and inlet and outlet temperatures of the heat-transfer fluid.

The efficiency of a flat plate solar collector can also be written by the Hottel–Whillier–Bliss equation:

$$\eta = F_R(\tau\alpha)_e - F_R U_L \frac{\left(T_i - T_{\text{amb}}\right)}{I} \tag{8.10.3}$$

where F_R, called the collector heat-removal factor, is a convenient parameter that gives the ratio of the actual useful energy gain of a flat plate collector to the useful gain if the whole collector surface were at the inlet fluid temperature; $(\tau\alpha)_e$ = effective transmittance absorptance product; and U_L = collector heat-loss coefficient.

Equation 8.10.2 suggests that if the efficiency, η, is plotted against $(T_i - T_{\text{amb}})/I$, the resulting curve will have a y intercept equal to $F_R\,(\tau\alpha)_e$ and a slope of $F_R U_L$. A linear curve usually gives an adequate approximation. Figure 8.10.3 shows an example of a performance curve for a water-heating flat plate collector, which is a linear least square curve fit to the actual test data.

Solar Ponds

A solar pond combines collector and energy storage in one unit. Solar radiation that enters the pond travels some distance through the water before being totally absorbed, thus increasing the temperature of the water at that depth. The heat thus collected can be stored in the pond by creating a stagnant, transparent, insulating layer in the top part of the pond. The most common method is by the addition of a salt into the pond to increase the density of water in the lower section of the pond. This type of pond is called a **salt gradient solar pond**. Reid (1987) reviewed the progress in the design and applications of salt gradient solar ponds. Figure 8.10.4 shows a schematic of a salt gradient pond along with a density profile in the pond.

The theory of salt gradient solar ponds has been described by Tabor (1981). The most important aspect of such ponds is the creation and maintenance of the salt gradient. Theory shows that the condition for maintaining stability is

$$\frac{\partial S}{\partial Z} > -1.14 \frac{\partial \rho}{\partial T} \times \frac{\partial T}{\partial Z} \frac{\partial \rho}{\partial S} \tag{8.10.4}$$

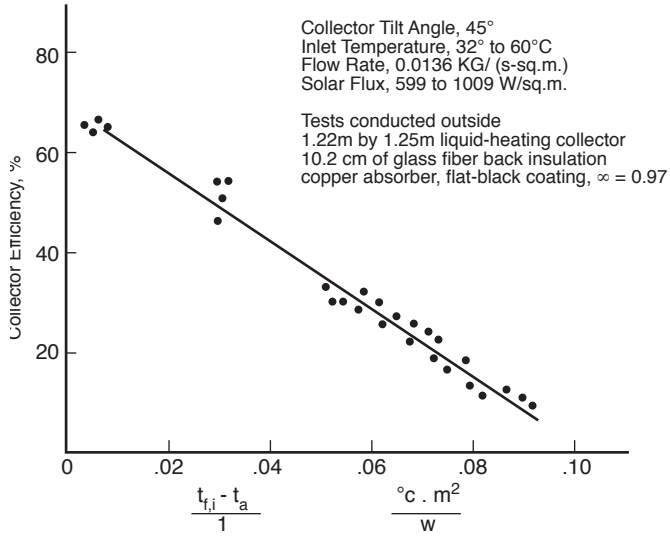


FIGURE 8.10.3 Thermal efficiency curve for a double-glazed flat plate liquid-type of solar collector. (Reprinted by permission of the American Society of Heating, Refrigerating and Air-Conditioning Engineers, Inc., Atlanta, GA, from ASHRAE Standard 93-77, *Methods of Testing to Determine the Thermal Performance of Solar Collectors*, 1977.)

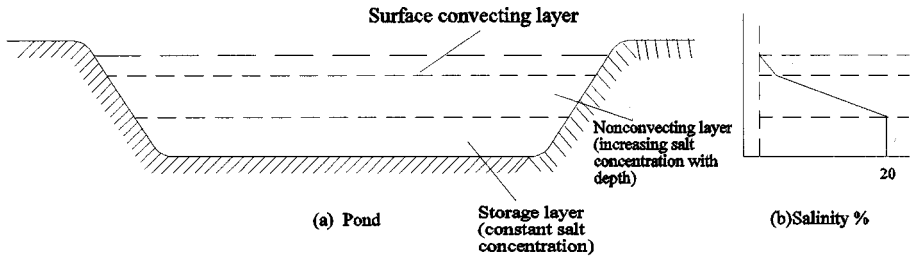


FIGURE 8.10.4 Schematic of a salt gradient solar pond.

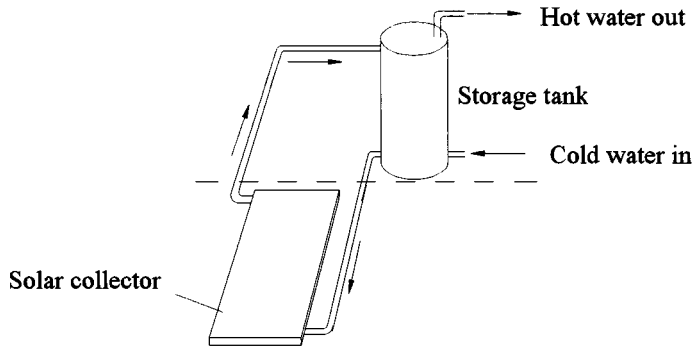
where S is the salt concentration in kilograms per cubic meter, Z is the depth from the surface in meters, ρ is the density in kilograms per cubic meter, and T is the temperature in Kelvin. The two most common salts considered for solar pond applications are NaCl and $MgCl_2$. According to the above criteria, there is no difficulty in obtaining stability with $MgCl_2$ and it is somewhat difficult but possible to get stability with NaCl.

Solar Water-Heating Systems

Solar water-heating systems represent the most common application of solar energy at the present time. Small systems are used for domestic hot water applications while larger systems are used in industrial process heat applications. There are basically two types of water-heating systems: **natural circulation or passive solar system** (thermosyphon) and **forced circulation or active solar system**.

Natural Circulation

Figure 8.10.5 shows a schematic of a natural circulation solar water-heating system. It is also called a thermosyphon or passive solar water heater because it does not require a pump to circulate water. The storage tank is located above the collector. When the water in the collector gets heated, it rises into the tank, because of density change, setting up a circulation loop.



FIGURES 8.10.5 Schematic of a thermosyphon solar water-heating system.

Forced Circulation

Figure 8.10.6 shows three configurations of forced circulation systems: (1) open loop, (2) closed loop, and (3) closed loop with drainback. In an open loop system, the collectors are empty when they are not providing useful heat and the storage tank is at atmospheric pressure. A disadvantage of this system is the high pumping power required to pump the water to the collectors every time the collectors get hot. This disadvantage is overcome in the pressurized closed loop system (Figure 8.10.6B) since the pump has to overcome only the resistance of the pipes. Because water always stays in the collectors of this system, antifreeze (propylene glycol or ethylene glycol) is required for locations where freezing conditions can occur. During stagnation conditions (in summer), the temperature in the collector can become very high, causing the pressure in the loop to increase. This can cause leaks in the loop unless some fluid is allowed to escape through a pressure-relief valve. In both cases, air enters the loop causing the pump to run dry. This disadvantage can be overcome in a closed loop drainback system (Figure 8.10.6C). In this system, when the pump shuts off, the water in the collectors drains back into a small holding tank, which can be located where freezing does not occur. The holding tank can be located at a high level to reduce pumping power.

Industrial Process Heat Systems

For temperatures of up to about 100°C, required for many industrial process heat applications, forced circulation water-heating systems described above can be used. The systems, however, will require a large collector area, storage and pumps, etc. For higher temperatures, evacuated tube collectors or concentrating collectors must be used.

Space-Heating Systems

Solar space-heating systems can be classified as active or passive depending on the method utilized for heat transfer. A system that uses pumps and/or blowers for fluid flow in order to transfer heat is called an *active system*. On the other hand, a system that utilizes natural phenomena for heat transfer is called a *passive system*. Examples of passive solar space-heating systems include direct gain, attached greenhouse, and storage wall (also called Trombe wall).

Active space-heating systems store solar collected heat in water or rocks. Heat from water storage can be transferred to the space by convertors or by fan-coil units. A system using a fan-coil unit can be integrated with a conventional air system as shown in Figure 8.10.7. Heat from rock storage can be transferred to air by simply blowing air over the rocks.

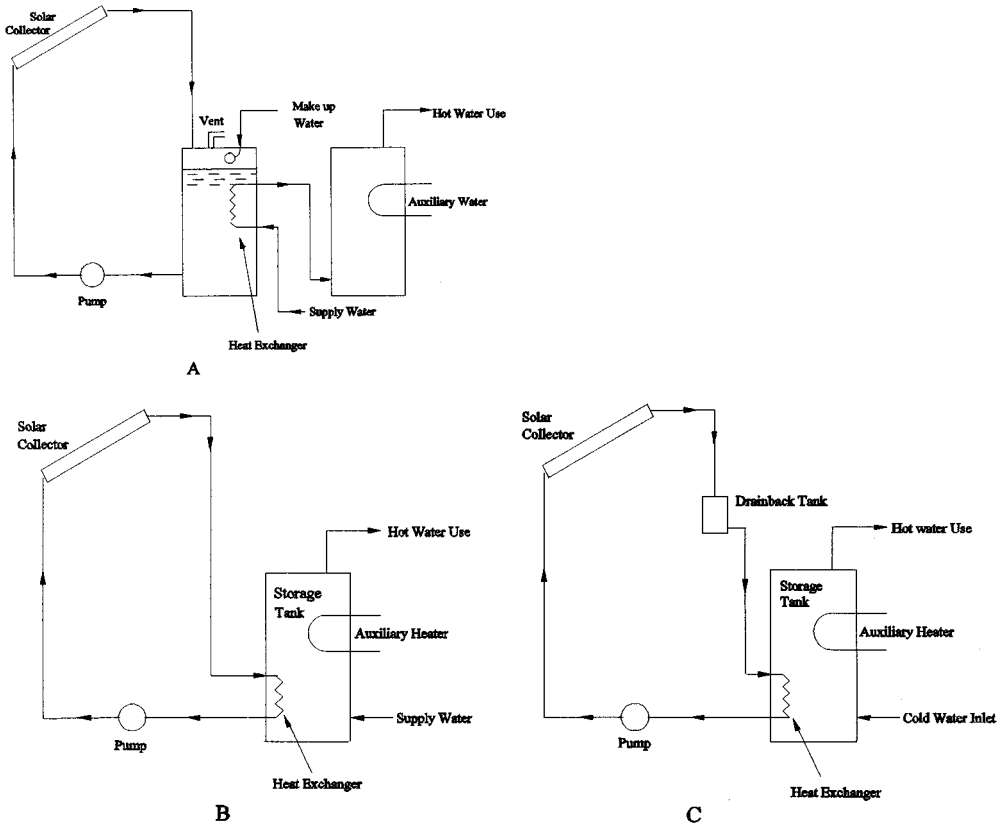


FIGURE 8.10.6 Typical configurations of solar water-heating systems: (A) open loop system, (B) closed loop system, (C) closed loop drainback system. (Adapted from Goswami, D.Y., *Alternative Energy in Agriculture,* Vol. 1, CRC Press, Boca Raton, FL, 1986.)

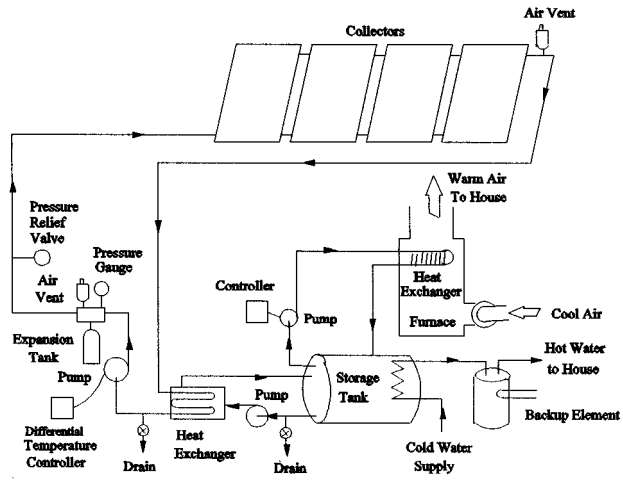


FIGURE 8.10.7 Schematic of an active solar space-heating system.

Solar Thermal Power

Solar thermal energy can be used to produce electrical power using conventional thermodynamic power cycles such as Rankine, Stirling, and Brayton cycles. The choice of power cycle and the working fluids depends on the temperature obtainable in the solar system, which depends on the type of solar collectors used. At present, developed solar thermal power systems include

Parabolic trough systems
Central receiver systems
Parabolic dish-Stirling engine system

Parabolic Trough Systems

Parabolic trough systems are simple in concept and, therefore, the most developed commercially. In 1984, Luz Company installed a Solar Electric Generating System (SEGS I) of 14 MW_e capacity in Southern California, utilizing parabolic trough solar collectors and natural gas fuel for superheat and backup. From 1984 to 1991, Luz Company installed eight additional plants, SEGS II to SEGS IX, with the total capacity of the nine plants being 354 MW_e. With each successive SEGS plant the technology was improved and the cost reduced. The cost of electricity was reduced from about 30¢/kWhr for the first plant to about 8¢/kWhr for the last plant. A schematic of the SEGS IX is shown in Figure 8.10.8, and some important data for the system are given in Table 8.10.3.

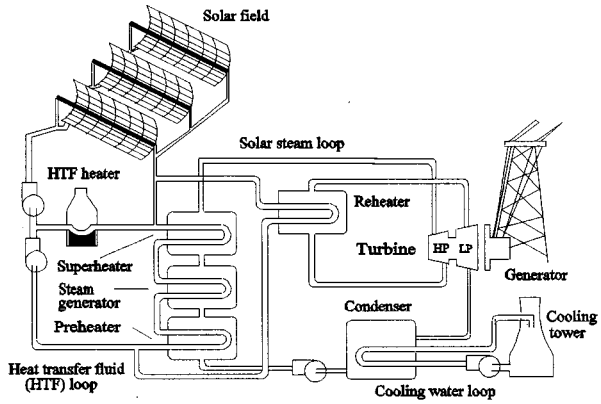


FIGURE 8.10.8 Schematic of SEGS IX power plant.

TABLE 8.10.3 Plant Characteristics — SEG IX

Power Block		Solar Field	
Gross power	88 MWe	Number of collectors	888
Net power	80 MWe	Aperature area	483,960 m^2
Steam inlet pressure	100 bar	Inlet temperature	293°C
Steam inlet temperature	371°C	Outlet temperature	390°C
Reheat pressure	17.2 bar	Annual thermal efficiency	50%
Reheat temperature	371°C	Peak optical efficiency	80%
Conversion efficiency	37.6%	Heat-transfer fluid (HTF)	Oil (VP-1)
Annual gas use	25.2×10^9 m^3	HTF volume	1289 m^3

Source: DeLaquil, P. et al., in *Renewable Energy Sources for Fuel and Electricity*, Island Press, Washington, D.C., 1993. With permission.

It has been recognized that this design does not utilize the energy of the natural gas efficiently. It has been suggested that energy of natural gas can be better utilized by combining the solar system with a natural gas turbine combined-cycle power plant (DeLaquil et al., 1993; Washom et al., 1994). Such a

hybrid system would use the exhaust of a natural gas turbine for superheating and preheating of water, while the solar field would be used for steam generation. Such a hybrid system can achieve conversion efficiency as high as 60%. A schematic of a **solar hybrid combined cycle** is shown in Figure 8.10.9.

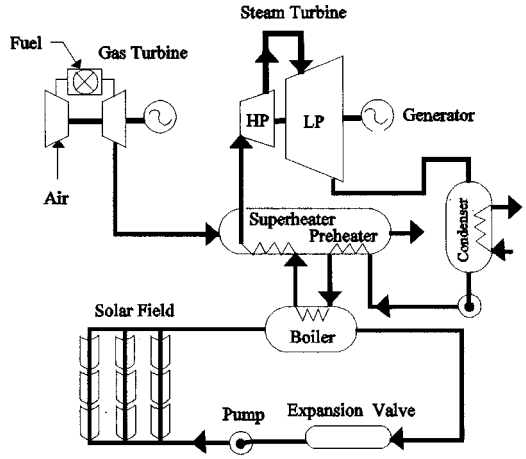


FIGURE 8.10.9 Solar hybrid combined cycle. (Adapted from Washom, B. et al., paper presented at the 1994 ASES Annual Conference, San Jose, CA, 1994.)

Central Receiver System

A central receiver system can potentially operate at very high temperature and therefore can have much higher efficiency than parabolic trough systems. However, the system can be economical only at larger capacities, such as 100 MW and above. The central receiver absorber can heat the working fluid or an intermediate fluid to a temperature as high as 600 to 1000°C which can be used to drive a Rankine cycle or Brayton cycle.

Solar One, a 10-MW_e central receiver power plant started operating in 1982 in Barstow, California. This plant generated superheated steam at 510°C and 10.3 MPa in the receiver absorber, which was used to drive a simple steam Rankine power cycle. The plant operated successfully for 6 years and provided a good learning experience. The plant has now been redesigned as Solar Two in which molten sodium nitrate is used as the heat-transfer fluid as well as for storage. Use of molten salt allows operation of the receiver absorber at much lower pressures The constraint is that the salt must always be above its melting point (220°C). Figure 8.10.10 shows a schematic of the Solar Two power plant.

Parabolic Dish Systems

Parabolic dish systems can achieve very high temperatures. The high temperatures produced by parabolic dishes can be used to drive Rankine, Stirling, and Brayton power cycles. So far, Rankine and Stirling cycles have been successfully demonstrated with parabolic dishes for electrical power production.

Early versions of parabolic dishes were made from die-stamped aluminum petals made reflective using a metallized polymer film. Later designs used simpler flat mirror facets fixed on a structure in such a way as to approximate a parabolic dish. The latest designs use a polymer film stretched on a circular frame (Mancini, 1994). The film is given a slight concave curvature by providing a vacuum behind it. These stretched polymer films are fixed on a structure to approximate a parabolic dish. Because of the low weight of the polymer film, the dish structure can be made out of light tubular members, thereby reducing the cost of the dish considerably. Parabolic dishes require two-axis tracking.

McDonnell Douglas Corporation successfully demonstrated a 25-KW_e parabolic dish system using a Stirling engine and a generator assembly fixed at the focal point of the dish in 1985 (Gupta, 1987). The concept is very attractive because it provides a modular design of stand-alone power plants for small

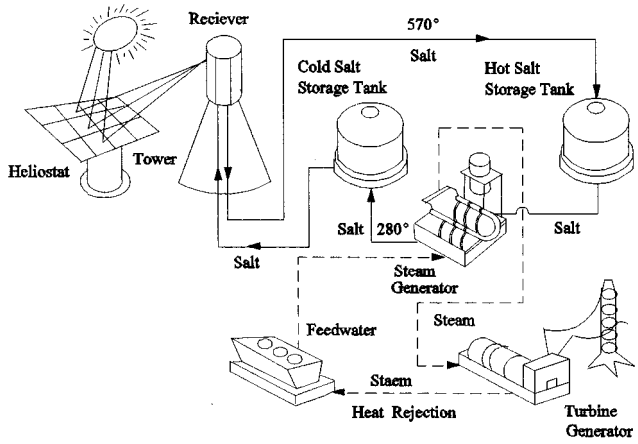


FIGURE 8.10.10 Schematic of Solar Two.

communities and independent power applications. Cummins Power Generation Company, in Indiana, has developed this concept further (Mancini, 1994). The Cummins system uses stretched polymer film facets for a parabolic dish, a heat pipe receiver absorber, a free-piston Stirling engine, and a linear alternator, resulting in a very compact power generation system. Figures 8.10.11 shows the latest version of the Cummins Dish Stirling power system. A detailed discussion of Stirling engines is given in Section 8.5.

Nomenclature

A = area of collector
C_p = specific heat
m = mass flow rate
T = temperature
I = incident solar radiation
Qn = useful heat collected
F_R = collector heat-removal factor
U_L = collector heat-loss coefficient
S = salt concentration
Z = depth

Greek Symbols

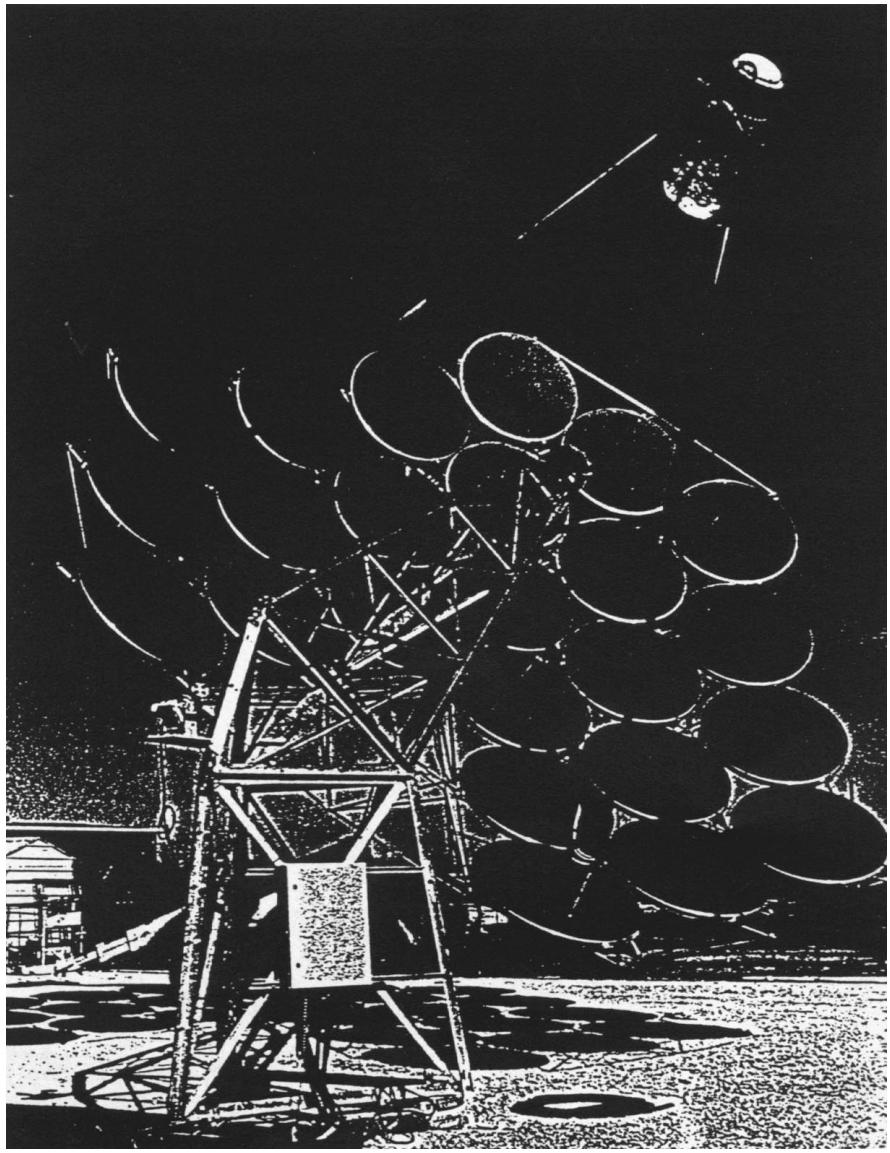
α = absorptance
τ = transmittance
η = efficiency
ρ = density

Subscripts

i = inlet
o = outlet
amb = ambient
e = effective

Defining Terms

Forced circulation or active solar system: A solar thermal system that uses active means, such as pumps, for fluid flow and heat transfer.



Figures 8.10.11 A parabolic dish — Stirling Engine Power System.

Glazing: A material used in a solar collector that transmits short-wavelength solar radiation and blocks longer-wavelength reradiation from the absorber plate.

Natural circulation passive solar system: A solar thermal system that uses natural means, such as the thermosyphon effect, for fluid flow and heat transfer.

Salt gradient solar pond: A solar pond that uses high salt concentration in the lowest layer and a concentration gradient in the middle layer to make those layers nonconvective.

Selective surface: A surface that has high absorptance in short wavelengths and low emittance in longer wavelengths.

Solar hybrid combined cycle: A hybrid of solar and natural gas turbine combined cycle.

Solar thermal collector: A solar collector that absorbs sunlight and converts it to heat.

References

ASHRAE. 1977. ASHRAE Standard 93-77, *Method of Testing to Determine the Thermal Performance of Solar Collectors*, ASHRAE, Atlanta, GA.

DeLaquil, P., Kearney, D., Geyer, M., and Diver, R. 1993. Solar thermal electric technology, Chapter 5 in *Renewable Energy Sources for Fuel and Electricity*, T.B. Johansson, M. Kelly, A.K.N. Reddy, and R.H. Williams, Eds., Island Press, Washington, D.C.

Duffie, J.A. and Beckman, W.A. 1980. *Solar Engineering of Thermal Processes*, John Wiley & Sons, New York.

Goswami, D.Y. 1986. *Alternative Energy in Agriculture,* Vol. I, CRC Press, Boca Raton, FL.

Gupta, B.P. 1987. Status and progress in solar thermal research and technology, in *Progress in Solar Engineering*, Ed. D.Y. Goswami, Hemisphere Publishing, Washington, D.C.

Mancini, T.R. 1994. The DOE solar thermal electric program, in *Proceedings of the 1994 IECEC*, pp. 1790–1975. AIAA, Washington, D.C.

Reid, R.L. 1987. Engineering design of salt gradient solar pond for thermal and electric energy, in *Progress in Solar Engineering*, Ed. D.Y. Goswami, Hemisphere Publishing, Washington, D.C.

Tabor, H. 1981. Solar ponds. *Solar Energy*, 27(3), 181.

Washom, B., Mason, W., Schaefer, J.C., and Kearney, D. 1994. Integrated Solar Combined Cycle Systems (ISCCS) Utilizing Solar Parabolic Trough Technology — Golden Opportunities for the 90s, paper presented at the 1994 ASES Annual Conference, San Jose, CA.

Winston, R. 1974. Principles of solar concentrators of novel design. *Solar Energy*, 16(2), 89.

Further Information

For solar heating and cooling:

Solar Engineering of Thermal Processes, by J.A. Duffie and W.A. Beckman, John Wiley & Sons, New York, 1980.

Principles of Solar Engineering, by F. Kreith and J.F. Kreider, Hemisphere Publishing, a division of Taylor and Francis, Washington, D.C., 1978.

For solar thermal power:

Solar Energy Fundamentals and Design, by W.B. Sine and R.W. Harrigan, John Wiley & Sons, New York, 1985.

8.11 Wind Energy Conversion*

Dale E. Berg

Introduction

Wind energy conversion machines have evolved over the past 2000 years, mostly by trial and error. Although there are many different configurations of wind machines, most of them can be classified as either horizontal-axis wind turbines (HAWTs), which utilize rotors that rotate about a horizontal axis parallel to the wind, or vertical-axis wind turbines (VAWTs), which have rotors that rotate about a vertical axis. Figure 8.11.1 illustrates the main features of both HAWTs and VAWTs. Figure 8.11.2 shows both types of turbines in a wind farm in the Altamont Pass area of California. HAWTs have all of their drivetrain equipment located on a tower, which makes servicing somewhat difficult, their blades are subjected to cyclic stresses due to gravity as they rotate, and they must be oriented with respect to the wind. However, they may be placed on tall towers to access the stronger winds typically found at greater heights. VAWTs, on the other hand, have most of their drivetrain on the ground, do not experience cyclic gravitational stresses, and do not require orientation with the wind. VAWTs, however, cannot be placed on tall towers to exploit the stronger winds at greater height, and their blades are subject to severe alternating aerodynamic loading due to rotation. The most common type of modern HAWT is the propeller-type machine, and these machines are generally classified according to the rotor orientation (upwind or downwind of the tower), blade articulation (rigid or teetering), and number of blades (generally two or three). The most common types of modern VAWTs are the Darrieus, with curved blades that are fixed in pitch, and the fixed-pitch, straight-bladed machines. The following discussion will focus on these types of turbines.

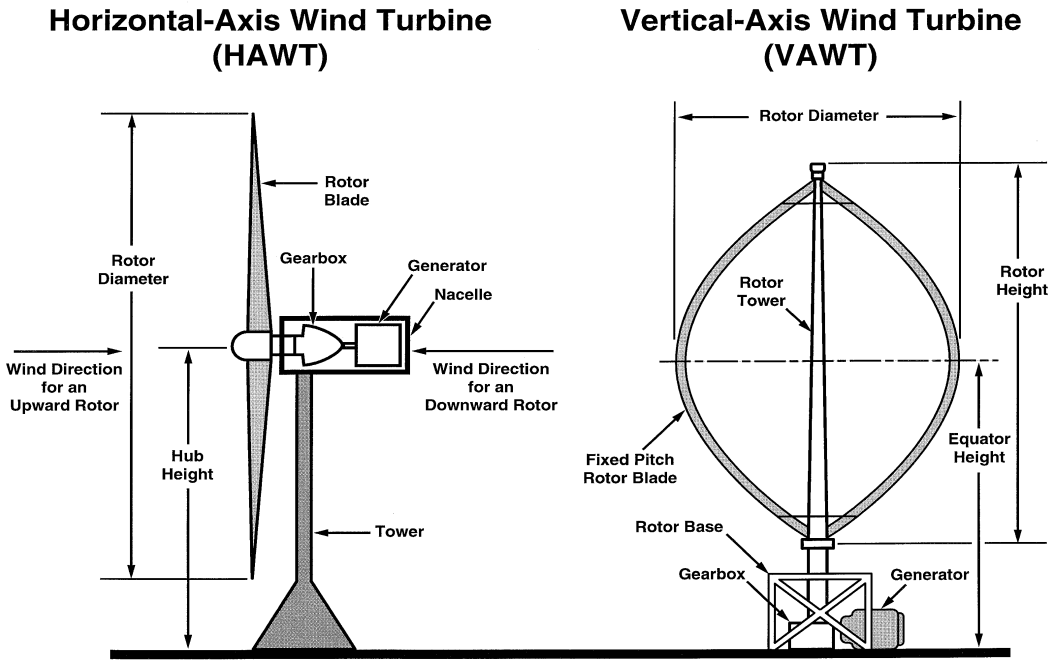


FIGURE 8.11.1 Wind turbine configurations.

* This work was supported by the United States Department of Energy under Contract DE-AC04-94AL 85000.

FIGURE 8.11.2 Wind farm both horizontal-axis and vertical-axis turbines.

While small and even medium-sized machines have been developed primarily through trial and error, developing larger, more-complex, and highly efficient machines this way becomes very expensive and time-consuming. A large cost-effective machine can be developed at a reasonable cost only if the designers can accurately predict the performance of conceptual machines and investigate the effects of design alternatives. In the past two decades numerous techniques to predict the aerodynamic and structural dynamic performance of wind turbines have been developed. These analytical models are not, in general, amenable to simple approximations, but must be solved with the use of computer codes of varying complexity. These models and codes will be summarized in the following sections.

Wind Turbine Aerodynamics

Items exposed to the wind are subjected to both drag (in the direction of the wind) and lift (perpendicular to the wind) forces. The earliest wind machines used drag to produce power. The European and American windmills discussed in Section 7 of Chapter 7 were primarily drag devices, but they did make some use of lift. Modern wind turbines rely on airfoil-shaped blades that generate large amounts of lift to produce power more efficiently than the drag machines. Let us consider how efficient these machines are at extracting energy from the wind.

Figure 8.11.3 illustrates the flow field about a translating drag device. The drag results from the relative velocity between the wind and the device, and the power that is generated by the device (the product of the drag force and the translation velocity) is given by

$$P = Dlv = \left[0.5\rho(U - v)^2\right]C_D clv \tag{8.11.1}$$

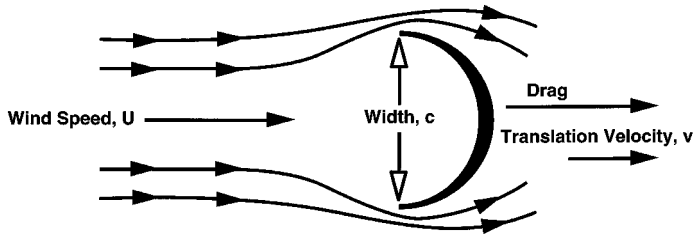


FIGURE 8.11.3 Schematic of translating drag device.

where

P = power extracted in watts
D = drag force per unit spanwise length in n/m
l = length of device (distance into the page) in m
v = translation velocity in m/sec
ρ = air density in kg/m^3
U = steady free-stream wind velocity in m/sec
C_D = drag coefficient; function of device geometry function
c = width of device (perpendicular to wind) in m

The velocity of the device must always be less than the wind velocity, or no drag is generated. The **power coefficient** (the ratio of the power extracted to the power available in the area occupied by the device) for this machine is

$$C_p = \frac{P}{0.5\rho U^3 cl} = \frac{v}{U}\left[1 - \frac{v}{U}\right]^2 C_D \tag{8.11.2}$$

Now consider a device that utilizes lift to extract power from the wind. Figure 8.11.4 illustrates an airfoil that is translating at right angles to the wind direction and is subject to both lift and drag forces. The relative velocity across this surface is the vector sum of the free-stream wind velocity and the wind speed induced by translation. The angle between the direction of the relative velocity and the chord line of the airfoil is termed the angle of attack α. In this case, the power is given by

$$P = 0.5\rho U^3 cl \frac{v}{U}\left[C_L - C_D \frac{v}{U}\right]\sqrt{1 + \left(\frac{v}{U}\right)^2} \tag{8.11.3}$$

where c = airfoil chord length in m and C_L, C_D = lift and drag coefficients, respectively; functions of airfoil shape and α. The power coefficient then is

$$C_p = \frac{v}{U}\left[C_L - C_D \frac{v}{U}\right]\sqrt{1 + \left(\frac{v}{U}\right)^2} \tag{8.11.4}$$

Figure 8.11.5 compares Equations (8.11.2) and (8.11.4) using $C_L = 1.0$ and $C_D = 0.10$ for the airfoil (easily achieved with modern airfoils) and a maximum drag coefficient of 2.0 for the drag machine. The airfoil has a maximum power coefficient of 15, compared with 0.3 for the drag device, or 50 times more power per unit of projected area. Moreover, operating a lifting device at velocities well in excess of the wind velocity is easily achieved with rotating machines.

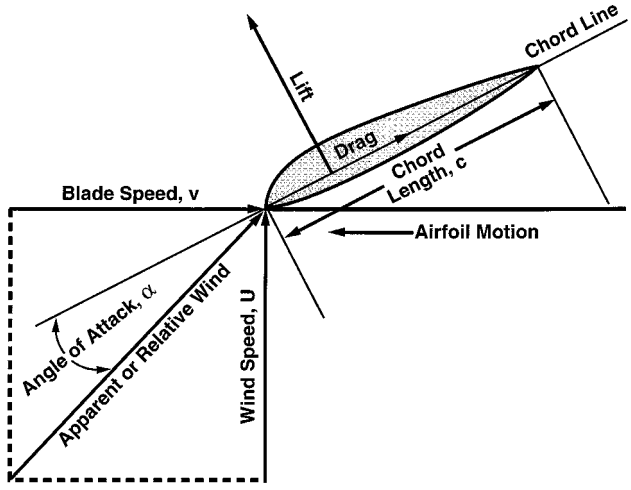


FIGURE 8.11.4 Schematic of translating lift device.

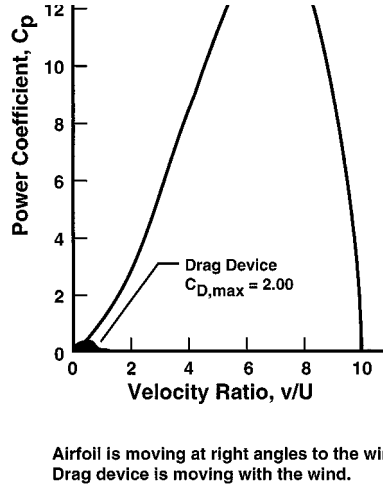


FIGURE 8.11.5 Comparison of power coefficients for a translating airfoil and a translating drag device. The airfoil is moving at right angles to the wind direction. The drag device is moving in the wind direction.

Machines discussed in this section utilize lift-producing blades to capture wind energy.

Aerodynamic Models

The aerodynamic analysis of a wind turbine has two primary objectives: (1) to predict the power produced by of the turbine and (2) to predict the detailed aerodynamic loads which will act on the turbine rotor blades. In general, the same models are used to accomplish both objectives. The aerodynamics of wind turbines are far too complex to model with simple formulas that can be solved with hand-held calculators; computer based models ranging from very simplified to very complex are required. The various models commonly used today are described below.

Momentum Models

The simplest aerodynamic model of a wind turbine is the actuator disk model or **momentum theory** in which the turbine is modeled as a single porous disk. In this model the axial force acting on the rotor or disk is equated to the time rate of change of momentum of the airstream passing through the rotor

or disk. By utilizing the conservation of mass, the conservation of axial momentum, the Bernoulli equation, and the first law of thermodynamics and by assuming isothermal flow, the power produced by the turbine (the product of the axial force and the air velocity at the disk) may be found to be

$$P = 2\rho AV^3 a(1-a)^2 \tag{8.11.5}$$

where V is the freestream wind velocity, $a = (V - v)/V$ and v is the wind velocity at the disk. The power coefficient for the turbine becomes

$$C_p = 4a(1-a)^2 \tag{8.11.6}$$

This is maximized for $a = {}^1/_3$, and we get $C_{p,\max} = 16/27 = 0.593$, the **Betz limit**, as the maximum fraction of available energy that can be extracted from the wind by a turbine.

The typical performance of various types of wind machines is compared with the Betz limit in Figure 8.11.6 where the variation of the turbine power coefficients with the **tip–speed ratio** (the ratio of the speed of the blade tip to the free-stream wind speed) are presented. Even though the maximum performance of modern HAWTs and VAWTs is well above that of the older machines, it still falls more than 10% below the Betz limit.

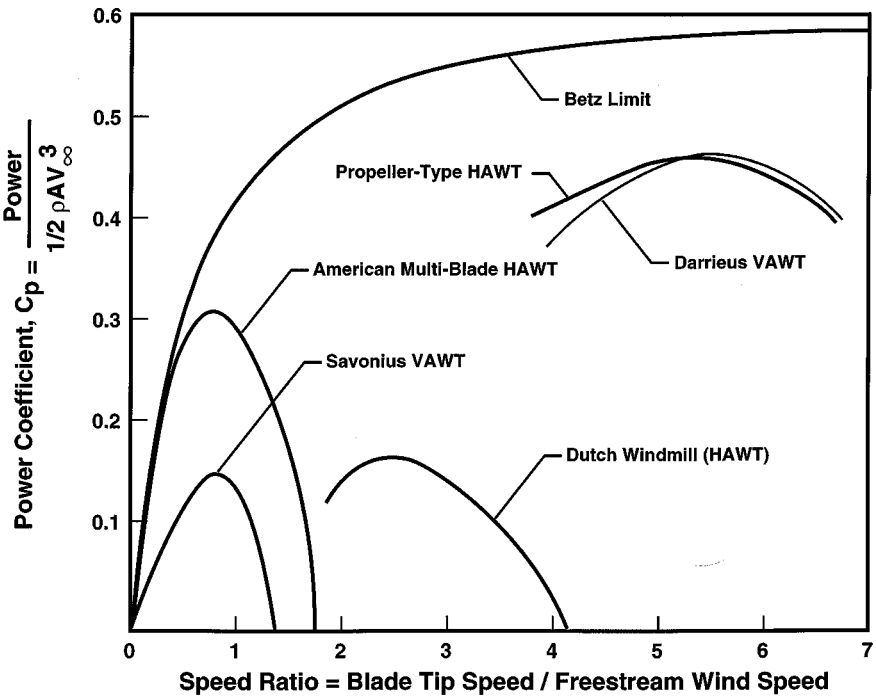


FIGURE 8.11.6 Typical performance of various types of wind turbines.

For HAWTs, momentum theory can be expanded to the blade element or strip theory, which includes the effects of blade lift and drag, wake rotation, and number and type of blades. Numerous corrections are applied to account for the three-dimensional flow near blade tips, the thick blade sections near the root, and gaps along the blade span. Additional information on these models may be found in Hansen and Butterfield (1993) and Wilson (1994).

Momentum theory may also be expanded for vertical-axis turbines into the multiple streamtube and the double-multiple streamtube theories that are the VAWT equivalent of the HAWT blade element theory. Additional information on these models may be found in Touryan et al. (1987) and Wilson (1994).

Wind shear and local Reynolds number variations may be readily accounted for with both the blade element and multiple streamtube methods. These models are extremely popular with wind turbine designers because they are simple, fast, and fairly accurate for performance prediction. However, they are very approximate methods based upon the assumption of steady flow and streamtubes that are fixed in time and space. More-complex models such as vortex and local circulation models are needed for the analysis of yawed flow, unsteady aerodynamics, and other complex flows, all of which can have large impacts on turbine performance and loads.

Vortex Models

Vortex models, based on the vorticity equation, can use either lifting line or lifting surface formulations for the blades with either free-wake or fixed- (or prescribed-) wake models, although there are a number of variations of these models. The three-dimensional, lifting-surface, free-wake formulation is the most physically realistic model, but a computer program implementing such a model will require a tremendous amount of computer resources and time. The problem with vortex codes is one of finding a balance between model simplification (and limitation) and computation time. Additional information on vortex models may be found in Strickland et al. (1981) and Kocurek (1987).

Local Circulation Method

The local circulation method (LCM) utilizes a balance between the force on the blade and the change in wind momentum as it passes through the rotor, similar to what is done with the momentum models. The blade, however, is represented as a superposition of imaginary blades of different spans with elliptical circulation distributions. Unlike the streamtube models, LCM models may be formulated to analyze unsteady flow, and are able to yield detailed flow field velocity and blade-loading information. The LCM yields better answers than the momentum models, avoids the convergence problems of the vortex models, and, with an appropriate wake model, requires far less computer time than the vortex models. However, this model has not been widely used. Additional information may be found in Nasu and Azuma (1983), Masse (1986), and Oler (1989).

Common Model Limitations

All of the aerodynamic models in use today use airfoil section characteristic tables (lift and drag coefficients as functions of angle of attack and Reynolds number) to determine the blade loading and turbine performance. Static two-dimensional wind tunnel test results or two-dimensional static airfoil design code predictions are modified with empirical, semiempirical, or analytic methods and used to estimate blade loads under three-dimensional, dynamic conditions. The greatest difficulty in obtaining accurate load predictions with any performance code is the determination of the appropriate airfoil section characteristics.

Additional information and references on turbine aerodynamics may be found in Hansen and Butterfield (1993) and Wilson (1994) for HAWTs and in Touryan et al. (1987) and Wilson (1994) for VAWTs.

Wind Turbine Loads

Wind turbine aerodynamic loads are of two types — the narrowband harmonic or cyclic loads resulting from the steady atmospheric wind, wind shear, rotor rotation, and other steady effects; and the broadband random loads resulting from nonuniformity or turbulence in the wind. The prediction of these loads is more complicated than the prediction of the aerodynamic performance and requires the use of computer-based models. The harmonic loads are generally predicted with the same codes that are used to predict wind turbine performance. The random loads are typically estimated with empirical relations, although a few analysts do utilize a performance code with a nonuniform wind model to predict them. A wind

turbine will experience hundreds of millions of loading cycles in a 30-year lifetime, and small errors that lead to underpredicting component loads can result in costly short-term component failure.

Accurate prediction of turbine performance does not guarantee accurate prediction of detailed aerodynamic loads — the performance predictions result from the integration of loads over the entire turbine, and significant errors may be present in the detailed loads but balance out in the performance predictions. While there is a considerable body of data showing good agreement of predicted performance with measured performance, there are very few data available against which to compare detailed aerodynamic load predictions.

Wind Turbine Dynamics

Horizontal-Axis Turbines

Horizontal-axis turbine designs usually use fairly rigid, high-aspect-ratio blades, cantilevered from a rigid hub and main shaft, although they may sometimes use relatively slender, quite flexible blades, attached to a less rigid hub and/or main shaft. This assembly rotates and yaws about a tower which may be flexible. These structures have many natural vibration modes, and some of them may be excited by the wind or the rotation frequency to cause a **resonance** condition in which the vibrations are amplified and cause large stresses in one or more components. Careful structural analysis during the design can ensure that the turbine that is built is dynamically stable under turbine operating conditions. Ignoring the structural analysis or failing to properly conduct parts of it will likely result in a machine that experiences resonances and fails very quickly. Relatively rigid systems are less likely to experience these stability problems than are very flexible, highly dynamic systems.

Detailed analysis of the structural response of a turbine is a rather daunting task requiring the formulation and solution of the full governing equations of motion, usually performed with a finite-element structural model. Those equations must account for the interaction between the blades and the steady centrifugal forces, the time-dependent gravitational forces, the steady and oscillatory aerodynamic forces, and the Coriolis forces. They must also model nonsteady airflow, the yaw motion of the nacelle, pitch control of the blades, teetering blades, the interaction between the rotor and the supporting tower, starting and braking sequences, etc. The rotor must be modeled in a rotating coordinate frame with time-dependent coefficients, while the tower must be modeled in a fixed coordinate frame, with the exciting forces arising from the nonlinear, time-dependent coupling of the rotor and the time-dependent loading of the wind.

Malcolm and Wright (1994) provide a list of some of the available HAWT dynamics codes that have been developed, together with their limitations.

Vertical-Axis Turbines

Darrieus turbine designs normally use relatively slender, high-aspect-ratio structural elements for the blades and supporting tower. The result is a very flexible, highly dynamic structure, with many natural modes of vibration which, again, must be carefully analyzed to ensure that the turbine is dynamically stable under all operating conditions. Typically, the guy cables and turbine support structure can be analyzed with conventional methods, but the tower and blades require a more refined analysis, usually performed with a finite-element structural program.

The blades and tower must be modeled in the rotating coordinate frame with time-independent coefficients. The equations of motion are determined by the steady centrifugal and gravitational forces, the steady and oscillatory aerodynamic forces, and the Coriolis forces, together with the turbine physical properties. Detailed information on the modeling may be found in Lobitz and Sullivan (1983).

Aerodynamic Loads/Blade Motion Coupling

The blades themselves are driven by aerodynamic as well as structural dynamic forces. The motion of slender, high-aspect-ratio blades may couple with the aerodynamic loads acting on them. This coupling may increase the motion of the blades, creating a potentially fatal condition known as flutter instability,

or it may decrease the motion of the blades, creating a beneficial condition known as aerodynamic damping.

Stochastic Wind Effects

The wind is **stochastic** in nature, with significant short-term variations in both direction and velocity. As turbines become larger, the relative extent of these variations becomes smaller than the size of the turbine, and the effects become more pronounced. Analysis of the effect of fluctuating wind loads on the response of the turbine shows an increase in the broadband response, accompanied by a decrease in the magnitude of the dominant narrowband responses at multiples of the rotation frequency. This increase in broadband response can include excitation of turbine vibration modes that are close to a narrowband response frequency, but that are not predicted to be excited by a uniform wind (Lobitz, 1984).

Wind Turbine Controls

In general, wind turbines are designed to operate when the incident wind is high enough to generate electricity and to shut down when the wind speeds exceed 25 to 30 m/sec. In spite of the tremendous amount of power which is present in the high winds, the amount of energy that can be captured is usually more than offset by the fatigue damage that is sustained by the turbine.

Brakes

Most turbines utilize mechanical brakes, frequently in conjunction with aerodynamic brakes, to stop the rotor and to keep it from rotating when the turbine is not generating electricity. A few turbines rely solely on aerodynamic braking devices to accomplish this. Whatever type of braking system is used, it should be a fail-safe design that will automatically activate to slow or stop the rotor in the event of an electrical system failure.

Yaw Systems

Virtually all upwind and a few downwind HAWT turbines incorporate an active yaw control system, using wind-direction sensors and electric or hydraulic drive motors, to orient the rotor with respect to the wind. VAWTs do not require yaw systems.

Peak Power Regulation

All turbines incorporate some method of limiting the peak power produced. This enables the generator to operate near its design power rating, where it is most efficient, over a range of wind speeds. The increase in generator efficiency at lower wind speeds, together with the lower cost of the drivetrain, more than offset the energy that is lost as a result of power limiting. Most horizontal-axis turbines use one of three common techniques to limit peak power — stall regulation with fixed-pitch blades (passive control), full- or partial-span pitch control, or partial-span control surfaces such as ailerons and/or flaps. With stall regulation, the blades are designed so that airfoil **stall**, which creates decreased lift and increased drag, limits the power output in high winds. However, rotor drag loads continue to increase as the wind speed increases. Another disadvantage is the difficulty of controlling aerodynamic loads in deep stall. With full- or partial-span blade pitch control, peak power is controlled by decreasing blade pitch angle as wind speed increases, limiting peak power and decreasing rotor drag loads. A major disadvantage of pitch control is the poor peak power control during high-wind stochastic (or turbulent) conditions — power excursions can exceed twice the rated power levels before the pitch-control system can respond. Partial-span control surfaces limit the peak power by decreasing the lift and increasing the drag of a portion of the airfoil. They can respond to wind changes somewhat faster than full-span pitch control systems can. Sample power curves for both stall and pitch-regulated turbines are shown in Figure 8.11.7.

Several other methods of pitch control have also been used, but on a limited basis. Passive pitch-control techniques automatically adjust the blade pitch angle using cams activated by centrifugal loads or using tailored blade materials that permit the blade to twist as the aerodynamic loads increase. The

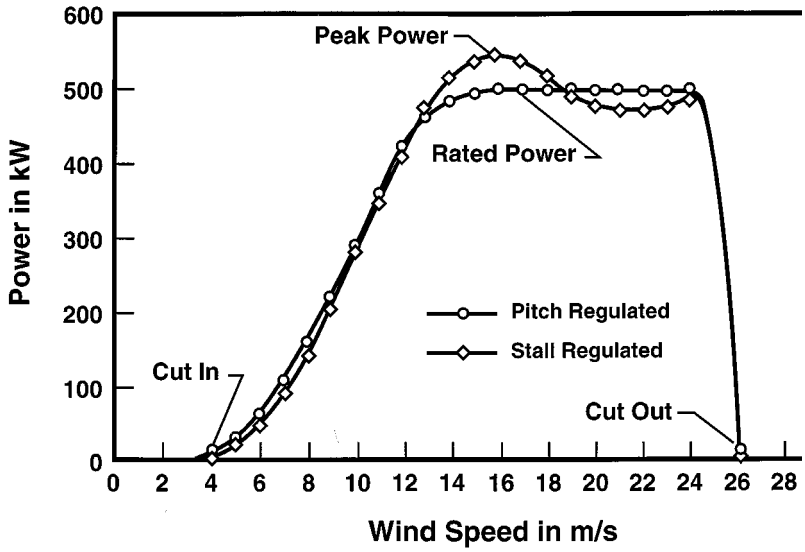


FIGURE 8.11.7 Sample power curves for stall-regulated and pitch-regulated wind turbines.

Italian Gamma 60 turbine utilizes a novel means to control peak power — it is the only large-scale turbine that yaws a fixed-pitch rotor out of the wind to limit rotor power.

Virtually all VAWTs utilize stall regulation with fixed-pitch blades to control peak power.

Controller

Every wind turbine contains a controller, usually a microprocessor-based system, to control turbine operations. The basic turbine controller will start and stop the machine and connect or disconnect the generator output lines to the grid, as needed; control the operation of the brake system; control the operation of the yaw and pitch systems, if present; perform diagnostics to monitor the operation of the machine; and perform normal or emergency shutdown of the turbine as required. However, the controller can incorporate many other functions as well, functions such as recording turbine performance characteristics and controlling the operating state of a variable-speed machine.

Wind Turbine Electrical Generators

Once a wind turbine has converted the kinetic energy in the wind into rotational mechanical energy, the energy is usually converted into electricity which can be readily transported to where it is needed. While small wind turbines may utilize permanent magnet alternators to generate electricity, most grid-connected turbines today use either synchronous or induction electrical generators. Induction machines are cheaper than synchronous machines, are easier to control, and provide some power train damping, but they require reactive power that must be supplied by the grid. This can cause problems, but those problems can be usually be solved fairly quickly and at low cost.

Generator efficiency drops off rapidly as the generated power falls below the rated generator capacity, and single-generator systems tend to be very inefficient at low wind speeds where there is little power available in the wind. Some systems address this by having a second, smaller generator which is used for low-wind operation, where it operates close to its rated power. At higher winds, the smaller generator is disconnected and the larger generator is used. Similar results can be obtained with a single generator utilizing pole switching or dual windings. The two-generator operation may yield a sizeable increase in energy capture, but the additional costs of the smaller generator or the generator modification and additional controls must be balanced against the increased energy capture to determine if this is cost-effective.

While most turbines operate at a single fixed rotational speed, some operate at two or more fixed rotational speeds, and some operate anywhere within a range of speeds. Variable-speed turbine operation offers two major advantages over fixed-speed operation:

1. The aerodynamic efficiency of the rotor at low to moderate wind speeds may be improved by more closely matching the rotor speed to the short-term average wind speed. At higher wind speeds, the blades are either in stall or are pitched to control peak power, so matching rotor speed to wind speed is less important.
2. System dynamic loads are attenuated by the "flywheel" action of the rotor as it speeds up and slows down in response to wind gusts.

In addition, variable speed permits the operation of the turbine in a variety of modes, including operation at maximum efficiency for all wind speeds to maximize energy capture or operation to minimize fatigue damage. However, certain rotational speeds within the operating-speed range will likely excite turbine vibration modes, causing resonance and increased rates of fatigue damage. These rotational speeds must be avoided during operation, and turbine control can become quite a complicated issue.

Variable-speed operation, in general, generates variable-frequency power. Most applications, including interfacing with power grids, require high-quality power at a reference frequency. Sophisticated power electronics may be used to accomplish this interface (Smith, 1989).

Wind-Diesel Systems

Wind turbines are increasingly being coupled with diesel engine-powered generators to create **wind-diesel systems**. In these systems, the diesel engine provides dependable, consistent power and the wind turbines generate some of the power, reducing the engine fuel consumption and the system cost of energy. A large variety of wind-diesel systems and concepts have been investigated over the past decade (Infield, et. al., 1992). Although quite a few of these have been technically successful, very few of them have been commercially successful. While the potential market for wind-diesel systems appears to be very large, it has yet to materialize, for the reliability of the technology has not yet been well proven.

Water-Pumping Applications

The multibladed, mechanical windmill (the so-called "American" windmill) with a mechanical piston pump has been used for over a century to pump water in remote areas. Over the past few years some changes to improve operation in low wind speeds have been made to the design of these machines.

Researchers at the U.S. Department of Agriculture facility at Bushland, Texas have reported considerable success in using variable-voltage, variable-frequency electricity produced by small stand-alone wind turbines to directly power submersible water pumps. Clark (1994) compares the pumping performance of one of these wind–electric systems and that of the traditional American windmill with a piston pump. He found that the wind–electric system performed significantly better than the mechanical system, even though the price of the two systems was nearly identical. The wind–electric system offers another advantage as well — while the windmill for the mechanical pump must be mounted directly over the well, the wind turbine for a wind–electric system may be mounted some distance away, at a better wind location.

Defining Terms

Betz limit: Maximum fraction of available wind energy that can be extracted by a wind turbine rotor, according to momentum theory.

Momentum theory: A method of estimating the performance of a turbine by equating the time rate of change of air stream momentum through the turbine to the force acting on turbine blades.

Power coefficient: The ratio of captured energy to the energy available in the reference area.

Resonance: A vibration of large amplitude caused by a relatively small excitation at or near a system natural frequency.
Stall: A condition in which an airfoil experiences a decrease in lift and a large increase in drag.
Stochastic: Containing variations from a smooth, uniform flow.
Tip-speed ratio: The ratio of the speed of the blade tip to the free-stream wind speed.
Wind–diesel system: An electrical-generation system that utilizes both diesel engine–powered generators and wind turbines to create a dependable, consistent power system.

References

Clark, R.N. 1994. Wind-electric water pumping systems for rural domestic and livestock water, in *Proceedings of the 5th European Wind Energy Association Conference and Exhibition,* Macedonia, Greece, pp. 1136–1140.

Hansen, A.C. and Butterfield, C.P. 1993. Aerodynamics of horizontal-axis wind turbines, *Ann. Rev. Fluid Mech.,* 25, 115–149.

Infield, D., Scotney, A., Lunsager, P., Binder, H., Uhlen,K., Toftevaag, T., and Skarstein, O. 1992. Wind diesel systems — design assessment and future potential, paper presented at Sixth International Wind–Diesel Workshop, Prince Edward Island, Canada.

Kocurek, D. 1987. Lifting surface performance analysis for horizontal axis wind turbines, *SERI/STR*-217, 3163.

Lobitz, D.W. 1984. NASTRAN-based software for the structural dynamic analysis of VAWTs and HAWTs, paper presented at European Wind Energy Conference, Hamburg, p. 385

Lobitz, D.W. and Sullivan, W.N. 1983. A comparison of finite element prediction and experimental data for forced response of DOE 100 kW VAWT, in *Proceedings of the Sixth Biennial Wind Energy Conference and Workshop*, Minneapolis, MN, pp. 843–853.

Malcolm, D.J. and Wright, A.D. 1994. The use of ADAMS to model the AWT-26 prototype, in *Proceedings of 1994 ASME Wind Energy Symposium*, New Orleans, LA, pp. 125–132

Masse, B. 1986. A local-circulation model for Darrieus vertical-axis wind turbines, *J. Propulsion Power,* 2 (March-April) 135–141.

Nasu, K. and Azuma, A. 1983. An experimental verification of the local circulation method for a horizontal axis wind turbine, paper presented at 18th Intersociety Energy Conversion Engineering Conference, Orlando, FL.

Oler, J.W. 1989. A discrete local circulation model for Darrieus turbines, in *Proceedings of the Eighth ASME Wind Energy Symposium,* Houston, TX, pp. 65–69.

Smith, G.A. 1989. Electrical control methods for wind turbines, *Wind Eng.,* 13(2), 88–98.

Strickland, J.H., Smith, T., and Sun, K. 1981. A vortex model of the Darrieus turbine: an analytical and experimental study, SAND81-7017, Sandia National Laboratories, Albuquerque, NM.

Touryan, K.J., Strickland, J.H., and Berg, D.E. 1987. Electric power from vertical-axis wind turbines, *J. Propulsion Power,* 3(6), 481–493.

Wilson, R.E. 1994. Aerodynamic behaviour of wind turbines, in *Wind Turbine Technology, Fundamental Concepts of Wind Turbine Engineering,* D. Spera, Ed., ASME Press, New York, 215–282.

Further Information

Excellent summaries of HAWT and VAWT aerodynamics, together with extensive reference lists, are presented by Craig Hansen and Sandy Butterfield in their paper "Aerodynamics of Horizontal-Axis Wind Turbines" in the *Annual Review of Fluid Mechanics*, 1993, and by Ken Touryan, Jim Strickland, and Dale Berg in their paper "Electric Power from Vertical-Axis Wind Turbines" in the *Journal of Propulsion,* Volume 3, Number 6, 1987.

The latest developments in the field of wind energy in the U.S. and Europe may be found in the following:

Proceedings of the ASME Wind Energy Symposium, published annually by the American Institute of Aeronautics and Astronautics, 1801 Alexander Bell Drive, Suite 500, Reston, VA 20191–4344.

Proceedings of Windpower, the annual American Wind Energy Association (AWEA) conference, published annually by AWEA, 122 C St. NW, 4th Floor, Washington, D.C. 20001.

Proceedings of the European Wind Energy Association, published annually by EWEA, Eaton Court, Maylands Avenue, Hemel Hempstead, Hertforshire HP2 7TR, England.

8.12 Energy Conversion of the Geothermal Resource

Carl J. Bliem and Gregory L. Mines

This section discusses the uses of the geothermal resource. The primary use of the energy from geothermal resources to date has been in the production of electrical energy. Other applications, such as process heat and space conditioning, have also been made and will be discussed under the topic of direct use. This section begins with a discussion of the geothermal resource as it applies to the use of the energy. Then discussion of the three types of electrical generating facilities presently in use: — the **direct steam system**, the **flashed steam system**, and the **binary system** — is given. Finally, some discussion of direct-use applications is given.

Geothermal Resource Characteristics Applicable to Energy Conversion

Geothermal energy as defined here applies to hot fluids under pressure found at a reasonable depth (1 to 2 km) in the earth's crust. If one disregards the complex geological details relating to the formation of such naturally occurring reservoirs of hot fluids, Figures 8.12.1 and 8.12.2 present schematic representations of these reservoirs. High-temperature fluid (200 to 300°C) is created by the convection of water through the porous rock. As the water circulates, it dissolves various amounts of minerals containing sodium, potassium, calcium, silica, carbonates, and chlorides and gases such as nitrogen and carbon dioxide. In **geopressured resources** of the Gulf of Mexico, high pressures and significant amounts of dissolved methane are seen.

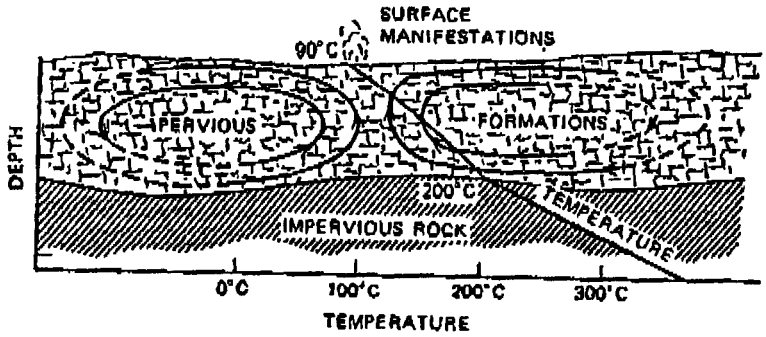


FIGURE 8.12.1 Schematic diagram of the convective cells in a geothermal reservoir. (From Kestin, J., Ed., Sourcebook on the Production of Electricity from Geothermal Energy, U.S. DOE, DOE/RA/4051-1, Washington, D.C., 1980).

The convective cells operate over large horizontal distances of as much as 30 km. The time in which the transfer of energy from the magma to the water takes place is of the order of 10^5 to 10^6 years. At the present time, it is difficult to say whether or not the resource can be considered "renewable." If natural circulation or the injection of spent geothermal liquid into the reservoir can make up for the liquid extracted during the energy conversion process, the reservoir can be considered at least of a very long life. (Individual wells generally have a life of about 10 years.)

The resources considered in this section are said to be **hydrothermal**. (Work is being done on creating artificial reservoirs by injecting water into hot dry rock, but this development is in its early stages and will not be considered here. The geopressured resource will not be considered either.)

As the geofluid is extracted from a reservoir, it flows to a region of lower static pressure. If this pressure falls below the saturation pressure for the temperature of the geofluid (close to but not equal to the saturation pressure of pure water because of the presence of the dissolved solids and gases), the

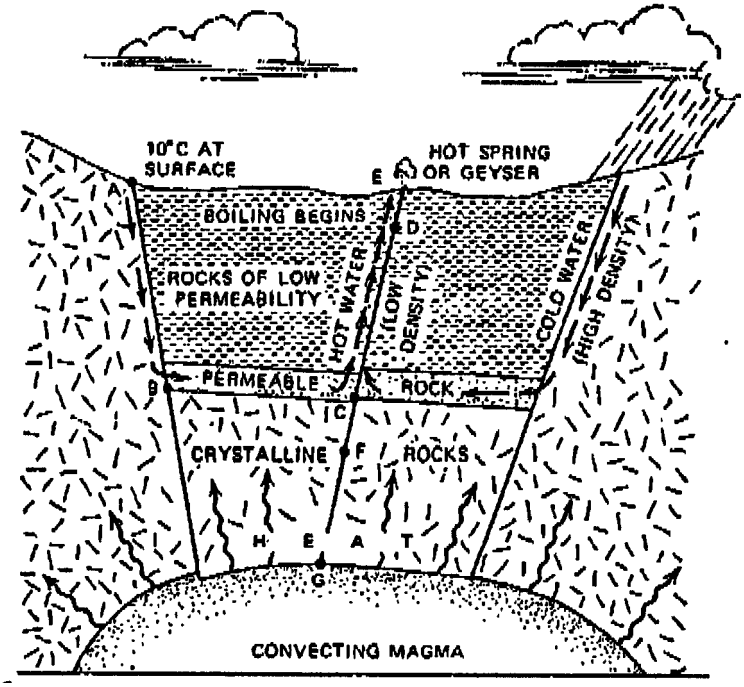


FIGURE 8.12.2 Schematic diagram of a characteristic geothermal reservoir. (From Kestin, J., Ed., Sourcebook on the Production of Electricity from Geothermal Energy, U.S. DOE, DOE/RA/4051-1, Washington, D.C., 1980).

geofluid will flash into steam. Therefore, the person using this energy source may have a number of different physical forms to consider:

1. Wet steam from a *vapor-dominated resource*;
2. Superheated or saturated steam from a vapor-dominated resource;
3. Liquid at a pressure above the saturation pressure from a *liquid-dominated resource*;
4. A mixture of liquid and vapor at a relatively low quality from a liquid-dominated resource.

Electrical Energy Generation from Geothermal Resources

The type of energy conversion system used to produce electrical power depends on the type and quality (temperature) of the geothermal resource. Vapor-dominated resources use systems where steam is expanded directly through a turbine, similar to conventional fossil fuel steam plants. Liquid-dominated resources use flash steam systems and binary systems, with binary systems predominantly used for the lower-quality resources. The term **binary system** is used to describe a power cycle where the geothermal fluid provides the source of thermal energy for a closed-loop Rankine cycle using a secondary working fluid. In this closed loop, the working fluid is vaporized using the energy in the geofluid, expanded through a turbine, condensed, and pumped back to the heater completing the closed loop.

Hydrothermal resources typically contain varying amounts of numerous dissolved minerals and dissolved gases. In power cycles where steam is extracted from the geothermal resource directly (vapor dominated) or indirectly (flashing liquid dominated) and expanded through a condensing turbine, the design and operation of the power cycle must account for the removal of the noncondensable gases. If the gases are not removed from the condenser, they will accumulate in the condenser, raising the turbine back pressure and decreasing the power output. In systems where the liquid geofluid is handled (binary cycle heat exchangers and piping and flash steam flash tanks and piping), measures must be taken to

prevent the precipitation of the dissolved solids and/or to provide a means of removal of the resulting scale.

Direct Steam Systems — Vapor-Dominated Resources

For a geothermal resource producing a superheated or saturated vapor steam (Case 2), the vapor from the geothermal production well is sent to a conventional steam turbine as shown in Figure 8.12.3. (This is done after appropriate removal of rocks and debris and possibly after scrubbing with water to remove corrosive substances.) Normally, the turbine is a condensing type, as shown in the figure, although in some applications a back-pressure turbine is used, exhausting the steam to the atmosphere. The back-pressure turbine is typically used for small systems with the possible later addition of another turbine and a condenser to increase the power generated by the geofluid flow from the wells.

Figure 8.12.3 shows a system with a direct-contact condenser and a wet cooling tower. In this type of system, the condensate from the condenser is more than enough to make up the evaporation and blowdown from the cooling tower. Therefore, the figure shows some of the condensate being injected into the reservoir. In many cases, direct-contact condensers are not feasible because of the hydrogen sulfide in the steam which would be released in the cooling tower exhaust. When hydrogen sulfide is in the steam, the majority of it appears as noncondensable and the noncondensable gas from the condensers must be treated. For these systems, surface condensers are normally used in conjunction with wet cooling towers. The actual hardware configuration is dictated by the process for removal of the sulfur. Again, some of the condensate can be used for cooling tower makeup if the sulfur is removed from the process. A number of processes have been developed to remove sulfur from the process.

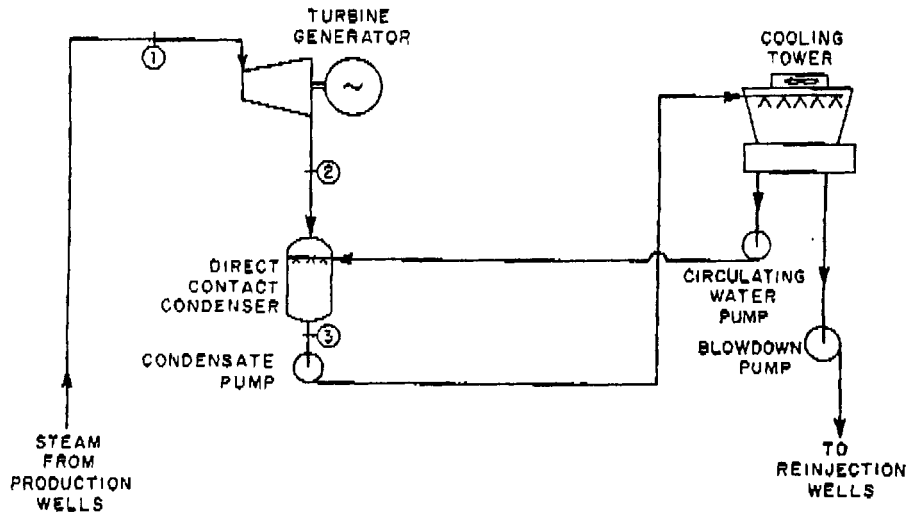


FIGURE 8.12.3 Schematic diagram of a direct dry-steam plant. (From Kestin, J., Ed., Sourcebook on the Production of Electricity from Geothermal Energy, U.S. DOE, DOE/RA/4051-1, Washington, D.C., 1980).

Figure 8.12.4 depicts a system which is similar to the one described above, but one which receives wet steam (Case 1). Here, the liquid is separated from the vapor prior to the entry of the vapor into the turbine. Otherwise, the system is the same as the one in Figure 8.12.3 and the same comments apply.

Flash Steam Systems — Liquid Dominated Resources

When the geofluid is flashed before it leaves the well, flash steam systems are generally used. This indicates that the resource is at a relatively high temperature. Figures 8.12.5 and 8.12.6 depict single- and dual-flash systems schematically. The single-flash system in Figure 8.12.5 is quite similar to the system in Figure 8.12.4. The only difference is that the geofluid pressure is dropped further before the steam is separated and sent to the turbine. An optimum flash pressure exists because the lower the flash

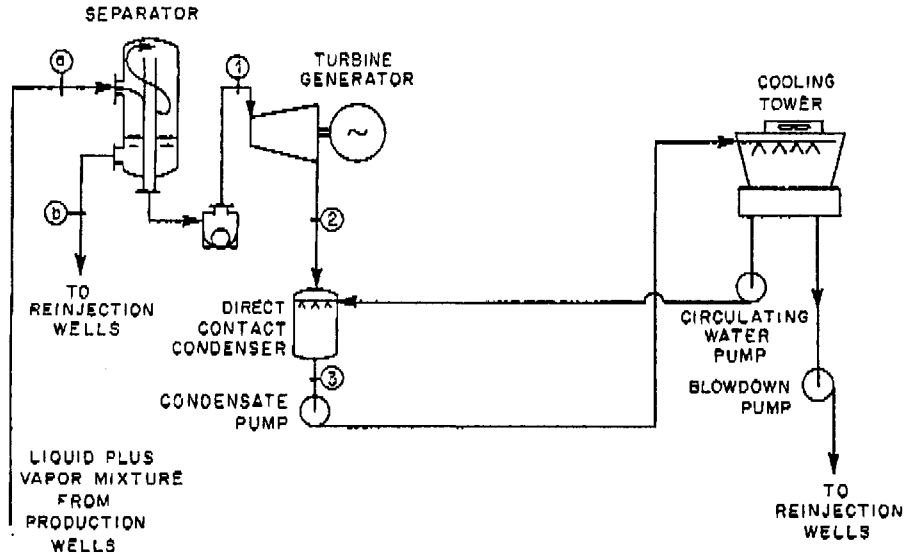


FIGURE 8.12.4 Schematic diagram of a plant using a two-phase resource. (From Kestin, J., Ed., Sourcebook on the Production of Electricity from Geothermal Energy, U.S. DOE, DOE/RA/4051-1, Washington, D.C., 1980).

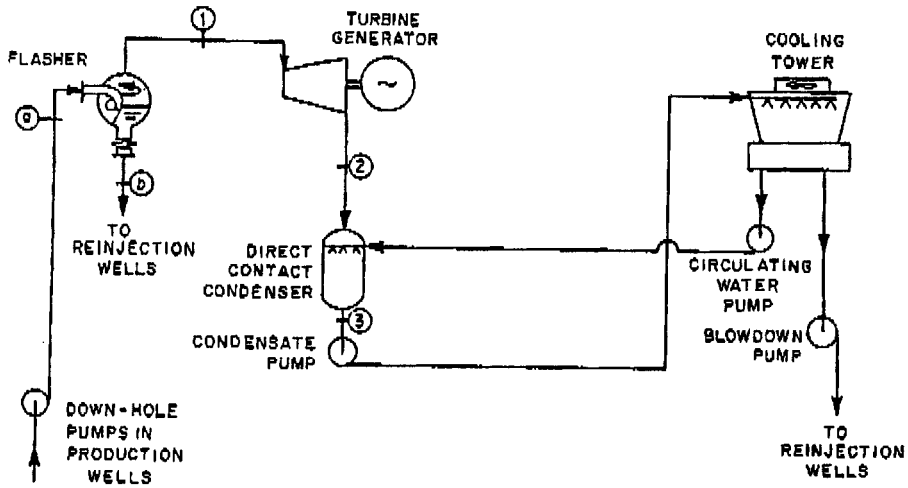


FIGURE 8.12.5 Schematic diagram of a single-flash plant. (From Kestin, J., Ed., Sourcebook on the Production of Electricity from Geothermal Energy, U.S. DOE, DOE/RA/4051-1, Washington, D.C., 1980).

pressure, the more steam which is evolved. However, the work done per unit mass of steam flowing through the turbine will also decrease with the lower flash pressure. For a given set of geofluid conditions entering the plant, a flash pressure exists that will maximize the energy produced per unit mass of geofluid and also minimize the levelized energy cost (LEC). The performance and cost optima will be near, but not generally at the same pressure.

The flash steam system can also be utilized in applications where the fluid enters the plant as a liquid (single phase). In these systems, the geothermal fluid is throttled with an expansion valve to the desired flash pressure. This flashing process can be considered adiabatic, where the amount of steam evolved can be determined from energy and mass balances of a simple throttling calculation.

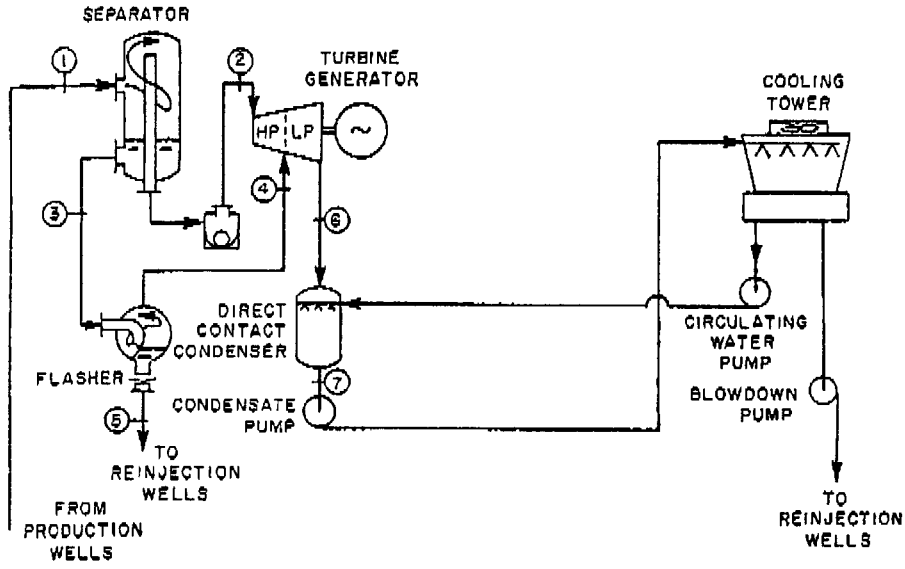


FIGURE 8.12.6 Schematic diagram of a dual-flash plant. (From Kestin, J., Ed., Sourcebook on the Production of Electricity from Geothermal Energy, U.S. DOE, DOE/RA/4051-1, Washington, D.C., 1980).

Most successful flash systems are dual flash. The first flash is generally near the well-head pressure and the second flash near atmospheric pressure. The low-pressure flash is normally kept above atmospheric pressure to prohibit leakage of air into the system in the flasher. Some recent studies have indicated that for low-temperature resources, a subatmospheric second flash would produce a cost-effective system. Again, optimization of the two flash pressures is necessary to minimize the LEC. In cases in which the geofluid has a high dissolved solid content, flash crystallizers are used to remove the precipitated dissolved solids. The flashing process releases carbon dioxide dissolved in lowering the geofluid pH, which causes the precipitation of insoluble carbonates. The solubility of silica is temperature dependent; lowering the geofluid temperature causes the precipitation of silica.

None of the steam cycles depicted provides for the removal of the noncondensable gases from the condenser. This removal is typically accomplished with steam ejectors or compressors which continuously remove the small stream of vapor from the condenser. Some steam is lost in this process of removing the noncondensable gases.

Binary Systems — Liquid-Dominated Resources

Recent studies have shown that for resources below 200°C, current technology binary systems have lower LEC than flash steam plants for liquid-dominated resources. Figure 8.12.7 shows a typical binary system with an evaporative heat-rejection system. This type of heat-rejection system has been replaced by air-cooled condensers in most applications. In the areas where the geothermal resource exists, there is little excess water for makeup in the cooling tower, as shown in Figure 8.12.7. All of the cooled geofluid in a binary system is typically injected back into the reservoir. This provides an environmentally acceptable means of disposal of the fluid and, more important, provides a recharge of the reservoir to maintain the reservoir productivity.

The binary cycle is an attempt to reduce the scaling potential of the geofluid. Carbonates are precipitated when the pressure of the geofluid is reduced and carbon dioxide comes out of solution as the geofluid flashes. With downhole pumps in the wells, this can be eliminated by keeping the fluid pressurized. Some resources do not require pumps to maintain the flow and pressure necessary to eliminate flashing (artesian flow). Similarly, if the exit temperature of the geofluid remains above some minimum value, silica will not be precipitated. These two operational strategies limit the scaling in a binary plant. Any constraint imposed on the geofluid exit temperature will impact the design of the

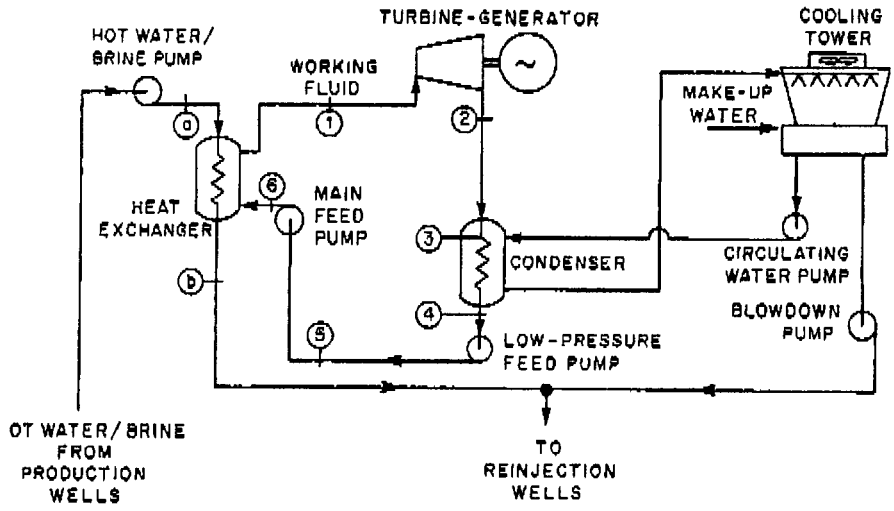


FIGURE 8.12.7 Schematic diagram of a binary plant. (From Kestin, J., Ed., Sourcebook on the Production of Electricity from Geothermal Energy, U.S. DOE, DOE/RA/4051-1, Washington, D.C., 1980).

binary plant, affecting the selection of turbine inlet conditions as well as the possible choice of working fluid.

The binary cycle consists of a closed loop of a working fluid normally performing a Rankine cycle. Most existing binary cycles use isobutane, pentane, or isopentane as working fluids. Studies have indicated that mixtures of hydrocarbons, e.g., 96% isobutane/4% hexane, will produce better utilization of a 180°C resource and in some instances lower LEC.

The performance of the binary system depends on a number of factors. Some plant designs incorporate multiple or staged boiling cycles, where the working fluid flow is split and boiling occurs at multiple pressures. In these cycles, multiple or staged turbines are required. The advantage of these cycles is the fact that the working fluid heat-addition process more closely matches the sensible cooling of the liquid geofluid (as shown on the T-Q or T-h diagram). This lowers temperature differences through the cycle, reducing cycle irreversibilities and increasing performance. The same effect can be achieved by heating the working fluid at supercritical pressures (pressures above the critical pressure). While the supercritical cycle will have higher component, material, and pumping costs because of higher operating pressures, they have fewer components because they are less complex than multiple boiling cycles. (In many cases, the maximum pressure can be kept below 600 psi with hydrocarbons such as isobutane.) Lowering the mean temperature difference in heat exchangers tends to require larger units so that capital costs are increased. In general, the LEC is reduced because the effects of increased performance more than outweigh the increase in capital cost.

The choice of the working fluid for the power cycle will also impact the cycle performance. In general, as resource temperature decreases, the more volatile fluids will produce more power per unit mass of geofluid. Power cycles using the more volatile fluids typically operate at higher pressures and have higher associated material and equipment cost. These higher costs may offset the gains in performance and produce higher LECs in some cases.

Working fluid mixtures have been shown to provide superior performance to the single component or pure working fluids. This performance improvement is due to the nonisothermal phase changes of this type of fluid at constant-pressure (both boiling and condensing), which allows the working fluid to match more closely the sensible cooling of the geofluid. More importantly in the reduction of irreversibilities, the desuperheating and condensing process more closely matches the sensible heating of cooling water or air in the heat-rejection process.

One additional type of binary cycle that has been proposed uses an ammonia-water mixture for the working fluid. A great deal of recuperative preheat of the working fluid is accomplished by splitting the duty of the geofluid, turbine exhaust, and preheated liquid flows through a more complex heat-transfer train than is shown in Figure 8.12.7. These systems are known as **Kalina systems**. In general, these systems do not change the composition of the mixture in the cycle as the Kalina cycle for applications such as is the case for gas turbine bottoming.

There is some consideration of using a binary cycle as a bottoming cycle for a flash steam or direct steam system. Similarly, a binary cycle could be used to bottom another binary system, perhaps with a different working fluid.

Design Considerations

The selection of the working fluid in binary cycles imposes safety considerations to be considered in the design of the power plant. Equipment and facility designs must take into account the flammable characteristic of the hydrocarbon working fluids.

The selection of materials of construction for the piping and components exposed to the geofluid will be resource specific. Typically, carbon steel is used for piping and pressure vessels. Turbines that use the steam directly may have stainless steel components, although the use of stainless may be limited by the presence of chlorides and the potential for stress cracking. The standard design for heat exchangers in binary cycles is for the geofluid to be on the tube side. This facilitates the cleaning of the exchanger if scaling or fouling occurs on the surfaces exposed to the geofluid. If the geofluid has a high scaling potential, components and piping should be designed to allow for periodic cleaning.

Direct Use of the Geothermal Resource

A number of direct-use applications of the heat in a geothermal resource have been successfully implemented. These include

1. Space conditioning (heating with the resource or a secondary fluid and cooling with heat pumps);
2. Heating of greenhouses;
3. Aquaculture;
4. Process heating (drying vegetable products);
5. Ground coupled heat pumps.

Although the United States is one of the world leaders for the production of electrical power from geothermal energy, other nations take the lead for the direct use of this energy source. In Iceland, over 85% of the buildings are supplied with heat and domestic hot water from geothermal systems (Ragnorson, 1995).

Typical direct-use applications are either closed systems with produced fluids being injected back into the geothermal reservoir or systems where the produced water is pure enough for beneficial use or disposal to surface waterways. Experience has shown that it is usually worthwhile to inject as much of the cooled fluid as possible back into the reservoir to maintain pressure and production rates.

Defining Terms

Binary system: A binary system that uses thermal energy from the geofluid to vaporize a secondary working fluid in a Rankine cycle.

Direct steam system: A geothermal energy conversion system that utilizes steam directly from a geothermal well.

Flashed steam system: A geothermal energy conversion system that utilizes steam flashed from the liquid geofluid.

Geopressurized resource: Naturally occurring reservoirs of hot pressurized fluid created by convection of water through hot porous rock.

Hydrothermal resource: Artificial reservoirs created by injecting water into hot dry rock in the earth's core.

Kalina system: A binary system using a mixture of ammonia and water as the working fluid in the power cycle.

Reference

Ragnorson, A. Iceland country update, in *Proceedings of the World Geothermal Congress, 1995*, Florence, Italy, May 1995, 145–161.

Further Information

Kestin, J. Ed., *Sourcebook on the Production of Electricity from Geothermal Energy,* U.S.DOE, DOE/RA/4051-1, Washington, D.C., 1980.

Lienau, Paul J. and Ben C. Lunis, Eds., *Geothermal Direct Use Engineering and Design Guidebook,* USDOE, Idaho Falls, ID, 1991.

Transactions of the Geothermal Resources Council, Vol. 1–19, (1977–1995), Geothermal Resources Council, Davis, CA.

8.13 Direct Energy Conversion

Solar Photovoltaic Cells

Kitt C. Reinhardt

Introduction

Solar photovoltaic cells convert sunlight directly into electrical energy via the collection of solar photon-generated semiconductor charge carriers. The collection of charge carriers within the cell produces a voltage across the terminals of the cell, called the **photovoltaic effect**, that can drive an external electrical circuit or charge a storage battery. Photovoltaic cells are useful in both space and terrestrial power applications. Silicon, Si, photovoltaic cells have provided the main source of electrical power to virtually all Earth-bound satellites since the advent of the space program in the late 1950s. In the early 1970s, photovoltaics generated a significant amount of interest for use in terrestrial power systems when oil supplies to the industrial world were disrupted. Today, while photovoltaic power remains the primary energy source for most communication and surveillance satellites, issues concerning system efficiency, reliability, and cost currently prevent its widespread use in residential and power utility applications. For example, in the United States the average price for conventional utility electricity is 6¢/kWhr, compared with ~35¢/kWhr for terrestrial photovoltaic electricity (Zweibel, 1995). Thus, the cost of photovoltaic power must be reduced by a factor of ~6 for it to become economically viable. At present, photovoltaic power is generally only cost-competitive for use in remotely located systems where conventional power is cost-prohibitive, such as in remote water-pumping and communications stations, signal and emergency lighting, and for village power. Factors that influence photovoltaic system energy costs include cell panel efficiency, total system lifetime, and cost per unit area. The present discussion will focus on issues concerning photovoltaic cells and panels. Detailed literature on power conditioning electronics and energy storage systems can be found elsewhere. A large number of different photovoltaic cell designs have been demonstrated by researchers over the years. However, the most common and practical cell designs are fabricated using single-crystal Si. Consequently, Si will be used to describe basic principles of semiconductors and photovoltaic cell operation.

Introduction to Semiconductors

We begin with a description of the concept of covalent bonding, valence electrons, and energy bands, which relates to conduction in semiconductors (Sze, 1981). The crystalline structure of Si is diamond, where each Si atom in the lattice is covalently bonded to four equidistant nearest neighbors that lie at the corners of a tetrahedron. Each Si atom has four electrons in its outer orbit, called valence electrons, and each atom shares these electrons with its four neighbors to form four covalent bonds. The atomic configuration of the 14 electrons of Si is $1s^2 2s^2 2p^6 3s^2 3p^2$. At practical temperatures, only the $3s^2 3p^2$ valence electrons contribute to the electrical conductivity; the $1s^2 2s^2 2p^6$ core electrons are too tightly bounded to the nucleus to participate. In a simplified model, as N Si atoms are brought together at 0 K to form a crystal, two distinct and nearly continuous bands of electronic energy levels form that are separated by an energy gap called the semiconductor **band gap, E_g**. The resulting upper **conduction band** contains 4N states, as does the lower **valence band**. The 4N electrons that come from the Si $3s^2 3p^2$ states completely fill the 4N states in the valence band at 0 K, and the conduction band states are completely empty. Since there are no unoccupied states in the valence band for electrons to move and the conduction band is empty, Si is a perfect insulator at 0 K.

As the temperature of the crystal increases, electrons in the valence band gain sufficient thermal energy ($>E_g$) to be excited across the band gap into the conduction band, leaving holes (missing electrons) behind in the valence band. When current conduction in a semiconductor is dominated by thermally generated electrons and holes, it is called intrinsic. In this case, the resulting number of electrons per unit volume in the conduction band, n, equals the number of holes per volume in the valence band, p, that is $n = p = n_i$, where n_i is called the intrinsic carrier concentration. In the presence of an electric

field, intrinsic electrons and holes gain kinetic energy and conduct electricity. However, since at room temperature n_i for Si is only 1.45×10^{10} cm^{-3}, compared with a free-electron density of more than 10^{22} cm^{-3} in metals, Si behaves as a very good insulator, i.e., electrical conductivity, σ, is given by $\sigma = q(n\mu_n + p\mu_p)$, where q is the electronic charge and μ is the respective carrier mobility.

In order to increase the conductivity to values useful for solid-state devices, the level of n and p can be increased by purposely adding impurity atoms into the crystal, called doping, that liberate extra electrons or holes. In the case of Si, which is in column IV of the periodic table, and hence has four valence electrons for bonding, doping is achieved using either column III elements (boron, aluminum, gallium, or indium), which have three valence electrons, or column V elements (phosphorus, arsenic, or antimony), which have five valence electrons. When an arsenic atom with five valence electrons replaces (substitutes) an Si atom, four of its electrons are used to form covalent bonds with the four neighboring Si atoms. The fifth electron is loosely bound to the arsenic nucleus, and at room temperature is ionized and "donated" to the conduction band. Arsenic is therefore called a donor, and Si becomes an n-type (mostly electrons) semiconductor. Similarly, when a boron atom with three valence electrons substitutes for an Si atom, one of the boron four covalent bonds becomes deficient of one electron. Boron can then accept one electron from the valence band to satisfy the bond requirement, which creates a positively charged hole in the valence band. Boron is therefore called an acceptor, and Si becomes a p-type (mostly holes) semiconductor. In this way the electrical conductivity of semiconductors can be precisely controlled by varying the concentration of donor and acceptor impurities. In practical solid-state devices, typical values of n and p range between 10^{15} and 10^{19} cm^{-3}.

The p-n Junction Diode

The p-n junction is a basic structure used for solid-state device rectification, amplification, and switching, as well as for photocarrier collection in photovoltaic cells. A p-n junction is formed when a p-type semiconductor is metallurgically joined with an n-type semiconductor (Streetman, 1980). Before they are joined, the p-material has a large concentration of holes and very few electrons, whereas the converse is true for the n-material. Upon joining the two materials, holes instantaneously diffuse from the p-side into the n-side and electrons diffuse from the n-side into the p-side. The transport of these carriers constitutes a "diffusion" current from the p-side to n-side; electron current is opposite in direction to electron flow by convention. As shown in Figure 8.13.1, negative acceptor ions are left behind as holes leave the p-side of the junction, creating a negative space-charge region (SCR), and positive donor ions are left behind as electrons leave the n-side of the junction, creating a positive SCR. Consequently, an electric field directed from the positive SCR to the negative SCR results that opposes the further diffusion of electrons and holes; i.e., the electric field creates a drift component of current from the n-side to p-side that opposes the diffusion component. In the absence of any external fields a condition of equilibrium is established, and the net current flow across the junction is zero. As will be discussed, the p-n junction electric field is also responsible for separating and collecting photon-generated carriers in photovoltaic cells.

When a voltage is applied across a p-n junction, the balance between the electron and hole drift and diffusion currents is disturbed and a net current results. Under forward bias, a positive voltage is applied to the p-side relative to the n-side, and the electric field across the junction is reduced; i.e., the electric field associated with the applied voltage subtracts from the zero-bias field. The reduced field enhances hole diffusion from the p-side to the n-side and electron diffusion from the n-side to the p-side, thereby increasing the "positive" current; the transport of current from the p-side to the n-side is positive by convention. Conversely, under reverse bias a negative voltage is applied to the p-side relative to the n-side, and the electric field across the junction increases. Consequently, the diffusion component of current decreases relative to the drift component, and a net "negative" current results across the junction.

The dark current-voltage (I-V) characteristics for an Si p-n junction are generally well described by the ideal Shockley diode equation (Sze, 1981),

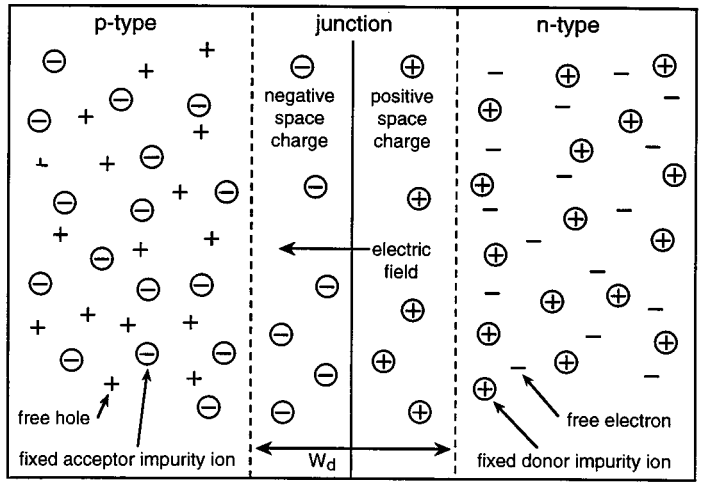


FIGURE 8.13.1 Schematic diagram illustrating an abrupt p-n junction with a unifiorm concentration of donor impurities in the n-region and acceptor impurities in the p-region.

$$I_D = I_o\left[\exp\left(qV/nkT\right) - 1\right] \tag{8.13.1}$$

where I_D is the junction dark current, I_o is the reverse saturation current, V is the forward-bias voltage, n is the diode ideality factor, and T is the absolute temperature. The value of n is ~1.0 when the current is dominated by carrier diffusion, but increases and approaches values of ~2 or greater when other current mechanisms become important, such as carrier recombination or tunneling. In high-quality Si p-n junction photovoltaic cells, the value of n is ~1.0 near the relevant operating voltage. The parameter I_o varies with T and E_g according to

$$I_o = qA\left[D_n n_p/L_n + D_p p_n/L_p\right] \propto T^3 \exp\left(-E_g/kT\right) \tag{8.13.2}$$

where A is the junction area and D_n and D_p, n_p and p_n, and L_n and L_p are the diffusion coefficients, minority carrier densities, and diffusion lengths for electrons and holes, respectively. The value of I_o decreases strongly as E_g increases, which, as will be shown, increases the photovoltage obtainable from a photovoltaic cell.

Cell Operation and Efficiency

Cell Operation. Photovoltaic energy conversion in a p-n junction is a two-step process where free electrons and holes (photocarriers) are generated in the semiconductor via the absorption of solar energy and then simultaneously collected across the junction (Fahrenbruch and Bube, 1983). Consider the schematic of a typical photovoltaic cell shown in Figure 8.13.2 which consists of a p-n junction formed very close to the top surface of the cell. Front metal ohmic contact grid fingers allow solar energy to pass into the absorber layers. The entire top surface is covered with an antireflection coating to minimize reflective losses, and the entire back surface is covered with an ohmic contact. The ohmic contacts form n and p region terminals that transfer (conduct) current from the semiconductor to the external circuit with a negligible amount of voltage drop.

When the photovoltaic cell is exposed to solar radiation, photons with energies greater than E_g (super-band-gap photons) are absorbed in the n and p layers, and free electrons and holes are generated via the breaking of covalent bonds. These electron-hole pairs, or **photocarriers**, are shown in Figure 8.13.1. The energy of the free photocarriers is converted directly into a current and voltage via photocarrier collection by the junction. The absorbed photons effectively contribute an energy E_g to the cell output,

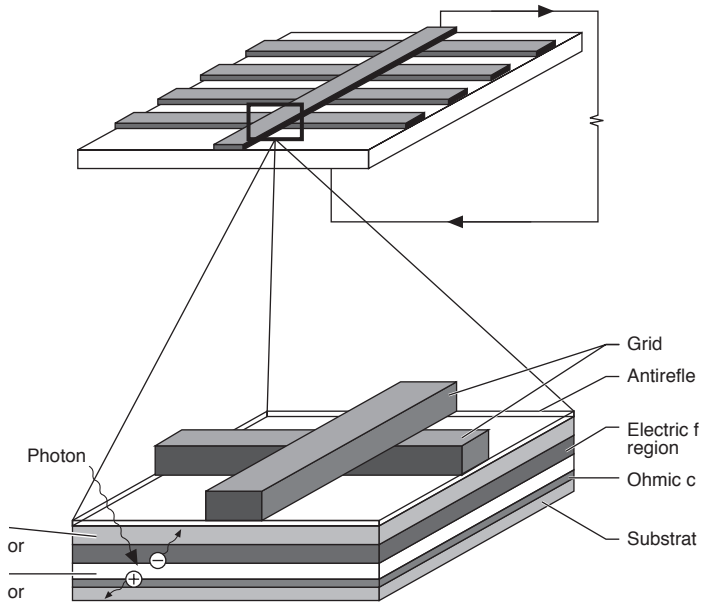


FIGURE 8.13.2 Schematic diagram of a typical p-n junction photovoltaic cell.

and energy greater than E_g is lost as heat. Photons with energies less than E_g (sub-band-gap photons) are transmitted through the cell. After generation, minority photocarriers, that is, holes on the n-side and electrons on the p-side, diffuse toward the edges of the junction due to a gradient of carriers that exists there. If the minority carriers are generated within a diffusion length, L, of the junction, they will reach it and be swept across it by the electric field of the junction. Hence, electrons are swept from the p-side to the n-side and holes from the n-side to the p-side, and thus they are separated. The minority carrier gradient present at the edges of the junction is due to the depletion of minority carriers that results from their transfer across the junction. The diffusion flux of minority carriers toward and across the junction constitutes a light-generated current, I_L, or photocurrent, that is directed from the n-side to the p-side of the cell. The build-up of positive holes on the p-side and negative electrons on the n-side gives rise to a **photovoltage** across the junction. The polarity of both the photovoltage and photocurrent is identical to that of a battery, and power is delivered from the junction to the external circuit.

Cell Efficiency. In order to derive the solar conversion efficiency, it is convenient to model the photovoltaic cell as an ideal p-n diode in parallel with a **light-generated** (constant) **current** source, I_L, as shown in the equivalent circuit of Figure 8.13.3. Parasitic series and shunt resistance losses, R_s and R_{sh}, respectively, are also shown, where R_s is due to ohmic contact and semiconductor resistances, and R_{sh} is due to defect-related carrier recombination and/or tunneling phenomena (Stirn, 1972). A qualitative expression for I_L is given by Tada et al. (1982):

$$I_L = qAG\left(L_n + W_d + L_p\right) \tag{8.13.3}$$

where G is the photocarrier generation rate in carriers/cm^3-sec due to solar photon absorption, which depends on E_g and the photon energy (wavelength) and intensity (concentration), and W_d is the sum of the negative and positive SCR widths. As mentioned, I_L is directed from the n-side to the p-side. In

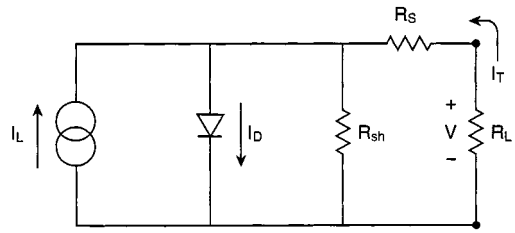


FIGURE 8.13.3 Schematic of equivalent circuit model for a p-n photovoltaic cell.

contrast, the dark diode current given by Equation (8.13.1) is directed oppositely from the p-side to the n-side. The **dark current** is due to the forward-bias photovoltage that appears across the cell p-n junction when it is illuminated. Thus, the dark current opposes the light current. In the ideal case, that is, when $R_s = 0$ and $R_{sh} = \infty$, the total forward current, I_T, is given by

$$I_T = I_D - I_L = I_o\left[\exp(qV/nkT) - 1\right] - I_L \tag{8.13.4}$$

A plot of dark and light current-voltage (I-V) curves resulting from Equations (8.13.1) and (8.13.4), respectively, is shown in Figure 8.13.4 for a typical p-n solar cell. Under illumination, the forward-bias dark I-V curve is displaced downward into the fourth quadrant by the **photocurrent**, $\boldsymbol{I_L}$. It is noted from Figure 8.13.3 that the voltage drop across the load resistance, R_L, is $V = -I_T R_L$. Under short-circuit conditions, that is, when the n and p terminals are tied to each other, R_L is negligible. The resulting voltage drop across the p-n junction will also be negligible, and from Equation (8.13.1), $I_D \approx 0$. As shown in Figure 8.13.4, the resultant current is termed the **short-circuit current**, $\boldsymbol{I_{sc}}$, or $I_T = -I_{sc} = -I_L$. As the value of R_L increases, a voltage appears across the junction, $V = -I_T R_L$, called the **photovoltage**, and I_D increases in accordance with Equation 8.13.1. Under this condition the cell is operating in the fourth quadrant of the I-V characteristic (i.e., the junction voltage is positive and the current is negative), and, consequently, the cell delivers power (product of the current and voltage) to R_L. As the value of R_L continues to increase, so too does V and I_D. When the value of R_L approaches infinity, that is, under open-circuit conditions, I_D approaches I_L and I_T goes to zero. The resulting **open-circuit voltage**, $\boldsymbol{V_{oc}}$, is shown in Figure 8.13.4. V_{oc} can be obtained by setting $I_T = 0$ in Equation (8.13.4) and solving for V. For $V \gg kT/q$,

$$V_{oc} = nkT/q \ln\left[I_L/I_o\right] \tag{8.13.5}$$

Thus, the operating point on the I-V curve in the fourth quadrant can be swept from (I_{sc}, 0) to (0, V_{oc}) by varying the value of R_L. When the optimum load is chosen (i.e., $R_L \approx V_{oc}/I_{sc}$), approximately 80% of the product $I_{sc}V_{oc}$ can be extracted as useful power as shown by the shaded maximum-power rectangle in Figure 8.13.4. Also shown in Figure 8.13.4 are the parameters I_m and V_m, which correspond to values of current and voltage, respectively, that yield the maximum cell power, P_m, where $P_m = I_m V_m$. The knee that appears at P_m is due to the parasitic effects of R_s and R_{sh}. The curvature of the knee at P_m is described by the **fill factor**, **FF**, where

$$\text{FF} = I_m V_m / I_{sc} V_{oc} \tag{8.13.6}$$

The photovoltaic cell **conversion efficiency**, η, is defined as

$$\eta = I_m V_m / P_{in} A$$

or

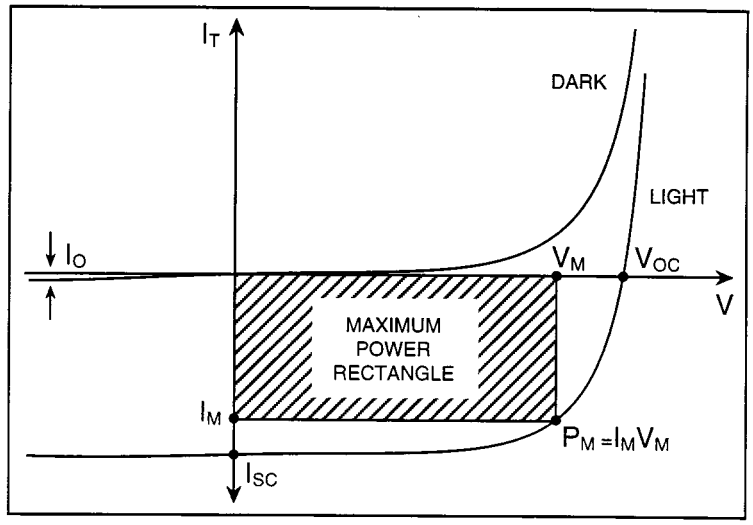


FIGURE 8.13.4 Typical dark and light current-voltage characteristics for a p-n photovoltaic cell.

$$\eta = FF\, I_{sc} V_{oc} / P_{\text{in}} \tag{8.13.7}$$

where P_{in} is the incident power in W/m^2 equal to the sum in energy of the incident photons per time per area. Values for V_{oc}, I_{sc}, FF, and η can be obtained in the laboratory under various air mass conditions from light *I-V* curves measured using a carefully controlled (calibrated) light source to illuminate the cell.

Cell Material vs. Efficiency

The optimum value of material E_g for solar photovoltaic conversion is ~1.0 to 1.5 eV. To understand how the choice of cell material affects conversion efficiency, an ideal expression can be derived using Equation 8.13.4 for the theoretical conversion efficiency. The output power can be expressed as (Henry, 1980)

$$P = IV = I_o V\left[\exp\left(qV/nkT\right) - 1\right] - I_L V \tag{8.13.8}$$

The maximum output power is obtained when $dP/dV = 0$, and an expression for I_m and V_m can be obtained from Equation (8.13.8) and multiplied to give P_m, where

$$P_m = I_m V_m \approx I_L\left[V_{oc} - kT/q \ln\left(qV_m/kT + 1\right) - kT/q\right] \tag{8.13.9}$$

In practical cells, values for V_m and V_{oc} are typically $^1/_2E_g/q$ to $^2/_3E_g/q$. Thus, for materials with $E_g \sim 1 - 2$ eV, the quantity in the large brackets of Equation (8.13.9) becomes ~V_{oc}, and the factors that determine I_L and V_{oc} also determine P_m. From Equations (8.13.2) and (8.13.5), it is clear that V_{oc} increases with E_g through the reduction in I_o. In contrast, as E_g increases, I_L decreases because a smaller portion of the solar spectrum is energetic enough to be absorbed; i.e., I_L is the product of q and the number of available photons with energy greater than E_g. Hence, for a given solar spectrum there is an optimum value of E_g that maximizes the product of V_{oc} and I_L. A plot of ideal AM1 conversion efficiency vs. E_g is shown in Figure 8.13.5 for "one sun" (925 W/m^2) and "1000 suns" (925 kW/m^2) concentrations (Henry, 1980). The efficiency curves were obtained using Equations (8.13.1) through (8.13.7) at 300 K. A maximum in efficiency occurs for $E_g \sim 1.0 - 1.5$ eV.

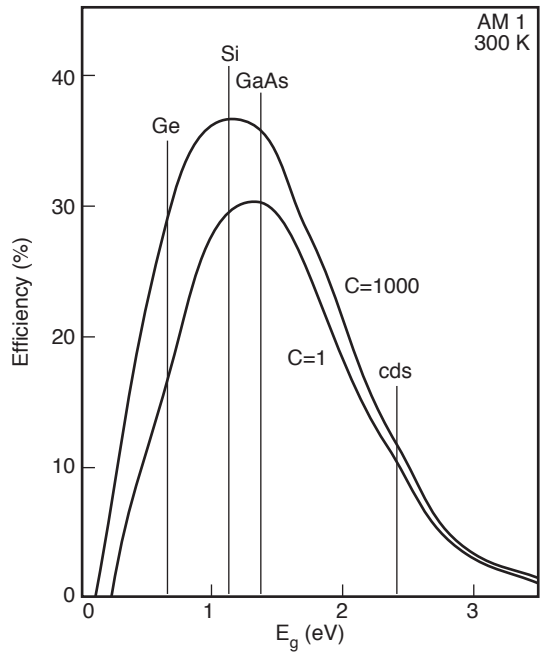


FIGURE 8.13.5 Theoretical AM1 efficiency vs. semiconductor band gap E_g for an ideal photovoltaic cell under 1 sun and 1000 suns concentrations.

Manufacture of Cells and Panels

There are basically five important solar cell design concepts, and each offers a trade-off between efficiency and cost: (1) Large-area single-crystal planar cells, typically 1 × 1 cm² to 6 × 6 cm², yield high efficiencies under normal light conditions, 2) single-crystal small-area concentrator cells, typically less than 1 × 1 cm², are potentially less costly and yield higher efficiencies under concentrated light, i.e., concentration ratios of 20 to 1000 are typical; (3) more-complex single-crystal multijunction cells yield the highest efficiencies measured to date, but are substantially more expensive; (4) cells made from polycrystalline materials are less expensive than single-crystalline cells, but are less efficient; and (5) cells made from thin film amorphous materials provide the lowest-cost approach yet, but are generally less efficient than polycrystalline cells.

A typical 15% 4 × 4 cm² photovoltaic cell produces only ~0.25 W under AM1.5 conditions. Therefore, individual cells must be electrically wired together to form larger submodules or panels to increase the total output power. The cells can be connected in series to increase the total voltage or in parallel to increase the current. The modular nature of photovoltaic power allows the design of systems that can deliver electrical power from a few watts to many megawatts. In terrestrial applications the cells are typically supported and held in place with a rigid substrate, i.e., typically aluminum, Plexiglas, fiberglass, or glass, and are encapsulated with glass or a polymeric material; in space applications the support structure may be rigid or flexible, and the cells are protected from the space environment with quartz cover slides. The electrical power generated by the cells is conducted to an electrical load or storage battery. Metal interconnects soldered to the ohmic contacts of the cells conduct electrical current from one cell to the next. Current is then conducted from the network of series- and parallel-connected cells by wires to a distribution terminal "bus" that transfers the power to either the load or battery.

Single-Crystal Cells. The p-n photovoltaic cells made from **single-crystal** Si dominate in space and terrestrial applications because of their high efficiency and reliability. The formation of single-crystal p-

type Si results from the selective cooling of pure molten Si to form large cylindrical crystal ingots, called boules, from which thin wafers are sliced and polished. The p-type impurities, usually boron, are added to the melt, to give the desired impurity concentration. A large-area p-n junction is then formed by diffusion of n-type impurity atoms, usually phosphorus. Front and back metal ohmic contacts and an antireflection coating are then formed using standard photolithography thermal evaporation or sputtering techniques (Sze, 1985). The resulting cell structure is shown in Figure 8.13.2, and typical cell areas range from 1 to 36 cm^2. Different semiconductors absorb sunlight more efficiently than others, described by a factor called the absorption coefficient. Si has a relatively small absorption coefficient compared with other materials, such as InP, GaAs, and amorphous-Si, and consequently requires an absorption layer thickness of ~100 μm to maximize conversion efficiency. Conventional Si cells have a thickness on the order of 250 μm, but can be chemically or mechanically polished to a thickness of 100 μm.

Single-crystal III-V photovoltaic cells, such as InP and GaAs, are made from elements in the III and V columns of the periodic table. The band gaps of these cells, 1.35 and 1.42 eV, respectively, are close to the optimum value. These materials involve the growth of single-crystal semiconductor layers upon a single-crystal semiconductor substrate. This technique is called epitaxy and it provides a method to produce both n-type and p-type layers to form the p-n junction. Epitaxial growth of n and p layers is required for InP and GaAs because diffusion of impurities at high temperatures is confounded by the high vapor pressure of the material. The formation of ohmic contacts and antireflection coating employ the same techniques as used for Si cells. The required absorption layer thickness for these cells is only a few microns because of their large absorption coefficients. However, issues concerning yield and mechanical strength limit their minimum thickness to ~100 μm The best reported efficiencies for single-crystal Si, GaAs, and InP cells under AM1.5 conditions are 24, 25, and 22%, respectively (Green et al., 1995).

Polycrystalline Cells. In the case of **polycrystalline** Si cells, molten Si is directly deposited into either cylindrical or rectangular ingots. As the material solidifies, individual crystalline regions form that are separated by grain boundaries which contain large numbers of structural defects. When the cell is illuminated, these defects capture a portion of the light-generated electron-hole pairs through recombination processes before they can reach the junction and be collected. Thus, the grain boundaries diminish the light-generated current and overall efficiency of the cell. However, polycrystalline silicon cells are sufficiently inexpensive to be commercially viable (Stone, 1993). An area that requires improvement is the slicing of polycrystalline ingots, where yields as low as 50% are common. An approach that eliminates the expense of sawing and polishing altogether is the growth of polycrystalline Si directly into the form of thin ribbons using a technique called edge-defined film-fed growth (EFG) (Fahrenbruch and Bube, 1983). In this approach, a carbon die with a slot-shaped aperture is immersed in a crucible of molten Si. The liquid Si wets the die and flows through the slot where it cools and is pulled to form a thin ribbon. This material also has high crystalline defect densities, but has good overall yields. An additional approach involves the growth of films of nearly single-crystal quality, where two parallel supporting dendrites form the boundaries of a web or ribbon pulled from a supercooled melt of Si. The best efficiency for polycrystalline Si cells under AM1.5 conditions is ~18%; that for Si cells grown by the EFG technique is 14%; and that for Si dendritic web cells is 15.5% (Stone, 1993). It is important to note that although these lower-cost films yield lower cell efficiencies compared with single-crystal cells, the cost of cells depends on the cost of the starting material and the cost per watt is more important than efficiency (Zweibel, 1995).

Thin Film Cells. Thin films cells provide an even lower-cost (and lower-efficiency) approach because they require a very small amount of semiconductor. An excellent review on thin film photovoltaic technologies, particularity on present and future cost issues, is given by Zweibel (1995). The general approach involves depositing only a few microns of material on a low-cost substrate using various vacuum deposition techniques, although a multitude of other deposition techniques have also been demonstrated. The top thin film cell candidates are amorphous Si, a-Si, cadmium telluride, CdTe, and copper indium diselenide, CIS (and related alloys). The highest reported thin film AM1.5 efficiencies

are 17% for CIS, followed by 15.8% for CdTe, and ~11% for a-Si (Zweibel, 1995). However, the relative level of maturity of each design for commercial application must be put into perspective. While the best CIS cell efficiency is quite high, the best CIS square foot panel efficiency reported back in 1988 was, and still is today, only 11%. Significant manufacturing problems have plagued the CIS cell, and currently it is still not commercially available. CdTe is believed to be the easiest of the thin film cells to fabricate, and probably represents the closest to large-scale commercialization. Two U.S. companies have publicly announced CdTe manufacturing plants, and commercial efficiencies are likely to be in the range of 6 to 8% in the first plants. The future of a-Si cells is currently believed to be limited if it cannot overcome a 10% efficiency at the module level. Development problems include electrochemical instability to light that results in a 20 to 40% degradation. However, it appears that the use of multijunction thin film a-Si layers may solve the problem, and modules of 7 to 9% are expected in the near term.

Concentrator Cells. Photovoltaic modules are typically either of the flat plate or concentrator configuration. Flat plate modules can be fixed with respect to the sun or mounted to track the sun in one or two axis. Concentrator modules use large-area mirrors or lenses to concentrate sunlight onto smaller-area cells. Concentrator cells operate at higher efficiencies. However, concentrator modules require one- or two-axis tracking which adds system complexity and cost that generally offsets the module efficiency and lower area cost benefits. The increase in conversion efficiency with illumination intensity is shown in the 1000-sun concentration curve of Figure 8.13.5. Values for I_L increase linearly with concentration through the factor G in Equation (8.13.3) and V_{oc} increases logarithmically with concentration through I_L in Equation (8.13.5). Under solar concentration of "20 suns" or greater, a significant amount of cell heating can occur. While J_{sc} increases slightly with increasing temperature, values of V_{oc} and FF drop strongly. Thus, on adequate heat sink or active cooling is required at high concentrations. The reported AM1.5 efficiency for Si cells increases from 24 to 26.5% under a concentration of 255; and that for GaAs cells increases from 25 to 27.6% under a concentration of 140 (Green et al., 1995).

Multijunction Cells. Another approach to increase photovoltaic cell efficiency is through the use of multijunction tandem cells. The simplest multijunction cell is a two-junction, two-terminal device. In this design, a high-band-gap (E_{g1}) p-n junction is vertically stacked (mechanically) or epitaxially grown atop a bottom lower-band-gap (E_{g2}) p-n junction as shown in Figure 8.13.6. The top junction absorbs photons with energy $\geq E_{g1}$, and the bottom junction absorbs photons with energy E, where $E_{g1} > E \geq E_{g2}$, that passed through the top cell. This increases the utilization of the solar spectrum, since the excess energy of the high-energy photons is not wasted. The values of E_{g1} and E_{g2} must be chosen to achieve maximum solar absorption and current matching; i.e., the two cells must generate equal currents when illuminated. The band gap combinations of ~0.7 and 1.4, 1.0 and 1.4, 1.1 and 1.4, and 1.4 and 1.9 eV are current matched, where E_g for GaSb, $CuInSe_2$, Si, GaAs, AlGaAs, and $GaInP_2$ are ~0.7, 1.0, 1.1, 1.4, and 1.9 eV, respectively. Multijunction concentrator cells have reported efficiencies in excess of 30% (Green, 1995).

Design of a Photovoltaic Generating System

A schematic diagram depicting the basic components of a typical photovoltaic power generation system is shown in Figure 8.13.7 (Pulfrey, 1978). The system includes a photovoltaic array that consists of many smaller submodules, each containing many hundreds or thousands of photovoltaic cells. The DC output power from the array is controlled by a power conditioning unit that contains an inverter for developing AC power and an input power tracking device to maintain the optimal array load to achieve maximum output power. Power is directly fed to the electrical load and/or storage system by the conditioning unit. The storage system is needed to save energy when power is generated in excess of the immediate demand, or when the load demand exceeds the immediate generation level. Total photovoltaic power system efficiency is the product of the efficiencies of the individual components. Typical efficiencies for the power conditioning unit (determined by the inverter) and energy storage system are about 95% and 50-80%, respectively. Thus, an array efficiency of 10% would result in a total system

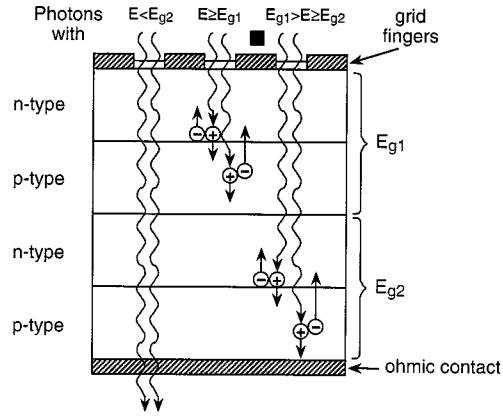


FIGURE 8.13.6 Schematic diagram of a multijunction (two-junction) photovoltaic cell under illumination.

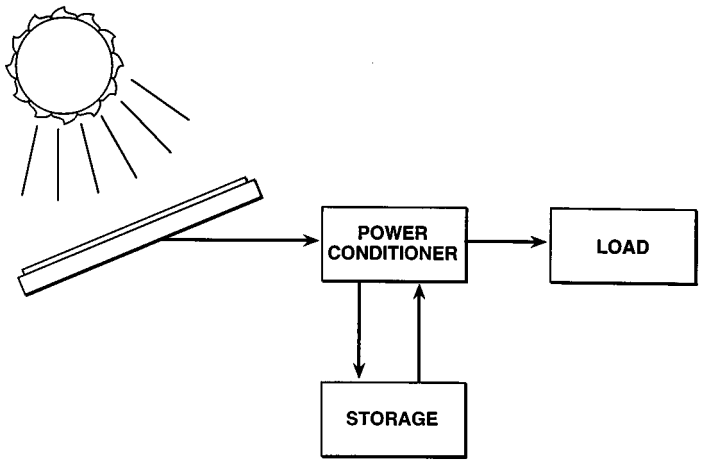


FIGURE 8.13.7 Schematic diagram depicting the basic components of a typical photovoltaic power system.

efficiency of about 5 to 8%, and a total system efficiency of 10% would require an array efficiency in excess of 13%.

Defining Terms

Band gap (E_g): The difference in energy between the energy level of the bottom of the conduction band and the energy level of the top of the valence band.

Conduction band: A range of allowable energy states in a solid in which electrons can move freely.

Conversion efficiency (η): The ratio of the available power output photovoltaic cell to the total incident radiant power.

Dark current: Any current that flows through the p-n junction in the absence of external irradiation.

Fill factor (ff): The ratio of the maximum photovoltaic cell output power to the product of the open-circuit voltage and short-circuit current.

Light-generated current (I_L): The electrical current obtained from an illuminated p-n junction resulting from the collection of photocarriers across the junction.

Open-circuit voltage (V_{oc}): The voltage obtained across the terminals of an illuminated p-n photovoltaic cell under open-circuit conditions.

Photocarriers: Electrons and holes generated within a semiconductor via the absorption of photon energy.

Photocurrent: Synonymous with light-generated current.

Photovoltaic effect: The production of a voltage difference across a p-n junction resulting from the absorption of photon energy.

Photovoltage: The voltage resulting from the photovoltaic effect.

Polycrystalline: A material characterized by an array or agglomerate of small single-crystal sections of various crystal orientations separated from one another by grain boundaries, which are localized regions of very severe lattice disruptions and dislocations.

Short-circuit current (I_{sc}): The electrical current measured through the terminals of an illuminated p-n photovoltaic cell under short-circuit conditions.

Single crystal: A material characterized by a perfect periodicity of atomic structure; the basic arrangement of atoms is repeated throughout the entire solid.

Valence band: A range of allowable energy states in a solid crystal in which lie the energies of the valence electrons that bind the crystal together.

References

Fahrenbruch, A.L. and Bube, R.H. 1983. *Fundamentals of Solar Cells — Photovoltaic Solar Energy Conversion,* Academic Press, New York.

Green, M.A., Emery, K., Bucher, K., and King, D.L. 1995. *Short communication: solar cell efficiency tables (version 5), Prog. Photovoltaics Res. Dev.*, 3, 51–55.

Henry, C.H. 1980. Limiting efficiency of ideal single and multiple energy gap terrestrial solar cells, *J. Appl. Phys.,* 51, 4494.

Pulfrey, D.L. 1978. *Photovoltaic Power Generation,* Van Nostrand Reinhold, New York.

Stirn, R.J. 1972. *Junction characteristics of Si solar cells,* in Proceedings of the 9th IEEE Photovoltaics Specialists Conference, p.72.

Stone, J.L. 1993. Photovoltaics: unlimited electrical power from the sun, *Phys. Today*, Sepember, 22–29.

Streetman, B.G. 1980. *Solid State Electronic Devices,* Prentice-Hall, Englewood Cliffs, NJ.

Sze, S.M. 1981. *Physics of Semiconductor Devices,* 2nd ed., John Wiley & Sons, New York.

Sze, S.M. 1985. *Semiconductor Devices: Physics and Technology,* John Wiley & Sons, New York, 341–457.

Tada, H.Y., Carter, J.R., Anspaugh, B.E., and Downing, R.G. 1982. *Solar Radiation Handbook*. JPL Publication 82-69, 2-11

Zweibel, K. 1995. *Thin Films: Past, Present, Future*. NREL/IP-413-7486 Publication (DOE UC Category 1260 DE95004084).

Further Information

An excellent presentation of the basic theory of the various photovoltaic cell designs is given in *Fundamentals of Solar Cells: Photovoltaic Solar Energy Conversion*, by Alan L. Fahrenbruch and Richard H. Bube. This text covers the basics of solar insolation, semiconductors, p-n junctions, and single-crystal, polycrystalline, thin film, and concentrator photovoltaic cells.

An excellent review of the progress achieved in terrestrial and space photovoltaics can be traced in the *Proceedings of the IEEE (Institute of Electrical and Electronics Engineers) Photovoltaics Specialists Conference* (PVSC) that dates back to 1961. These proceedings include thousands of papers that address nearly every aspect of photovoltaic cell development: basic theory, design, fabrication, and application.

The monthly journal *Solar Energy Materials and Solar Cells* covers many aspects of improving device efficiency, reducing costs, and testing and applications.

The monthly journal *Progress in Photovoltaics* documents recent results of research work conducted in photovoltaics worldwide. This journal is an excellent source for currently reported cell conversion efficiencies.

Proceedings of the conference *Space Photovoltaics Research and Technology* is an excellent source for the reader interested in the development of photovoltaics for use in space.

Fuel Cells

D. Yogi Goswami

Introduction

A *fuel cell* is an electrochemical device in which a fuel and an oxidant react in such a controlled manner that the chemical energy of reaction is converted directly into electrical energy. Ordinarily, a fuel reacts violently with an oxidant in a combustion reaction resulting in the release of heat of combustion. The heat of combustion can, then, be converted into electrical energy via mechanical work with the constraint of the second law of thermodynamics. The overall efficiency of the series of conversion processes is of the order of 40%. A fuel cell bypasses these processes resulting in potential efficiencies of the order of 80%. As an example, when hydrogen is burned in an atmosphere of oxygen it results in the following reaction:

$$H_2 + \frac{1}{2}O_2 \rightarrow H_2O + \Delta H \quad (\text{heat of reaction}) \tag{8.13.10}$$

In this reaction, two hydrogen atoms bond with an oxygen atom by sharing their electrons with the outermost orbit of oxygen, which becomes full, resulting in a stable structure.

The reactants H_2 and O_2 may be combined to form the same product (H_2O) by first stripping the electrons away from the hydrogen atoms and allowing the electrons to pass through an external circuit before combining with oxygen. Table 8.13.1 gives typical electrochemical reactions in fuel cells.

$$H_2 \rightarrow 2H^+ + 2e^- \quad (\text{electrons}) \tag{8.13.11}$$

$$2H^+ + O + 2e^- \rightarrow H_2O \tag{8.13.12}$$

Figure 8.13.8 shows a schematic of an arrangement that would allow the above reaction to proceed.

TABLE 8.13.1 Electrochemical Reactions in Fuel Cells

Fuel Cell	Anode Reaction	Cathode Reaction	Overall Reaction
Proton exchange	$H_2 \rightarrow 2H^+ + 2e^-$	$^1/_2O_2 + 2H^+ + 2e^- \rightarrow H_2O$	$H_2 + {}^1/_2O_2 \rightarrow H_2O$
Alkaline	$H_2 + 2(OH)^- \rightarrow 2H_2O + 2e^-$	$^1/_2O_2 + H_2O + 2e^- \rightarrow 2(OH)^-$	$H_2 + {}^1/_2O_2 \rightarrow H_2O$
Phosphoric acid	$H_2 \rightarrow 2H^+ + 2e^-$	$^1/_2O_2 + 2H^+ + 2e^- \rightarrow H_2O$	$H_2 + {}^1/_2O_2 \rightarrow H_2O$
Molten carbonate	$H_2 + CO_3^- \rightarrow H_2O + CO_2 + 2e^-$ $CO + CO_3^- \rightarrow 2CO_2 + 2e^-$	$^1/_2O_2 + CO_2 + 2e^- \rightarrow CO_3^-$	$H_2 + {}^1/_2O_2 + CO_2$ (cathode) $\rightarrow H_2O + CO_2$ (anode)
Solid oxide	$H_2 + O^- \rightarrow H_2O + 2e^-$ $CO + O^- \rightarrow CO_2 + 2e^-$ $CH_4 + 4O^- \rightarrow 2H_2O + CO_2 + 8e^-$	$^1/_2O_2 + 2e^- \rightarrow O^-$	$H_2 + {}^1/_2O_2 \rightarrow H_2O$ $CO + {}^1/_2O_2 \rightarrow CO_2$ $CH_4 + 2O_2 \rightarrow H_2O + CO_2$

Source: Hirschenhofer, J.H. et al., *Fuel Cells, A Handbook*, rev. 3, Gilbert/Commonwealth, Morgantown, WV, 1994. With permission.

In order for the above reactions to occur according to the schematic of Figure 8.13.8:

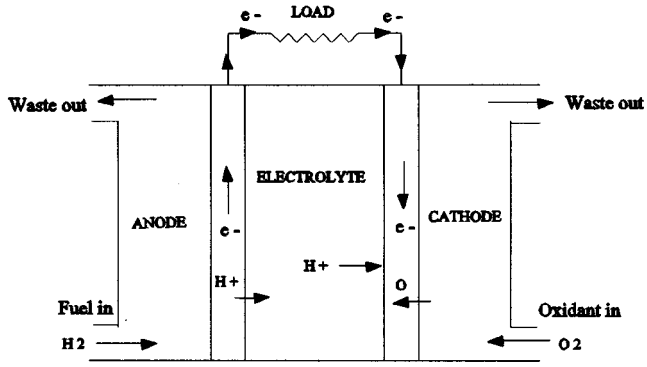


FIGURE 8.13.8 Conceptual schematic of a hydrogen fuel cell.

1. Electrodes must be porous to let the fuel and electrolyte penetrate.
2. The electrolyte must be permeable to H^+ and $(OH)^-$ ions.
3. Electrode materials must be catalysts for the reaction (Pt, Ni, etc.).

In 1839, William Grove, an English investigator, constructed a chemical battery in which he noticed that the water-forming reaction of hydrogen and oxygen generated an electrical current. However, it was not until 50 years later that two English chemists, Ludwig Mond and Carl Langer, developed a device they actually called a fuel cell (Angrist, 1982). There has been a strong resurgence in research and development of fuel cells in the last four decades.

Thermodynamic Performance

The energy released or needed in any chemical reaction (ΔH) is equal to the difference between the enthalpy of formation of the products and the reactants.

$$\Delta H = \sum (\Delta H)_{\text{products}} - \sum (\Delta H)_{\text{reactants}} \tag{8.13.13}$$

In an exothermic reaction the change in enthalpy of formation (ΔH) is negative. Table 8.13.2 gives values of ΔH for various compounds at 25°C at 1 atm. All naturally occurring elements have a ΔH value of zero.

TABLE 8.13.2 Enthalty of Formation ΔH^0 and Gibbs Free Energy ΔG^0 of Compounds at 1 atm and 298 K

Compound or Ion	Enthalpy of Formation, ΔH^0, J/kg · mol	Gibbs Free Energy ΔG^0, J/kg · mol
CO	-110.0×10^6	-137.5×10^6
CO_2	-394.0×10^6	-395.0×10^6
CH_4	-74.9×10^6	-50.8×10^6
Water	-286.0×10^6	-237.0×10^6
Steam	-241.0×10^6	-228.0×10^6
C, H_2, O_2	0	0
O(g)	$+249.2 \times 10^6$	$+231.8 \times 10^6$
H(g)	$+218.0 \times 10^6$	$+203.3 \times 10^6$

Source: Adapted from Wark, K., *Thermodynamics*, McGraw–Hill, New York, 1988. With permission.

In a combustion reaction all of the change in the enthalpy of formation (ΔH) is converted to heat and is, therefore, called the higher heating value (HHV).

$$-(\Delta H)_{\text{reaction}} = \text{HHV of fuel} \tag{8.13.14}$$

For example, for complete combustion of hydrogen according to the following reaction:

$$H_2 + \frac{1}{2}O_2 \rightarrow H_2O \quad (\text{liquid}) \tag{8.13.15}$$

Change in the enthalpy of formation is

$$\Delta H = \Delta H_{H_2O} - \Delta H_{H_2} - \frac{1}{2}\left(\Delta H_{O_2}\right) = \left(-286 \times 10^6\right) - 0 - 0 = -286 \times 10^6 \ \text{J/kg} \cdot \text{mol } H_2 \tag{8.13.16}$$

In a fuel cell, most of ΔH can be converted to electricity directly. The part that cannot be converted to work directly gets converted into heat. The minimum amount that must be converted to heat is represented by reversible heat transfer $\int T\,dS$. If a fuel cell operates isothermally, the maximum amount of electrical work (W_e) produced is given by

$$W_e, \max = \Delta H - T\Delta S \tag{8.13.17}$$

In an irreversible reaction

$$W_e, \max < \Delta H - T\Delta S \tag{8.13.18}$$

Gibbs free energy, G, is given by:

$$G = H - TS \tag{8.13.19}$$

Therefore, in a reversible isothermal process

$$\Delta G = \Delta H - T\Delta S = W_e, \max \tag{8.13.20}$$

The actual electrical work in a fuel cell is given by

$$W_e \leq \Delta G \quad (\text{change in Gibbs free energy for the reaction}) \tag{8.13.21}$$

$$\Delta G_{\text{reaction}} = \sum (\Delta G)_{\text{products}} - \sum (\Delta G)_{\text{reactants}} \tag{8.13.22}$$

The electrical work, W_e, is associated with the work of electrons passing through an external resistance. 1 g · mol of electrons is equal to Avogadro's number (6.022×10^{23}) and the charge of these electrons is equal to 96,439 C which is called a **faraday** (F). If n g · mols of electrons are generated and E is the internal reversible cell voltage, then the maximum electrical work is given by

$$W_e, \max = \Delta G = -nFE \tag{8.13.23}$$

denoting values under standard conditions (25°C, 1 atm) by superscript 0, we have

$$\Delta G^0 = -nFE^0 \tag{8.13.24}$$

Values of G^0 for various compounds are given in Table 8.13.2. For fuel cell reactions as below:

$$aA + bB \rightarrow cC + dD \tag{8.13.25}$$

if the reactants (A and B) and the products (C and D) are assumed to be ideal gases with partial pressures P_A, P_B, P_C, and P_D the change in Gibbs free energy ΔG and the internal reversible cell voltage, E, are given by

$$\Delta G = \Delta G^0 + RT \ln \frac{(P_C)^c (P_D)^d}{(P_A)^a (P_B)^b} \tag{8.13.26}$$

$$E = E^0 + RT \ln \frac{(P_A)^a (P_B)^b}{(P_C)^c (P_D)^d} \tag{8.13.27}$$

Equation 8.13.27 is also called the Nernst equation.

Table 8.13.3 gives Nernst equations for the electrochemical reactions listed in Table 8.13.1.

TABLE 8.13.3 Fuel Cell Reactions and the Corresponding Nernst Equations

Cell Reactions	Nernst Equation
$H_2 + {}^1/_2O_2 \rightarrow H_2O$	$E = E^0 + (RT/2F)\ln[P_{H_2}(P_{O_2})^{1/2}/P_{H_2O}]$
$H_2 + {}^1/_2O_2 + CO_2(c) \rightarrow H_2O + CO_2(a)$	$E = E^0 + (RT/2F)\ln[P_{H_2}(P_{O_2})^{1/2}(P_{CO_2})_c/(P_{H_2O}(P_{CO_2})_a]$
$CO + {}^1/_2O_2 \rightarrow CO_2$	$E = E^0 + (RT/2F)\ln[P_{CO}(P_{O_2})^{1/2}/P_{CO_2}]$
$CH_4 + 2O_2 \rightarrow H_2O + CO_2$	$E = E^0 + (RT/8F)\ln[P_{CH_4}(P_{O_2})^2/P_{H_2O}^2 P_{CO_2}]$

Note: (a) = anode; (c) = cathode; E = equilibrium potential.

Source: Hirschenhofer, J.H. et al., *Fuel Cells, A Handbook,* Gilbert/Commonwealth, Morgantown, WV, 1994. With permission.

The maximum conversion efficiency of a fuel cell is given by

$$\eta_{max} = \frac{W_e,max}{\Delta H} = \frac{\Delta G}{\Delta H} = 1 - \frac{T\Delta S}{\Delta H} = \frac{-nFE}{\Delta H} \tag{8.13.28}$$

As current is drawn through an external circuit, the actual voltage drop (V) will be less than the internal cell voltage (E). Therefore, the actual conversion efficiency of a fuel cell will be lower than above and may be calculated as

$$\eta_{actual} = \frac{-nFV}{\Delta H} = \frac{ItV}{\Delta H} \tag{8.13.29}$$

where I is the current drawn through an external circuit for a period of time, t.

Types of Fuel Cells

Fuel cells are primarily classified by type of electrolyte, since many other characteristics, particularly operating temperatures, are limited by the electrolyte properties (Hirschenhofer et al., 1994). Major fuel cells under active development at this time are the **phosphoric acid fuel cell (PAFC)**, the **molten carbonate fuel cell** (**MCFC**), the **solid oxide fuel cell** (**SOFC**), the **polymer electrolyte fuel cell** (**PEFC**), and the **alkaline fuel cell** (**AFC**).

Phosphoric Acid Fuel Cell. PAFC uses concentrated phosphoric acid (H_3PO_4) as the electrolyte, hydrogen as the fuel, and oxygen (from air) as the oxidant. Table 8.13.4 provides information on the electrodes and other materials for PAFC as well as other fuel cells.

The reactions take place at the porous electrodes on highly dispersed electrocatalyst Pt particles supported on carbon black and a polymeric binder, usually polytetrafluoroethylene (PTFE) (about 30 to 50% by weight) (Kinoshita et al., 1988; Hirschenhofer et al., 1994). A porous carbon paper substrate provides structural support for the electrocatalyst and serves as the current collector. A typical carbon paper electrode impregnated with the electrocatalyst has a porosity of about 60%, consisting of micropores of about 34 Å diameter and macropores of 3 to 50 μm diameter.

Dipolar plates (usually graphite) are used to separate the individual cells and electrically connect them in series in a fuel cell stack. In PAFC stacks, provisions are included to remove the waste heat, by liquids (usually water) or gases (usually air) which flow through channels provided in the cell stack.

Molten Carbonate Fuel Cells. MCFCs use as electrolytes mixtures of molten carbonates of lithium (Li_2CO_3), potassium (K_2CO_3), and sodium (Na_2CO_3) in proportions as shown in Table 8.13.4. The operating temperature of the cell (~650°C) is higher than the melting temperature of the carbonate electrolytes. Besides H_2 (fuel) and O_2 (oxidant from air) the cell uses CO_2 which transfers from the cathode to the anode according to the reactions in Table 8.13.1.

According to the reactions, 1 g · mol of CO_2 is transferred along with 2 g · mol of electrons ($2F$). The reversible potential for an MCFC, taking into account the transfer of CO_2, is given by the Nernst equation in Table 8.13.3.

It is usual practice to recycle the CO_2 generated at the anode to the cathode where it is consumed. An early process used to fabricate electrolyte structures involved hot processing mixtures of $LiAlO_2$ and alkali carbonates at temperatures slightly below the melting point of carbonates. These electrolyte structures had problems of void spaces, nonuniformity of microstructure, poor mechanical strength, and high iR drop. To overcome these problems, processes have been developed recently that include tape casting (Man et al., 1984; Hirschenhofer et al., 1994) and electrophoretic deposition (Baumgartner et al., 1985; Hirschenhofer et al., 1994).

Increasing the operating pressures results in enhanced cell voltages. Increasing the operating temperature above 550°C also enhances the cell performance. However, beyond 650°C the gains are diminished and the electrolyte loss and material corrosion are increased. Therefore, 650°C is about the optimum operating temperature.

Solid Oxide Fuel Cell. SOFCs offer a number of advantages over MCFCs for high-temperature operation, since there is no liquid electrolyte. Solid electrolyte allows flexibility in cell shape design based on application. Cells of several shapes, as shown in Figure 8.13.9, are being developed. Because of the high temperature of operation (~1000°C) carbon monoxide (CO) and hydrocarbons such as methane (CH_4) can be used as fuels. At 1000°C, these fuels can easily produce H_2 that is used at the anode by steam reforming and water-gas reactions as

$$CO + H_2O \rightarrow CO_2 + H_2 \quad (\text{water-gas}) \tag{8.13.30}$$

$$CH_4 + H_2O \rightarrow CO + 3H_2 \quad (\text{steam reforming}) \tag{8.13.31}$$

Because of very high operating temperatures the choice of cell materials is limited by (1) chemical stability in oxidizing and reducing atmosphere; (2) chemical stability of contacting materials; and (3) conductivity and thermomechanical compatibility. A detailed description of the current status is given by Minh (1991; 1993) and Appleby and Foulkes (1989). Present SOFC designs make use of thin film wall concepts where films of electrodes, electrolyte, and interconnect material are deposited on each other and sintered to form cell structure. Electrochemical vapor deposition (EVD) is now used to deposit thin layers.

TABLE 8.13.4 Cell Components for Various Fuel Cells

Component	PAFC	MCFC	SOFC[b]	PEFC	AFC
Anode	• PTFE-bonded Pt/C • Vulcan XC-72[a] • 0.1 mg Pt/cm^2	• Ni-10 wt% Cr • 3–6 μm pore size • 50–70% initial porosity • 0.5–1.5 mm thickness • 0.1–1 m^2/g	• Ni/ZrO_2 cermet[c] (30 mol% Ni) • Deposit slurry • 12.5 × 10^{-6} cm/cm °C • ~150 μm thickness • 20–40% porosity	• 10% Pt thin film	• Dual Porosit Ni • 16 μm max pore on electrolyte side • 30 μm pore on gas side
Cathode	• PTFE-bonded Pt/c • Vulcan XC-72[a] • 0.5 mg Pt/cm^2	• Lithiated NiO • 7–15 μm pore size • 60–65% after lithiation and oxidation • 70–80% initial porosity • 0.5–0.75 mm thickness • 0.5 m^{2}g	• Sr-doped lanthanum manganite (10 mol% Sr) • Deposit slurry, sinter • ~1 mm thickness • 12 × 10^{-6} cm/cm °C expansion from room temperature to 1000°C[d] • 20–40% porosity	• 10% Pt thin film	• Porous lithinated NiO
Electrode support	• Carbon paper	—	—	• Carbon paper with Teflon coating on one side	
Electrolyte support	• PTFE-bonded SiC	• γ-$LiAlO_2$ • 0.1–12m^2/g • 0.5 mm thickness	—	—	—
Electrolyte[a]	• 100% H_3PO_4	• 62 Li-38 K • 50 Li-50 Na • 50 Li-50 K • Tape cast • 0.5 mm thickness	• Yttria-stabilized ZrO_2 (8 mol% Y) • EVD[d] • 10.5 × 10^{-6} cm/cm °C expansion from room temperature to 1000°C[d] • ~40 μm thickness	• Proton conducting membrane of perfluoro sulfonic acid polymer	• KOH (45% to 85%)

[a] Conductive oil furnace black, product of Cabot Corp. Typical properties: 002 d-spacing of 3.6 Å by X-ray diffusion, surface area of 220 m^2/g by nitrogen adsorption, and average particle size of 30 μm by electron microscopy.

[b] Specifications for Westinghouse SOFC.

[c] Y_2O_3 stabilized ZrO_2.

[d] EVD = electrochemical vapor deposition.

Source: Hirschenhofer, J.H. et al., *Fuel Cells, Handbook,* Gilbert/Commonwealth, Morgantown, WV, 1994. With permission.

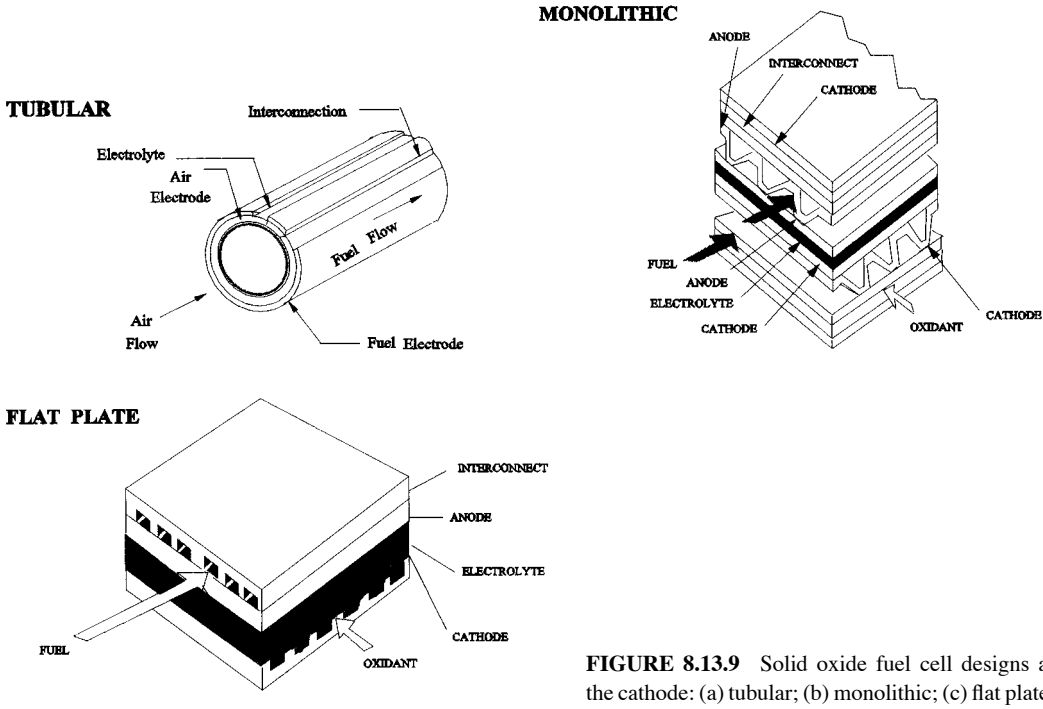


FIGURE 8.13.9 Solid oxide fuel cell designs at the cathode: (a) tubular; (b) monolithic; (c) flat plate.

Increasing pressure and temperature enhances the performance of SOFC.

Polymer Electrolyte Fuel Cell. The basic cell consists of a proton-conducting membrane such as perfluoro sulfonic acid polymer sandwiched between two Pt-impregnated porous electrodes. The backs of the electrodes are made hydrophobic by coating with Teflon®, which provides a path for gas to diffuse to the catalyst layer.

The electrochemical reactions for PEFC are similar to PAFC as given in Table 8.13.1.

The protons from the anode diffuse through the membrane with the help of water molecules soaked in the membrane. The cell operates at low temperature (80°C) and can have very high current densities. Therefore, the cell can be made very compact and can have fast start. There are no corrosive fluids (acids or alkalis) in the cell. Because of these attributes the cell is particularly suited for vehicle-power operation. Present research includes investigation of using methanol and natural gas as the fuel for the cell.

Alkaline Fuel Cell. Alkaline electrolytes have faster kinetics, which allows the use of non-noble metal electrocatalysts. However, AFCs suffer a drastic performance loss if CO_2 is present in the fuel or the oxidant, for example, air. Therefore, AFCs are restricted for use where pure H_2 and O_2 can be used. They have been used in the past in the space program.

Fuel Cell Performance

The performance of fuel cells is affected by the operating variables (e.g., temperature, pressure, gas composition, reactant utilization, and current density) and other factors that influence the reversible cell potential (impurities) and the magnitude of the irreversible voltage losses (polarization, contact resistance, exchange current).

The cell voltage (V_{cell}) is given by

$$V_{cell} = E - iR - \mu_p \tag{8.13.32}$$

where i is the current through the cell, R is the cell resistance, and μ_p is the polarization loss.

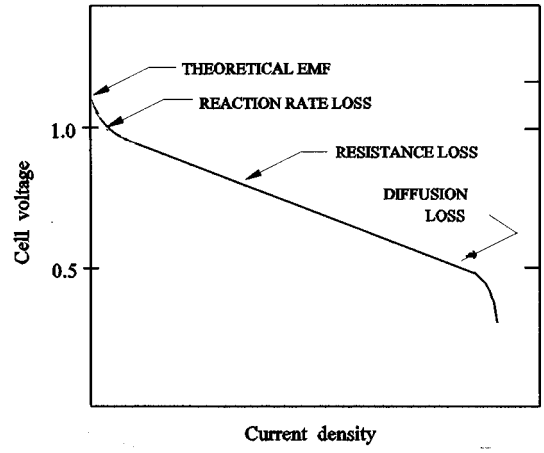


FIGURE 8.13.10 Losses affecting current-voltage characteristics.

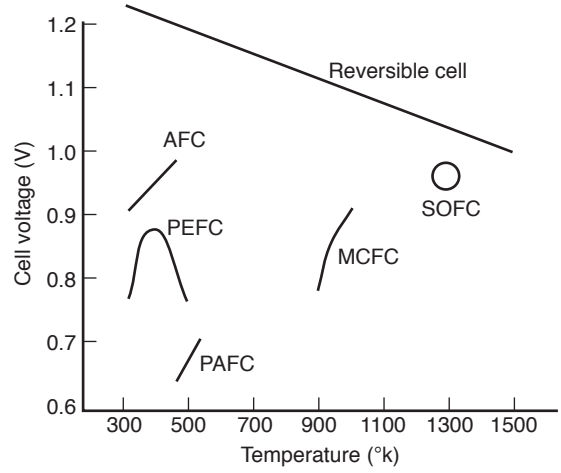


FIGURE 8.13.11 Dependence of the initial operating cell voltage of typical fuel cells on temperature. .

Current Density. Current density has a major impact on the cell voltage. Figure 8.13.10 shows how various losses affect the current-voltage characteristics.

Temperature and Pressure. Increase in pressure generally has a beneficial effect on the cell performance. Increased reactant pressure increases gas solubility and mass transfer rates. In addition, electrolyte loss due to evaporation is decreased.

Theoretically, the reversible potential of an H_2/O_2 fuel cell decreases with an increase in temperature. The practical effect of temperature is mixed, as shown in Figure 8.13.11.

Fuel Cell Power Systems

A general fuel cell power system consists of a fuel processor, fuel cell stack, power conditioner, and possibly a cogeneration or bottoming system to utilize the rejected heat. A schematic of a basic system is shown in Figure 8.13.12.

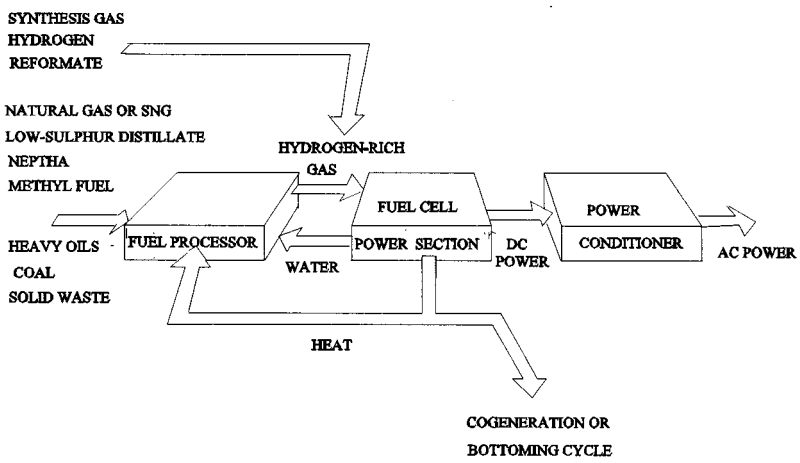


FIGURE 8.13.12 Basic fuel cell power system.

Fuel Processors. If pure hydrogen is available, no additional fuel processor is needed. However, in most applications hydrogen needs to be generated from other fuels, such as natural gas, methane, methanol, etc.

Natural Gas Processing. Natural gas is mostly methane (CH_4) with small amounts of other hydrocarbons. It can be converted to H_2 and CO in a steam-reforming reactor according to Equation (8.13.31). Fuels are typically steam reformed at temperatures of 760 to 980°C.

Liquid Fuel Processing. Liquid fuels such as distillate, naphtha, diesel oil, and fuel oil can be reformed by noncatalytic partial oxidation of the fuel by oxygen in the presence of steam with flame temperatures of 1300 to 1500°C.

Nomenclature

E = internal cell voltage
F = faraday, charge of 1 g-mol of electrons
G = Gibbs free energy
H = enthalpy
ΔH = heat of reaction, enthalpy of formation
I = current
n = number of g-mol
P = Pressure
R = gas constant
S = entropy
t = time
T = temperature
V = voltage drop
W_e = electrical work

Superscript

0 = values under standard conditions — 25°C, 1 atm

Defining Terms

Alkaline fuel cell (AFC): A fuel cell using KOH as the electrolyte.
Faraday: Change of 1 g · mol of electrons, which equals 96,439 C.
Molten carbonate fuel cell (MCFC): A fuel cell using molten carbonate as the electrolyte.

Phosphoric acid fuel cell (PAFC): A fuel cell using phosphoric acid as the electrolyte.
Polymer electrolyte fuel cell (PEFC): A fuel cell using Zerconia as the electrolyte.
Solid oxide fuel cell (SOFC): A fuel cell using potassium as the electrolyte.

References

Angrist, S.W. 1982. Chapter 8, in *Direct Energy Conversion*. Allyn and Bacon, Boston.
Appleby, A.J. and Foulkes, F.R. 1989. *Fuel Cell Handbook,* Van Nostrand Reinhold, New York.
Baumgartner, C.E., DeCarlo, V.J., Glugla, P.G., and Grimaldi, J.J. 1985. *J. Electrochem. Soc,* 132, 57.
Farooque, M. 1990. ERC, Development on Internal Reforming Carbonate Fuel Cell Technology, Final Report, prepared for United States DOE/METC, DOC/MC/23274-2941, pp. 3–19, October.
Hirschenhofer, J.H., Stauffer, D.B., and Engleman, R.R. 1994. *Fuel Cells, A Handbook,* rev. 3. Prepared by Gilbert/Commonwealth, Inc., under U.S. DOE Contract No. DE-AC01-88FE61684, United States Department of Energy, Office of Fossil Energy, Morgantown, WV.
Kinoshita, K., McLarnon, F.R., and Cairns, E.J. 1988. *Fuel Cells, A Handbook*. Prepared by Lawrence Berkeley Laboratory for the United States DOE under contract DE-AC03765F00098.
Maru, H.C., Paetsch, L., and Piegeaud, A. 1984. In *Proceedings of the Symposium on Molten Carbonate Fuel Technology,* R.J. Selman and T.D. Claar, Eds., The Electrochemical Society, Pennington, NJ, p. 20.
Minh, N.Q. 1991. High-temperature fuel cells, Part 2: The solid oxide cell, *Chem. Tech.,* 21, February.
Minh, N.Q. 1993. Ceramic fuel cells, *J. Am. Ceram. Soc.,* 76(3), 563–588.
Pigeaud, A., Skok, A.J., Patel, P.S., and Maru, H.C. 1981. *Thin Solid Films,* 83, 1449.
Wark, K. 1988. *Thermodynamics*, McGraw-Hill, New York, p. 873.

Further Information

Information presented in this section borrows heavily from Hirschenhofer et al. (1994) which lists original references of works published by thousands of researchers across the world. For those references and further information, readers are referred to the *Fuel Cell* handbooks by Hirschenhofer, Stauffer, and Engleman (1994), and Appleby and Foulkes (1989), listed in the References section.

Thermionic Energy Conversion

Mysore L. Ramalingam

Introduction

Thermionic energy conversion (TEC) is the process of converting heat directly to useful electrical work by the phenomenon of thermionic electron emission. This fundamental concept can be applied to a cylindrical version of the planar converter, considered the building block for space nuclear power systems (SNPS) at any power level. Space nuclear reactors based on TEC can produce power in the range of 5 kWe to 5 MWe, a spectrum that serves the needs of current users such as National Aeronautics and Space Administration (NASA), United States Air Force (USAF), United States Department of Energy (USDOE), and Ballistic Missile Defense Organization (BMDO). Electrical power in this range is currently being considered for commercial telecommunication satellites, navigation, propulsion, and planetary exploration missions.

The history of thermionic emission dates back to the mid-1700s when Charles Dufay observed that electricity is conducted in the space near a red-hot body. Although Thomas Edison requested a patent in the late 1800s, indicating that he had observed thermionic electron emission while perfecting his electric light system, it was not until the 1960s that the phenomenon of TEC was adequately described theoretically and experimentally (Hatsopoulos and Gryftopoulos, 1973). These pioneering activities have led to the development of thermionic SNPS that could potentially be augmented by Brayton and Stirling

cycle generators to produce additional power from waste heat in NASA manned lunar and martian exploration missions (Ramalingam and Young, 1993).

Principles of Thermionic Energy Conversion

Figure 8.13.13 represents a schematic of the essential components and processes in an elementary thermionic converter (TC). Electrons "boil-off" from the emitter material surface, a refractory metal such as tungsten, when heated to high temperatures (2000 K) by a **heat source**. The electrons then traverse the small interelectrode gap, to a colder (1000 K) collector surface where they condense, producing an output voltage that drives the current through the electrical load and back to the emitter. The flow of electrons through the electrical load is sustained by the temperature difference and the difference in **surface work functions** of the electrodes.

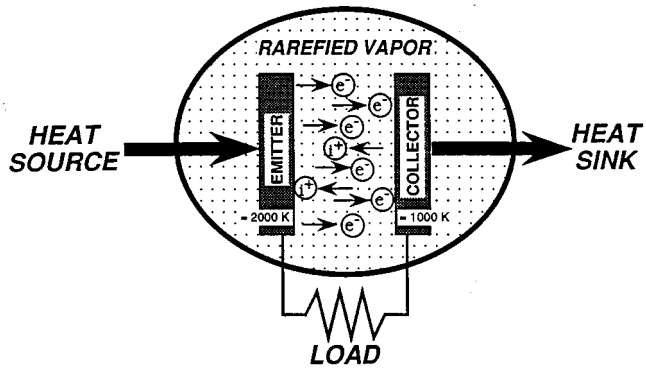


FIGURE 8.13.13 Schematic of an Elementary TEC.

Surface Work Function. In a simple form, the energy required to separate an electron from a metal surface atom and take it to infinity outside the surface is termed the electron work function or the work function of the metal surface. The force experienced by an electron as it crosses an interface between a metal and a rarefied vapor can be represented by the **electron motive**, Ψ, which is defined as a scalar quantity whose negative gradient at any point is a measure of the force exerted on the electron at that point (Langmuir and Kingdon, 1925). At absolute zero the kinetic energy of the **free electrons** would occupy quantum energy levels from zero to some maximum value called the Fermi level. Each energy level contains a limited number of free electrons, similar to the electrons contained in each electron orbit surrounding the nucleus of an atom. Fermi energy, μ, corresponds to the highest energy of all free electrons at absolute zero. At temperatures other than absolute zero some of the free electrons begin to experience energies greater than that at the Fermi level. Thus, the electron work function Φ, would be defined as

$$\Phi = \Psi_T - \mu \tag{8.13.33}$$

where Ψ_T represents the electron motive or energy at some temperature, T, above absolute zero.

Interelectrode Motive Distribution. Figure 8.13.14 provides a schematic representation of the electron motive distribution in the interelectrode space of a thermionic converter. Under ideal conditions of particle transport, the motive varies linearly from Ψ_{EM}, the motive just outside the emitter, to Ψ_{CO}, the motive outside the collector surface. The magnitudes of the Fermi energies of the emitter and collector relative to Ψ_{EM} and Ψ_{CO} are clearly indicated. The internal voltage drop of the converter is defined as;

$$\Delta V = \left(\Psi_{EM} - \Psi_{CO}\right)/e \tag{8.13.34}$$

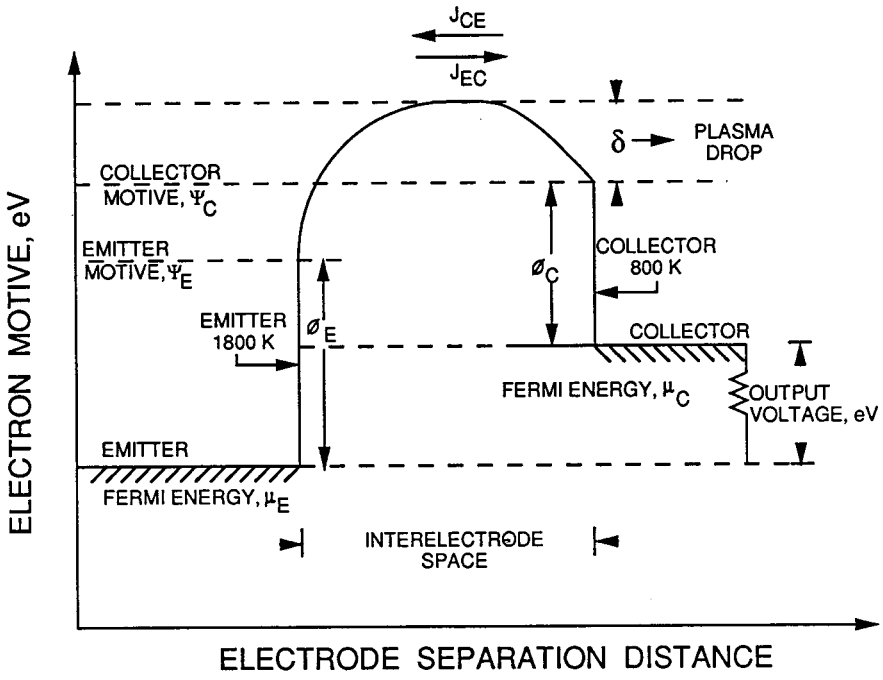


FIGURE 8.13.14 Electron motive distribution in the interelectrode gap.

In a conventional thermionic converter, the emitter and collector are not at the same temperature, but to a good approximation, the output voltage, neglecting **lead losses** and **particle interaction losses**, can be represented by the relationship.

$$V = \left(\mu_{CO} - \mu_{EM}\right)/e \tag{8.13.35}$$

Since a real thermionic converter has an ionizing medium to improve its performance, a similar motive distribution can be defined for the ions. It is sufficient to state that the ion interelectrode motive has a slope equal and opposite to the corresponding electron interelectrode motive. The ions are, therefore, decelerated when the electrons are accelerated and vice versa.

Electron Saturation Current. In the absence of a strong influence from an external electrical source, the electron current ejected from a hot metal at the emitter surface into the vacuum ionizing medium is termed the *electron saturation current*. As this quantity depends on the number of free electrons $N(\varepsilon_x)$, Fermi-Dirac statistics provide the means to compute the number of free electrons, $N(\varepsilon_x)\, d\varepsilon_x$, incident on a unit area within the metal in unit time with energies corresponding to the motion normal to the area, between ε_x and $\varepsilon_x + d\varepsilon_x$. For energies greater than the Fermi energy, the functional dependence of $N(\varepsilon_x)$ on ε_x is given by (Fowler, 1955)

$$N\left(\varepsilon_x\right) \approx \left[4\pi m_e kT/h^3\right]\left[\exp\left\{-\varepsilon_x - \mu/kT\right\}\right] \tag{8.13.36}$$

where m_e is the mass of the electron = 9.108×10^{-28} g and h is Planck's constant = 4.140×10^{-15} eV · sec.

The electron saturation current density, J_{sat}, for a uniform surface, is found by integrating $N(\varepsilon_x)$ in the range of ε_x from Ψ_T to infinity for all $\Psi_T - \mu > kT$, which is the case for almost all materials and practical temperatures. The result of the integration yields

$$J_{\text{sat}} = AT^2 \exp\left[-\left(\Psi_T - \mu\right)/kT\right] \tag{8.13.37}$$

or

$$J_{\text{sat}} = AT^2 \exp\left[-\left(\Phi\right)/kT\right] \tag{8.13.38}$$

where A is the Richardson constant ≈ 120 A/cm$^2 \cdot$ K^2.

Equation (8.13.38), which is the most fundamental and important relationship for the design of a thermionic converter, is called the Richardson-Dushmann equation (Richardson, 1912). On similar lines, the ion saturation current density for a converter with an ionizing medium is given by the relationship (Taylor and Langmuir, 1933):

$$f_{\text{iSat}} = ep_g \Big/ \left[\left(2\pi m_g kT_g\right)^{0.5}\left(1 + 2\exp\left\{\left(V_i - \Phi/kT\right)\right\}\right)\right] \tag{8.13.39}$$

where p_g, T_g, m_g, and V_i are the pressure, temperature, mass, and first ionization energy, respectively, of the ionizing medium.

Types of Thermionic Converters

Thermionic converters can be broadly classified as vacuum thermionic converters and vapor thermionic converters, depending on the presence of an ionizing medium in the interelectrode gap. In vacuum thermionic converters the interelectrode space is evacuated so that the space is free of particles other than electrons and the two electrodes are placed very close together, thereby neutralizing the negative space charge buildup on the electrode surface and reducing the total number of electrons in transit. Due to machining limitations, vacuum converters have been all but completely replaced by vapor-filled thermionic converters. In vapor-filled thermionic converters, the interelectrode space is filled with a rarefied ionizing medium at a vapor pressure generally on the order of 1 to 10 torr. The vapor generally used is cesium as it is the most easily ionized of all stable gases and this can be provided through an external two-phase reservoir or an internal graphite reservoir (Young et al., 1993). The vapor neutralizes the negative space charge effect by producing positive ions at the electrode surfaces and gets adsorbed on the surfaces, thereby altering the work function characteristics.

Converter Output Characteristics

Figure 8.13.15 represents the output current-voltage characteristics for various modes of operation of the vacuum and vapor-filled thermionic converters. Characteristics obtained by not considering particle interactions in the interelectrode gap are generally considered ideal output characteristics. The figure essentially displays three types of converter output current-voltage characteristics, an ideal characteristic, an ignited mode characteristic, and an unignited mode characteristic. For an ideal converter in the interelectrode space the net output current density consists of the electron current density, J_{EMCO} flowing from emitter to collector diminished by the electron current density J_{COEM} flowing from collector to emitter and the ion-current density $f_{i\text{EMCO}}$ flowing from emitter to collector. Thus,

$$J_{\text{net}} = J_{\text{EMCO}} - J_{\text{COEM}} - f_{i\text{EMCO}} \tag{8.13.40}$$

By expressing the individual terms as functions of ϕ, T, and V,

$$\begin{aligned} J_{\text{net}} = {} & AT_{\text{EM}}^2 \exp\left[-\left(\Phi_{\text{EM}}/kT_{\text{EM}}\right)\right] - AT_{\text{CO}}^2 \exp\left[-\left(\Phi_{\text{EM}} - eV\right)/kT_{\text{CO}}\right] \\ & - f_{\text{EMS}} \exp\left[-\left(\Psi_{\text{EM}} - \Psi_{\text{CO}}\right)kT_{\text{EM}}\right] \end{aligned} \tag{8.13.41}$$

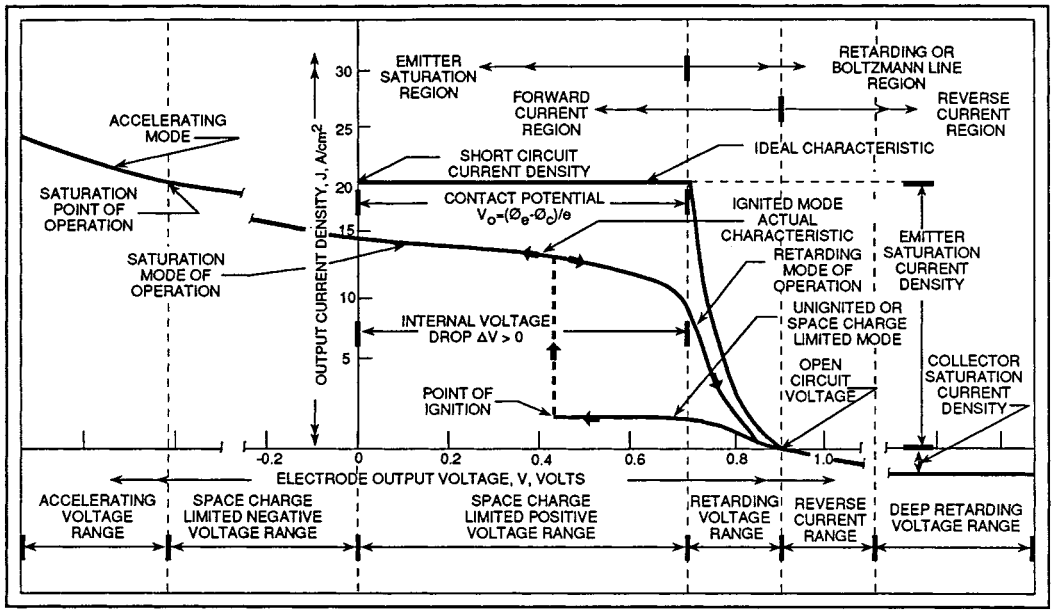


FIGURE 8.13.15 Thermionic diode output current density characteristics and nomenclature.

$$\text{for} \quad eV < \Phi_{EM} - \Phi_{CO}$$

Similar relationships can be generated for various types of thermionic converters.

Thermodynamic Analysis

In thermodynamic terms a thermionic converter is a heat engine that receives heat at high temperature, rejects heat at a lower temperature, and produces useful electrical work while operating in a cycle analogous to a simple vapor cycle engine. Based on the application of the first law of thermodynamics to the control volumes around the emitter (Houston, 1959; Angrist, 1976),

$$\text{Energy In} = \text{Energy Out} \tag{8.13.42}$$

i.e.,

$$q_{CB} + q_{JH} + q_{HS} = q_{EC} + q_{WB} + q_{CD} + q_{RA} \tag{8.13.43}$$

where, by using the terminology in Figure 8.13.14, each of the terms in Equation (8.13.43) can be elaborated as follows:

(a) Energy supplied by back emission of the collector:

$$q_{CB} = J_{COEM}\left[\Phi_{CO} + \delta + V + \left(2kT_{CO}/e\right)\right] \tag{8.13.44}$$

(b) Energy supplied by joule heating of lead wires and plasma:

$$q_{JH} = 0.5\left[J_{EMCO} - J_{COEM}\right]^2\left(R_{LW} + R_{PL}\right) \tag{8.13.45}$$

(c) Energy dissipated by electron cooling:

$$q_{EC} = J_{EMCO}\left[\Phi_{CO} + \delta + V - \Phi_{EM} + \left(2kT_{EM}\right)/e\right] \tag{8.13.46}$$

(d) Energy dissipated due to phase change by electron evaporation:

$$q_{WB} = J_{EM}\Phi_{EM} \tag{8.13.47}$$

(e) Energy dissipated by conduction through the lead wires and plasma:

$$q_{CD} = \Delta T\left[\left(K_{LW}A_{LW}/A_eL_{LW}\right) + \left(K_{PL}A_{PL}/A_eL_{IG}\right)\right] \tag{8.13.48}$$

Here, K represents thermal conductivity, LW = lead wires, PL = plasma, and IG = interelectrode gap.

(f) Energy dissipated by radiation from emitter to collector:

$$q_{RA} = 5.67 \times 10^{-12}\left(T_{EM}^2 - T_{CO}^4\right)\left(\varepsilon_{EM}^{-1} + \varepsilon_{CO}^{-1} - 1\right)^{-1} \tag{8.13.49}$$

Substitution for the various terms in Equation (8.13.42) yields q_{Hs}, the energy supplied to the emitter from the heat source.

The thermal efficiency of the thermionic converter is now expressed as

$$\eta_{TH} = \left[V\left(J_{EMCO} - J_{COEM}\right)/q_{HS}\right] \tag{8.13.50}$$

Design Transition to Space Reactors — Concluding Remarks

All the fundamentals discussed so far for a planar thermionic converter can be applied to a cylindrical version which then becomes the building block for space power systems at any power level. In a thermionic reactor, heat from the nuclear fission process produces the temperatures needed for thermionic emission to occur. The design of a thermionic SNPS is a user-defined compromise between the required output power and the need to operate reliably for a specified lifetime. Based on the type of contact the emitter has with the nuclear fuel, the power systems can be categorized as “in-core” or “out-of-core” power systems. At this stage it suffices to state that the emitter design for in-core systems is extremely complex because of its direct contact with the hot nuclear fuel.

Defining Terms

Electron motive: A scalar quantity whose negative gradient at any point is a measure of the force exerted on an electron at that point.

Free electrons: Electrons available to be extracted from the emitter for thermionic emission.

Heat source: Electron bombardment heating of the emitter.

Lead losses: Voltage drop as a result of the built-in resistance of the leads and joints.

Particle interaction losses: Voltage drop in the interelectrode gap as a result of particle collisions and other interactions.

Surface work function: A measure of the electron-emitting capacity of the surface.

Thermionic energy conversion: Energy conversion from heat energy to useful electrical energy by thermionic electron emission.

References

Angrist, S.W. 1976. *Direct Energy Conversion,* 3rd ed., Allyn and Bacon, Boston.

Fowler, R.H. 1955. *Statistical Mechanics,* 2nd ed., Cambridge University Press, New York.

Hatsopoulos, G.N. and Gyftopoulos, E.P. 1973. *Thermionic Energy Conversion,* Vol. 1, MIT Press, Cambridge, MA.

Houston, J.M. 1959. Theoretical efficiency of the thermionic energy converter, *J. Appl. Phys.,* 30:481–487.

Langmuir, I. and Kingdon, K.H. 1925. Thermionic effects caused by vapors of alkali metals, *Proc. R. Soc. London, Ser. A,* 107:61–79.

Ramalingam, M.L. and Young, T.J. 1993. The power of thermionic energy conversion, *Mech. Eng.,* 115(9):78–83.

Richardson, O.W. 1912. Some applications of the electron theory of matter, *Philos. Mag.,* 23:594–627.

Taylor, J.B. and Langmuir, I. 1933. The evaporation of atoms, ions and electrons from cesium films on tungsten, *Phys. Rev.,* 44:423–458.

Young, T.J., Thayer, K.L., and Ramalingam, M.L. 1993. Performance simulation of an advanced cylindrical thermionic fuel element with a graphite reservoir, presented at 28th AIAA Thermophysics Conference, Orlando, FL.

Further Information

Hatsopoulos, G.N. and Gryftopoulos, E.P. 1979. *Thermionic Energy Conversion,* Vol. 2, MIT Press, Cambridge, MA.

Cayless, M.A. 1961. Thermionic generation of electricity, *Br. J. Appl. Phys.,* 12:433–442.

Hernquist, K.G., Kanefsky, M., and Norman, F.H. 1959. Thermionic energy converter, *RCA Rev.,* 19:244–258.

Rasor, N.S. 1960. Figure of merit for thermionic energy conversion, *J. Appl. Phys.,* 31:163–167.

Ramalingam, M.L. 1993. The Advanced Single Cell Thermionic Converter Program, WL-TR-93-2112, USAF Technical Report, Dayton, OH.

Thermoelectric Power Conversion

Jean-Pierre Fleurial

Introduction

The advances in materials science and solid-state physics during the 1940s and 1950s resulted in intensive studies of thermoelectric effects and related applications in the late 1950s and through the mid-1960s (Rowe and Bhandari, 1983). The development of semiconductors with good thermoelectric properties made possible the fabrication of thermoelectric generators and refrigerators. Being solid-state devices, thermoelectric systems offer some unique advantages, such as high reliability, long life, small-size and no-vibrations refrigerators, and can be used in a wide temperature range, from 200 to 1300 K. However, because of their limited conversion efficiencies, these devices have remained confined to specialized applications. As the following sections will emphasize, the performance of those devices is closely associated with the magnitude of the **dimensionless figure of merit**, *ZT*, of the thermoelectric semiconductor.

ZT represents the relative magnitude of electrical and thermal cross-effect transport in materials. State-of-the-art thermoelectric materials, known since the early 1960s, have been extensively developed. Although significant improvements of the thermoelectric properties of these materials have been achieved, a maximum *ZT* value close to 1 is the current approximate limit over the whole 100 to 1500 K temperature range (Figure 8.13.16). To expand the use of thermoelectric devices to a wide range of applications will require improving *ZT* by a factor of two to three. There is no theoretical limitation on

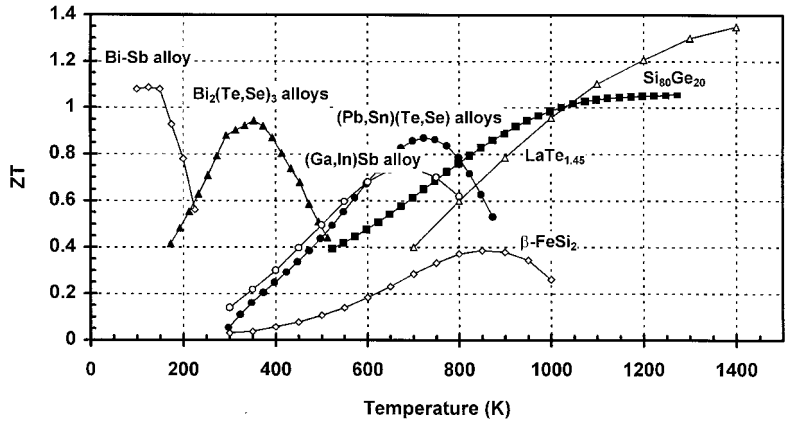


FIGURE 8.13.16 Typical temperature variations of *ZT* of state-of-the-art n-type thermoelectric alloys.

the value of *ZT*, and new promising approaches are now focusing on the investigation of totally different materials and the development of novel thin film heterostructure.

Thermoelectric Effects

Thermoelectric devices are based on two transport phenomena: the Seebeck effect for power generation and the Peltier effect for electronic refrigeration. If a steady temperature gradient is applied along a conducting sample, the initially uniform charge carrier distribution is disturbed as the free carriers located at the high-temperature end diffuse to the low-temperature end. This results in the generation of a back emf which opposes any further diffusion current. The open-circuit voltage when no current flows is the Seebeck voltage. When the junctions of a circuit formed from two dissimilar conductors (n- and p-type semiconductors) connected electrically in series but thermally in parallel are maintained at different temperatures T_1 and T_2, the open-circuit voltage V developed is given by $V = S_{pn}(T_1 - T_2)$, where S_{pn} is the Seebeck coefficient expressed in $\mu V \cdot K^{-1}$.

The complementary Peltier effect arises when an electrical current I passes through the junction. A temperature gradient is then established across the junctions and the corresponding rate of reversible heat absorption $\dot{Q}$ is given by $\dot{Q} = \Pi_{pn} I$, where Π_{pn} is the Peltier coefficient expressed in $W \cdot A^{-1}$ or V. There is actually a third, less-important phenomenon, the Thomson effect, which is produced when an electrical current passes along a single conducting sample over which a temperature gradient is maintained. The rate of reversible heat absorption is given by $\dot{Q} = \beta I(T_1 - T_2)$, where β is the Thomson coefficient expressed in $V \cdot K^{-1}$. The three coefficients are related by the Kelvin relationships:

$$S_{pn} = \frac{\Pi_{pn}}{T} \quad \text{and} \quad \frac{dS_{pn}}{dT} = \frac{\beta_p - \beta_n}{T} \tag{8.13.51}$$

Thermoelectric Applications

The schematic of a thermoelectric device, or module, on Figure 8.13.17, illustrates the three different modes of operation: power generation, cooling, and heating. The *thermoelectric module* is a standardized device consisting of several p- and n-type legs connected electrically in series and thermally in parallel, and bonded to a ceramic plate on each side (typically alumina). The modules are fabricated in a great variety of sizes, shapes, and number of thermoelectric couples and can operate in a wide range of currents, voltages, powers, and efficiencies. Complex, large-scale thermoelectric systems can be easily designed and built by assembling various numbers of these modules connected in series or in parallel depending on the type of applications.

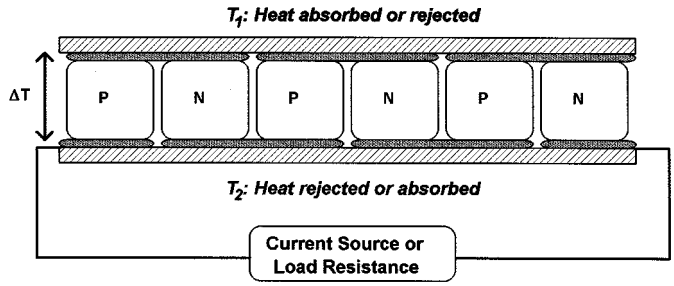


FIGURE 8.13.17 Schematic of a thermoelectric module.

Power Generation. When a temperature gradient is applied across the thermoelectric device, the heat absorbed at the hot junction (Figure 8.13.17, hot side $T_h - T_1$ and cold side, $T_c - T_2$) will generate a current through the circuit and deliver electrical power to the load resistance R_L (Harman and Honig, 1967). The conversion efficiency η of a thermoelectric generator is determined by the ratio of the electrical energy, supplied to the load resistance, to the thermal energy, absorbed at the hot junction, and is given by

$$\eta = \frac{R_L I^2}{S_{\text{pn}} I T_h + K\left(T_h - T_c\right) - \frac{1}{2} RI^2} \tag{8.13.52}$$

where K is the thermal conductance in parallel and R is the electrical series resistance of one p-n thermoelectric couple. The electrical power P_L generated can be written as

$$P_L = \frac{S_{\text{pn}}\left(T_h - T_c\right)^2 R_L}{\left(R_L + R\right)^2} \tag{8.13.53}$$

The thermoelectric generator can be designed to operate at maximum power output, by matching the load and couple resistances, $R_L = R$. The corresponding conversion efficiency is

$$\eta_P = \frac{T_h - T_c}{\frac{3}{2} T_h + \frac{1}{2} T_c + \frac{1}{4} Z_{\text{pn}}^{-1}} \tag{8.13.54}$$

where Z_{pn} is the figure of merit of the p-n couple given by

$$Z_{\text{pn}} = \frac{S_{\text{pn}}^2}{RK} \tag{8.13.55}$$

The figure of merit can be optimized by adjusting the device geometry and minimizing the RK product. This results in Z_{pn} becoming independent of the dimensions of the **thermoelectric legs**. Moreover, if the p- and n-type legs have similar transport properties, the figure of merit, $Z_{\text{pn}} = Z$, can be directly related to the Seebeck coefficient S, electrical conductivity σ or resistivity ρ, and thermal conductivity λ of the thermoelectric material:

$$Z = \frac{S^2}{\rho\lambda} = \frac{S^2\sigma}{\lambda} \tag{8.13.56}$$

The maximum performance η_{max} of the generator is obtained by optimizing the load-to-couple-resistance ratio, leading to the maximum energy conversion efficiency expressed as

$$\eta_{max} = \frac{T_h - T_c}{T_h} \frac{\sqrt{1 + Z_{pn} T_{av}} - 1}{\sqrt{1 + Z_{pn} T_{av}} + \frac{T_c}{T_h}} \tag{8.13.57}$$

It must be noted that the maximum efficiency is thus the product of the Carnot efficiency, less than unity, and of a material-related efficiency, increasing with increasing Z_{pn} values as illustrated in Figure 8.13.18.

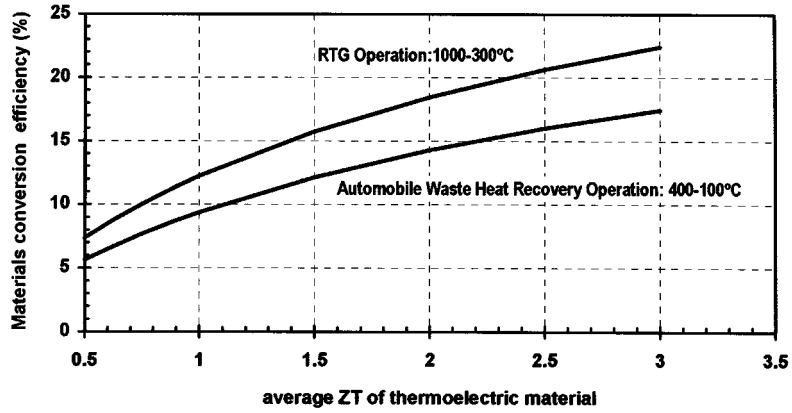


FIGURE 8.13.18 Maximum conversion efficiency η_{max} as a function of the average material figure of merit *ZT,* calculated using Equation (8.13.57) for two systems operating in different temperature ranges: the radioisotope generator (RTG) used for deep space missions and an automobile waste heat recovery generator.

Refrigeration. When a current source is used to deliver electrical power to a thermoelectric device, heat can be pumped from T_1 to T_2 and the device thus operates as a refrigerator (Figure 8.13.17, hot side $T_h = T_2$ and cold side, $T_c = T_1$). As in the case of a thermoelectric generator the operation of a thermoelectric cooler depends solely upon the properties of the p-n thermocouple materials expressed in terms of the figure of merit Z_{pn} and the two temperatures T_c and T_h (Goldsmid, 1986). The conversion efficiency or **coefficient of performance**, COP, of a thermoelectric refrigerator is determined by the ratio of the cooling power pumped at the cold junction to the electrical power input from the current source and is given by

$$\text{COP} = \frac{S_{pn} T_c I - \frac{1}{2} RI^2 - K\left(T_h - T_c\right)}{S_{pn}\left(T_h - T_c\right) I + RI^2} \tag{8.13.58}$$

There are three different modes of operation which are of interest to thermoelectric coolers. A thermoelectric cooler be designed to operate at maximum cooling power, Qc_{max}, by optimizing the value of the current:

$$I_{Qc_{max}} = \frac{S_{pn} T_c}{R} \quad \text{and} \quad Qc_{max} = \frac{1}{2} \frac{\left(ST_c\right)^2}{R} - K\left(T_h - T_c\right) \tag{8.13.59}$$

Similarly, the conditions required for operating at maximum efficiency, COP_{max}, across a constant temperature gradient, are determined by differentiating Equation (8.13.58) with respect to I, with the solution:

$$I_{COP_{max}} = \frac{K(T_h - T_c)_c}{S_{pn} T_{av}} \left(1 + \sqrt{1 + Z_{pn} T_{av}}\right) \tag{8.13.60}$$

$$COP_{max} = \frac{T_c}{T_h - T_c} \frac{\sqrt{1 + Z_{pn} T_{av}} - \frac{T_h}{T_c}}{\sqrt{1 + Z_{pn} T_{av}} + 1} \tag{8.13.61}$$

By reversing the input current to the device, the thermoelectric refrigerator can become a heat pump, with T_1 being the hot junction temperature. The expression of the maximum conversion efficiency of the heat pump is very similar to Equation (8.13.61) because of the following relationship:

$$\left(COP_{max}\right)^{\text{heat pump}} = 1 + \left(COP_{max}\right)^{\text{refrigerator}} \tag{8.13.62}$$

The maximum COP expression in Equation (8.13.61) is similar to the one derived for the conversion efficiency η of a thermoelectric generator in Equation (8.13.57). However, there is a major difference between the COP_{max} and η_{max} parameters. Clearly, η_{max} increases with increasing ΔT values but is limited by the Carnot efficiency (Equation 8.13.54) which is less than 1, while COP_{max} in Equation (8.13.52) increases with decreasing ΔT values and can reach values much larger than 1. Figure 8.13.19 represents the variations of the COP_{max} of a thermoelectric cooling device optimized for working voltage and geometry as a function of average ZT values and temperature differences (hot junction temperature at 300 K). The average ZT value for current state-of-the-art commercially available materials (Bi_2Te_3-based alloys) is about 0.8. For example, it can be seen that a COP_{max} of 4 is obtained for a $(T_h - T_c)$ difference of 10 K, meaning that to pump 8 W of thermal power only 2 W of electrical power needs to be provided to the thermoelectric cooling device. This also means that 10 W of thermal power will be rejected at the hot side of the cooler.

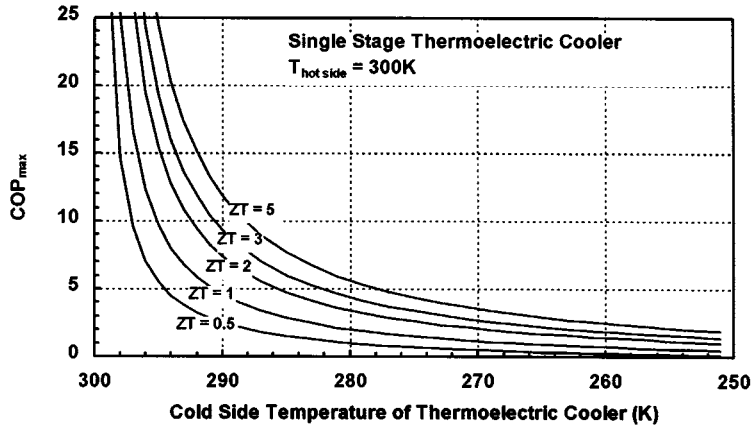


FIGURE 8.13.19 Maximum material coefficient of performance COP_{max} of a single-stage thermoelectric cooler calculated using Equation (8.13.61) as a function of the cold-side temperature (hot-side temperature of 300 K). Curves corresponding to various values of the average material figure of merit are displayed.

The operation of a thermoelectric refrigerator at maximum cooling power necessitates a substantially higher input current than the operation at maximum efficiency. This is illustrated by calculating the variations of the maximum COP and cooling power with the input current and temperature difference which have been plotted in Figures 8.13.20 and 8.13.21. The calculation was based on the properties of a thermoelectric cooler using state-of-the-art Bi_2Te_3-based alloys, and the arbitrary units are the same for both graphs. It can be seen that $I_{COP_{max}}$ increases while $I_{Qc_{max}}$ decreases with increasing ΔT. Also, it is possible to operate at the same cooling power with two different current values.

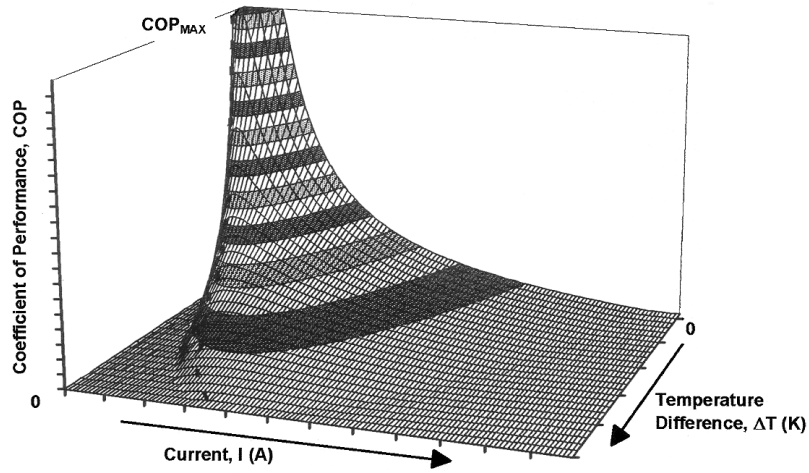


FIGURE 8.13.20 Three-dimensional plot of the variations of the COP of a thermoelectric cooler as a function of the operating current and the temperature difference.

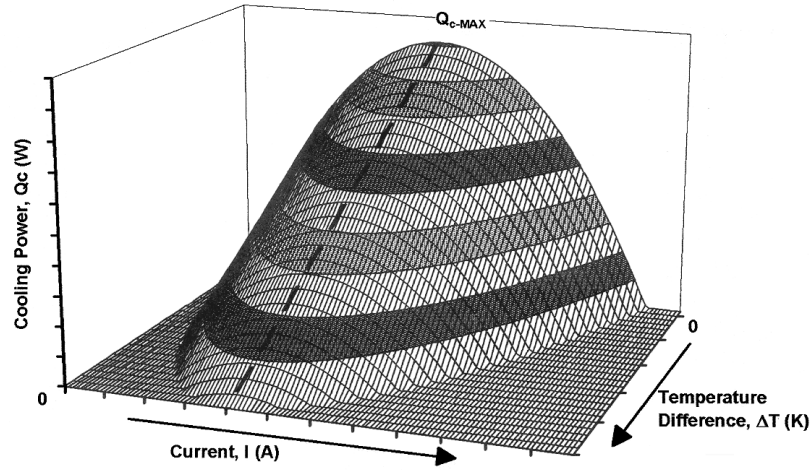


FIGURE 8.13.21 Three-dimensional plot of the variations of the cooling power of a thermoelectric cooler as a function of the operating current and the temperature difference.

Finally, the third problem of interest for thermoelectric coolers is to determine the maximum temperature difference, ΔT_{max}, that can be achieved across the device. As shown on Figure 8.13.21, by operating at maximum cooling power and extrapolating Equation (8.13.59) to $Qc_{max} = 0$, ΔT_{max} is given by

$$\Delta T_{max} = \frac{1}{2} Z_{pn} T_c^2 \quad \text{and} \quad T_{c\,min} = \frac{\sqrt{1 + 2Z_{pn} T_h} - 1}{Z_{pn}} \tag{8.13.63}$$

where $T_{c\,min}$ corresponds to the lowest cold-side temperature achievable. If the cooler operates at a ΔT close to ΔT_{max} or higher, it becomes necessary to consider a cascade arrangement with several **stages**. The COP of an n-stage thermoelectric cooler is optimized if the COP of each stage, COP_i, is the same, which requires $\Delta T_i / T_{i-1}$ to be the same for each stage. The overall maximum COP is then expressed as

$$COP_{max} = \frac{1}{\left(\prod_{i=1}^{n} \left(1 + \frac{1}{COP_i} \right) - 1 \right)} \tag{8.13.64}$$

Additional Considerations

When considering the operation of an actual thermoelectric device, several other important parameters must be considered. The thermal and electrical contact resistances can substantially degrade the device performance, in particular for short lengths of the thermoelectric legs. For example, the conversion efficiency of a radioisotope generator system is about 20% lower than the value calculated in Figure 8.13.18 for the thermoelectric materials only. The electrical contact resistance arises from the connection (see Figure 8.13.17) of all the legs in series. Typical values obtained for actual generators and coolers are 10 to 25 $\mu\Omega \cdot cm^2$. The thermal contact resistance is generated by the heat-transfer characteristics of the ceramic plates and contact layers used to build the thermoelectric module. The heat exchangers and corresponding heat losses should also be taken into account.

In addition, the transport properties of the thermoelectric materials vary with temperature, as illustrated in Figure 8.13.16. When a thermoelectric device is operating across a wide temperature range, these variations should be factored in the calculation of its performance.

Nomenclature

COP	coefficient of performance
COP_{max}	maximum coefficient of performance
COP_i	coefficient of performance of the *i*th stage of a multistage thermoelectric cooler
I	current intensity
$I_{COP_{max}}$	current intensity required to operate a thermoelectric cooler at maximum conversion efficiency
$I_{Qc_{max}}$	current intensity required to operate a thermoelectric cooler at maximum cooling power
K	thermal conductance
Q	rate of reversible heat absorption
R	electrical resistance
R_L	load resistance
P_L	electrical power delivered to the load resistance
S	Seebeck coefficient
S_{pn}	Seebeck coefficient of a p-n couple of thermoelements
T_1	temperature
T_2	temperature
T_{av}	average temperature across the thermoelectric device
T_c	cold-side temperature of a thermoelectric device
$T_{c\,min}$	minimum cold-side temperature which can be achieved by a thermoelectric cooler
T_h	hot-side temperature of a thermoelectric device
V	voltage; open-circuit voltage
Z	thermoelectric figure of merit

Z_{pn} thermoelectric figure of merit of a p-n couple of thermoelements
ZT dimensionless thermoelectric figure of merit
β Thomson coefficient
β_p Thomson coefficient for the p-type thermoelement
β_n Thomson coefficient for the n-type thermoelement
ΔT temperature difference across a thermoelectric device
ΔT_{max} maximum temperature difference which can be achieved across a thermoelectric cooler
η thermoelectric conversion efficiency
η_{max} maximum thermoelectric conversion efficiency
λ thermal conductivity
Π_{pr} Peltier coefficient
ρ electrical resistivity

Defining Terms

Coefficient of performance: Electrical to thermal energy conversion efficiency of a thermoelectric refrigerator, determined by the ratio of the cooling power pumped at the cold junction to the electrical power input from the current source.

Dimensionless figure of merit: The performance of a thermoelectric device depends solely upon the properties of the thermoelectric material, expressed in terms of the dimensionless figure of merit *ZT*, and the hot-side and cold-side temperatures. *ZT* is calculated as the square of the Seebeck coefficient times the absolute temperature divided by the product of the electrical resistivity to the thermal conductivity. The best *ZT* values are obtained in heavily doped semiconductors, such as Bi_2Te_3 alloys, PbTe alloys, and Si-Ge alloys.

Stage: Multistage thermoelectric coolers are used to achieve larger temperature differences than possible with a single-stage cooler composed of only one module.

Thermoelectric leg: Single thermoelectric element made of n-type or p-type thermoelectric material used in fabricating a thermoelectric couple, the building block of thermoelectric modules. The geometry of the leg (cross-section-to-length ratio) must be optimized to maximize the performance of the device.

Thermoelectric module: Standardized device consisting of several p- and n-type legs connected electrically in series and thermally in parallel, and bonded to a ceramic plate on each. The modules are fabricated in a great variety of sizes, shapes, and number of thermoelectric couples.

References

Goldsmid, H.J. 1986. *Electronic Refrigeration,* Pion Ltd., London.

Hannan, T.C. and Honig, J.M. 1967. *Thermoelectric and Thermomagnetic Effects and Applications,* McGraw-Hill, New York.

Rowe, D.M and Bhandari, C.M. 1983. *Modern Thermoelectrics,* Reston Publishing, Reston, VA.

Further Information

The *Proceedings of the Annual International Conference on Thermoelectrics* are published annually by the International Thermoelectric Society (ITS). These proceedings provide the latest information on thermoelectric materials research and development as well as thermoelectric devices and systems. The ITS also publishes a semiannual newsletter. For ITS membership or questions related to thermoelectrics, you may contact the current ITS secretary: Dr. Jean-Pierre Fleurial, Jet Propulsion Laboratory, MS 277-212, Pasadena, CA 91109. Phone (818)-354-4144. Fax (818) 393-6951. E-mail jean-pierre.fleurial@jpl.nasa.gov.

Also, the *CRC Handbook of Thermoelectrics,* edited by D.M. Rowe was published by CRC Press Inc., Boca Raton, FL, became available in 1996. This handbook covers all current activities in thermoelectrics.

Magnetohydrodynamic Power Generation

William D. Jackson

Introduction

The discipline known as magnetohydrodynamics (MHD) deals with the interactions between electrically conducting fluids and electromagnetic fields. First investigated experimentally by Michael Faraday in 1832 during his efforts to establish the principles of electromagnetic induction, application to energy conversion yields a heat engine which has its output in electrical form and, therefore, qualifies as a direct converter. This is generally referred to as an MHD generator, but could be better described as an electromagnetic turbine as it operates on a thermodynamic cycle similar to that of a gas turbine.

The operating principle is elegantly simple, as shown in Figure 8.13.22. A pressurized, electrically conducting fluid flows through a transverse magnetic field in a channel or duct. Electrodes located on the channel walls parallel to the magnetic field and connected through an external circuit enable the motionally induced "Faraday electromotive force" to drive an electric current through the circuit and thus deliver power to a load connected into it. Taking the fluid velocity as **u** and the magnetic flux density as **B**, the intensity of the motionally induced field is **u** × **B**. The current density, **J**, in the channel for a scalar conductivity σ is then given by Ohm's law for a moving conductor as

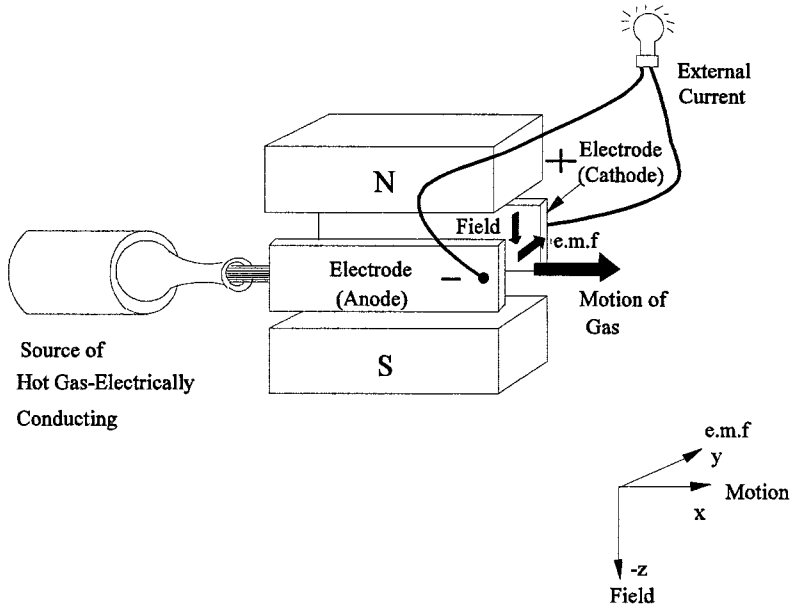


FIGURE 8.13.22 Principle of electromagnetic turbine or MHD generator.

$$\mathbf{J} = \sigma[\mathbf{E} + \mathbf{u} \times \mathbf{B}]$$

By taking the coordinates of Figure 8.13.22 and assuming that the quantities are constant, the power density flow from the MHD generator is, using **E** · **J**

$$w_e = \sigma u_x^2 B^2 k(1-k)$$

where $k = Ez/u \times B$ is the "loading factor" relating loaded electric field to open circuit induction and is used in the same manner as the regulation of an electrical machine is applied to its terminal voltage.

It is instructive at this point to determine the power density of an MHD generator using values representative of the most commonly considered type of MHD generator. Combustion gas with $\sigma = 10$ S/m, a flow velocity of 800 m/sec and an applied field of 5 T for maximum power transfer ($k = 0.5$) yields w_e as 40 MW/m^3. This value is sufficiently attractive to qualify MHD generators for bulk power applications. An intensive, worldwide development effort to utilize this and other MHD generator properties has been conducted since the late 1950s. However, this has not yet led to any significant application of MHD generators. To understand how this has come about and what still needs to be accomplished to make MHD attractive will now be summarized.

Electrical Conductivity Considerations

Two approaches have been followed to obtain adequate ionization and, therefore, conductivity in the working fluid. What may be termed the mainline approach to achieving electrical conductivity is to add a readily ionizable material to "seed" the working fluid. Alkali metals with ionization potentials around 4 V are obvious candidates, although a lower value would be highly desirable. A potassium salt with an ionization potential of 4.09 eV has been widely used because of low cost but cesium with a 3.89-eV value is the preferred seed material when the running time is short or the working fluid is recycled. There are two methods of ionization:

1. Thermal ionization in which recombination ensures a common temperature for electrons, ions, and neutrals; the mass action law (Saha equation) is followed; and the heat of ionization in electron volts is the ionization potential; and
2. Extrathermal or nonequilibrium ionization where electrons and heavy particles are at different temperatures and the concept of entwined fluids (electron, ion, and neutral gases) is involved.

The former is applicable to diatomic combustion gases while the latter occurs in monatomic (noble) gases but is also observed in hydrogen. Only a small amount of seed material is required and is typically around 1% of the total mass flow for maximum conductivity.

The existence of mutually perpendicular **E** and **B** fields in an MHD generator is of major significance in that the electrons are now subjected to the Hall effect. This causes electrons and, therefore, electric currents to flow at an angle with respect to the **E** field in the collision-dominated environment of the MHD generator. The presence of a significant Hall effect requires that the electrical boundary conditions on the channel be carefully selected and also introduces the possibility of working fluid instabilities driven by force fluctuations of electrical origin. A further source of fluctuations and consequent loss of conductivity occurs when nonequilibrium ionization is employed due to current concentration by Joule heating. This latter effect is controlled by operating only in a regime where, throughout the channel, complete ionization of the seed material is achieved.

Generator Configurations and Loading

The basic consequence of the Hall effect is to set up **E** fields in both transverse and axial directions in the generator channel and these are generally referred to as the Faraday and Hall fields, respectively. The direction of the Faraday field depends on the magnetic field; the Hall field depends on the Hall parameter and is always directed toward the upstream end of the channel. These considerations, in turn, lead to the MHD generator having the characteristics of a gyrator — a two-terminal pair power-producing device in which one terminal pair (Faraday) is a voltage source and the other (Hall) is a current source dependent in this case on the Hall parameter. Electric power can be extracted from either the Faraday or Hall terminals, or both.

This has resulted in several electrical boundary conditions being utilized with the axial flow channel as shown in Figure 8.13.23. These are most readily understood by treating each anode-cathode pair as a generating cell. The segmented Faraday configuration (Figure 8.13.23a) is then simply a parallel operation of cells which leads to the apparently inconvenient requirement of separate loading of individual

cells: the Hall connection (Figure 8.13.23b) utilizes a single load by series connection but depends on the Hall parameter for its performance; and the diagonal connection (Figure 8.13.23c) connects the cells in series-parallel and so avoids Hall parameter dependence while retaining the single load feature. In all three linear configurations, the channel walls are electrically segmented to support the Hall field, and experience has shown that this must be sufficiently finely graded so that no more than 50 V is supported by the interelectrode gaps to avoid electrical breakdown.

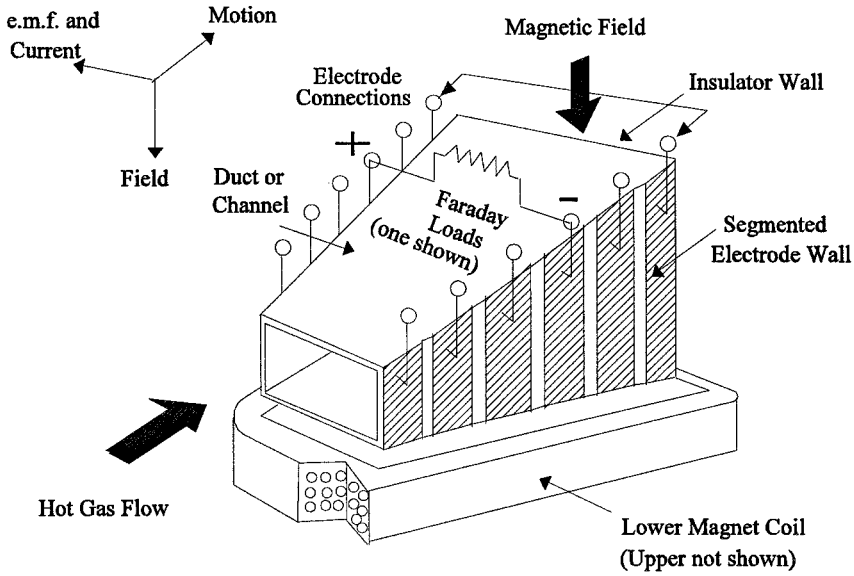


FIGURE 8.13.23 Basic Faraday linear MHD generator.

The MHD generator is a linear version of the homopolar machine originally demonstrated by Faraday and is, as a practical matter, confined to DC generation. Accordingly, some form of DC-AC conversion using power electronics is required for the vast majority of applications. The single load feature loses significance in this situation as the "power conditioning" can readily be arranged to combine (consolidate) the individual cell outputs before conversion to the required AC system conditions. Indeed, to maximize power extraction while limiting interelectrode voltages and controlling electrode currents (to ensure adequate lifetime), the power conditioning is arranged to extract power from both Faraday and Hall terminal pairs.

An alternative geometry is to set up a radial flow (usually but not necessarily outward) with the disk configuration of Figure 8.13.23d. The result is a Hall generator, which is generally favored for nonequilibrium ionization as it avoids the inevitable nonuniformities associated with electrode segmentation with their proclivity for promoting ionization instabilities. A measure of Faraday performance is achievable through the introduction of swirl, and additional ring electrodes enable power conditioning to control (and optimize) the radial electric field.

Components

An MHD generator per se requires several components to make up a complete powertrain. In addition to the power conditioning needed for DC-AC conversion these include a magnet, seed injector, combustor with fuel and oxidizer supply or an input heat exchanger, nozzle, compressor, diffuser, exhaust gas-cleaning system (for once-through systems), and controls. The need to accommodate a channel between the poles of a magnet qualifies the MHD generator as a large-air-gap machine.

Systems

Power systems incorporating MHD generators are either of the once-through (open-cycle) or working fluid recycle (closed-cycle) type, and the complete MHD system described in the previous section can either be stand-alone or incorporated into a more complex configuration such as a binary cycle. In the latter case, the high-temperature operation of the MHD unit makes it a topping cycle candidate and it is in this mode that MHD has received most system consideration. An MHD generator operated directly on ionized combustion gas is usually associated with an open cycle while nonequilibrium ionization with seeded noble gases and LMMHD are invariably considered for closed-cycle operation. An exception is nonequilibrium ionization in cesium-seeded hydrogen which has received some attention for open-cycle operation.

Heat Sources, Applications, and Environmental Considerations

A heat source capable of providing from 1000 K for LMMHD to over 3000 K for open-cycle systems for power production is a candidate. Rocket motor fuels, all fossil fuels, high-temperature nuclear reactors cooled with hydrogen and biomass are suitable for open cycles, while closed cycles can be driven through a heat exchanger by any of these combustion sources. A high-temperature nuclear reactor, probably helium cooled, is also a feasible source for MHD and in the early stages of development of the process received much attention. With the abandonment of efforts to develop commercial reactors to operate at temperatures above 1200 K, attention has focused on high-energy fuels for pulse power (few seconds) operation and coal for utility power generation.

While the debate over the role of fossil energy in the long-term electricity generation scenario continues, it is established that coal is a major resource which must be utilized at maximim efficiency to limit CO_2 production and must be combusted in a manner which reduces SO_2 and NO_x effluents to acceptable levels. The use of MHD generators significantly advances all of these objectives. Briefly, it was first observed that the "electromagnetic turbine" has the major advantage that it cannot only provide the highest efficiency of any known converter from the Carnot viewpoint but that its operation is not adversely affected by coal slag and ash. Indeed, slag provides an excellent renewable coating for the channel walls and increases lifetime.

System calculations have shown that, when coupled as a topping cycle to the best available steam plant technology, a thermal efficiency with coal and full environmental control is 52.5%. When coupled into a ternary cycle with either a gas turbine or fuel cells and a steam turbine, efficiencies upward of 60% are possible.

Technology Status

A pulse-type MHD generator was successfully built and operated by Avco (now Textron Defense Industries) in 1963 and a complete natural gas-fired pilot plant with a nominal 20-MW MHD generator was commissioned in the U.S.S.R. on the northern outskirts of Moscow in 1971. In the decade of the 1980s, development was focused on coal firing as a result of the oil crises of the 1970s and in the U.S. progressed to the point where the technology base was developed for a demonstration plant with a 15-MW MHD generator.

Future Prospects

The two particular attributes of the MHD generator are its rapid start-up to multimegawatt power levels for pulse power applications and its ability to provide a very high overall thermal efficiency for power plants using coal while meeting the most stringent environmental standards. The first has already been utilized in crustal exploration, and the second must surely be utilized when coal is the fuel of choice for electric power production. In the meantime, MHD has been established as a viable technology on which further development work will be conducted for advanced applications such as the conversion system for thermonuclear fusion reactors.

Further Information

The following proceedings series contain a full and complete record of MHD generator and power system development:

1. *Proceedings of the Symposia for the Engineering Aspects of Magnetohydrodynamics*, held annually in the U.S. since 1960 (except for 1971 and 1980).
2. *Proceedings of 11 International Conferences on Magnetohydrodynamic Electrical Power Generation*, held in 1962, 1964, 1966, 1968, 1971 , 1975, 1980, 1983, 1986, 1989, and 1992. The 12th conference will be held in Yokahama, Japan in October 1996.

8.14 Ocean Energy Technology

Desikan Bharathan and Federica Zangrando

The ocean contains a vast renewable energy potential in its waves and tides, in the temperature difference between cold, deep waters and warm surface waters, and in the salinity differences at river mouths (SERI, 1990; WEC, 1993; Cavanagh et al., 1993). Waves offer a power source for which numerous systems have been conceived. Tides are a result of the gravity of the sun, the moon, and the rotation of the Earth working together. The ocean also acts as a gigantic solar collector, capturing the energy of the sun in its surface water as heat. The temperature difference between warm surface waters and cold water from the ocean depths provides a potential source of energy. Other sources of ocean energy include ocean currents, salinity gradients, and ocean-grown biomass.

The following sections briefly describe the status and potential of the various ocean energy technologies, with emphasis placed on those with a near-term applicability.

Ocean Thermal Energy Conversion

Ocean thermal energy conversion (OTEC) technology is based on the principle that energy can be extracted from two reservoirs at different temperatures (SERI, 1989). A temperature difference as little as 20°C can be exploited effectively to produce usable energy. Temperature differences of this magnitude prevail between ocean waters at the surface and at depths up to 1000 m in many areas of the world, particularly in tropical and subtropical latitudes between 24° north and south of the equator. Here, surface water temperatures typically range from 22 to 29°C, while temperatures at a depth of 1000 m range from 4 to 6°C. This constitutes a vast, renewable resource, estimated at 10^{13} W, for potential baseload power generation.

Recent research has been concentrated on two OTEC power cycles, namely, **closed-cycle** and **open-cycle**, for converting this thermal energy to electrical energy (Avery and Wu, 1994). Both cycles have been demonstrated, but no commercial system is yet in operation. In a closed-cycle system, warm seawater is used to vaporize a working fluid such as ammonia flowing through a heat exchanger (evaporator). The vapor expands at moderate pressures and turns a turbine. The vapor is condensed in another heat exchanger (condenser) using cold seawater pumped from the ocean depths through a cold-water pipe. The condensed working fluid is pumped back to the evaporator to repeat the cycle. The working fluid remains in a closed system and is continuously circulated. In an open cycle system, the warm seawater is "flash" evaporated in a vacuum chamber to make steam at an absolute pressure of about 2.4 kPa. The steam expands through a low-pressure turbine coupled to a generator to produce electricity. The steam exiting the turbine is condensed by using cold seawater pumped from the ocean depths through a cold-water pipe. If a surface condenser is used, condensed steam provides desalinated water.

Effluent from either a closed cycle or an open cycle system can be further processed to enhance production of desalinated water through a flash evaporator/condenser system in a second stage.

For combined production of power and water, these systems are estimated to be competitive with conventional systems in several coastal markets.

Tidal Power

The energy from tides is derived from the kinetic energy of water moving from a higher to a lower elevation, as for hydroelectric plants. High tide can provide the potential energy for seawater to flow into an enclosed basin or estuary that is then discharged at low tide (Ryan, 1980). Electricity is produced from the gravity-driven inflow or outflow (or both) through a turbogenerator. The tidal resource is variable but quite predictable, and there are no significant technical barriers for deployment of this technology. Because costs are strongly driven by the civil works required to dam the reservoir, only a few sites around the world have the proper conditions of tides and landscape to lend themselves to this technology.

Although it has benefited from some recent developments in marine and offshore construction that significantly reduce construction time and costs, the economics of tidal power production still does not make it competitive with conventional energy systems.

The highest tides in the world can reach above 17 m, as in the Bay of Fundy between Maine and Nova Scotia, where it is projected that up to 10,000 MW could be produced by tidal systems in this bay alone. A minimum tidal range (difference between mean high and low tides) of 5 m is required for plants using conventional hydroelectric equipment (horizontal axial-flow water turbines housed in underwater bulbs or Straflo turbines). More recently, low-head hydroelectric power equipment has proved adaptable to tidal power and new systems for 2-m heads have been proposed.

There are a few tidal power stations operating in France, the former U.S.S.R., China, and Canada. The largest and longest-operating plant is the 240-MW tidal power station on the Rance River estuary in northern France (Banal and Bichon, 1981), which has operated with 95% availability since 1968. The 400-kW tidal plant in Kislaya Bay on the Barents Sea in the former U.S.S.R. has been operating since 1968; at this favorable site only a 50-m-wide dam was needed to close the reservoir. The 20-MW Canadian plant at Annapolis on the Bay of Fundy uses a Straflo turbine generator and has operated reliably since 1984. A number of small-bulb and Straflo turbine generator plants of up to 4 MW are also installed on the China coastline.

Wave Power

Waves contain significant power which can be harnessed by shore-mounted or offshore systems, the latter having larger incident power on the device but requiring more costly installations. A myriad of wave-energy converter concepts have been devised, transforming wave energy into other forms of mechanical (rotative, oscillating, or relative motion), potential, or pneumatic energy, and ultimately into electricity; very few have been tested at sea.

The power per unit frontal length of the wave is proportional to wave height squared and to wave period, with their representative values on the order of 2 m and 10 sec. The strong dependence on wave height makes the resource quite variable, even on a seasonal and a yearly average basis. The northeastern Pacific and Atlantic coasts have average yearly incident wave power of about 50 kW/m, while near the tip of South America the average power can reach 100 kW/m. Japan, on the other hand, receives an average of 15 kW/m. Waves during storms can reach 200 kW/m, but these large waves are unsafe for operation because of their severity, and they impose significant constraints and additional costs on the system. Overall, the amount of power that could be harvested from waves breaking against world coastlines is estimated to be on the order of the current global consumption of energy. However, total installed capacity amounts to less than 1 MW worldwide.

A commonly deployed device is the oscillating water column (OWC), which has so far been mounted on shore but is also proposed for floating plants. It consists of an air chamber in contact with the sea so that the water column in the chamber oscillates with the waves. This motion compresses the air in the chamber, which flows in and out of a Wells turbine. This turbine can be self-rectifying, i.e., it uses the airflow in both directions, and it consists of a simple rotor, with symmetrical airfoil blades attached tangentially to a central disk. Inertial energy storage (flywheels) is often used to even out the variable pneumatic energy delivered by the air.

Two of the largest wave-energy power plants were built at Toftestallen, near Bergen, Norway. A Norwegian company, Norwave A.S., built a 350-kW tapered channel (Tapchan) device in 1984, which survived a severe storm in 1989 (the 500 kW multi-resonant OWC plant built by Kvaerner Brug A.S. did not). The channel takes advantage of the rocky coastline to funnel waves through a 60-m-wide opening into a coastal reservoir of 5500 m^2, while maintaining civil engineering costs to a minimum. Wave height increases as the channel narrows over its 60-m length, and the rising waves spill over the 3-m-high channel walls, filling the reservoir. Continuous wave action maintains the reservoir level at a relatively constant elevation above sea level, providing potential energy for a low-head hydroelectric Kaplan turbogenerator. Estimates by Norwave to rebuild an identical plant at this site suggested capital

costs of $3500/kW installed, and energy costs of 8¢/kWhr, at a plant capacity factor of 25% (ASCE, 1992). In recent efforts, the National Institute of Ocean Technology has installed a 150-kW wave-energy conversion device in the southern tip of India.

Concluding Remarks

Among the many ocean energy prospects, OTEC, tides, and tapered channel wave-energy converters offer the most near-term potential and possess applicability for a large variety of sites. To realize their potential, additional research and development is required.

Defining Terms

Ocean thermal energy conversion (OTEC): A system that utilizes the temperature difference between the seawater at the surface and at depths.

Closed-cycle OTEC: Uses a working fluid in a closed cycle.

Open-cycle OTEC: Uses steam flashed from the warm seawater as the working fluid which is condensed and exhausted.

References

ASCE, 1992, *Ocean Energy Recovery, The State of the Art,* R.J. Seymour, Ed., American Society of Civil Engineers, New York.

Avery, W.H. and Wu, C. 1994, *Renewable Energy from the Ocean, a Guide to OTEC,* Oxford University Press, New York.

Banal, M. and Bichon, A. 1981. Tidal Energy in France: The Rance Estuary Tidal Power Station — Some Results after 15 Years of Operation, Paper K3, Second Symposium on Wave and Tidal Energy, Cambridge, England, September.

Cavanagh, J.E., Clarke, J.H., and Price, R. Ocean energy systems, in *Renewable Energy, Sources for Fuels and Electricity,* T.B. Johansson, H. Kelley, A.K.N. Reddy, and R.H. Williams (Eds.), Island Press, Washington, D.C., 1993, chap 12.

Ryan, P.R. 1979/80. Harnessing power from tides: state of the art, *Oceanus,* 22(4), 64–67.

SERI, 1989. *Ocean Thermal Energy Conversion: An Overview*, Solar Energy Research Institute, SERI/SP-220-3024, Golden, CO.

SERI, 1990. *The Potential of Renewable Energy: An Interlaboratory White Paper*, Solar Energy Research Institute, SERI/TP-260-3674, Golden, CO.

WEC, 1993. *Renewable Energy Resources — Opportunities and Constraints 1990–2020*, World Energy Council, London, England.

Further Information

CEC, *1992 Energy Technology Status Report*, California Energy Commission, Sacramento, CA, 1992.

Funakoshi, H., Ohno, M., Takahashi, S., and Oikawa, K. Present situation of wave energy conversion systems, *Civil Eng. Jpn.,* 32, 108–134, 1993.

8.15 Combined Cycle Power Plants

William W. Bathie

There is a considerable amount of energy available in the exhaust gases of a gas turbine engine that can be used as the energy source for another system. Table 8.15.1 lists exit temperatures and flow rates for several present-day gas turbine engines.

TABLE 8.15.1 Gas Turbine Exhaust Temperatures and Mass Flow Rates for Several Gas Turbines

Manufacturer Model	Year Available	ISO Base Rating, kW	Heat Rate, Btu/kWhr	Mass Flow, lb/sec	Exhaust Temperature
ABB					
GT 5	1993	2,650	12,544	33.5	445°C
GT 11N	1987	83,800	10,403	699.0	505°C
GT 13D	1970	97,900	10,564	869.0	490°C
GT 26	1994	240,000	9,030	1195.0	608°C
GE Power Systems					
PG 5371 (PA)	1987	26,300	11,990	270.0	909°F
PG 6541 (B)	1978	38,340	10,780	302.0	1002°F
PG 7161 (EC)	1994	116,000	9,890	769.0	1030°F
PG 9311 (FA)	1992	226,500	9,570	1327.0	1093°F
Siemens (KWU)					
V 64.3	1990	62,200	9,720	423.0	529°C
V 84.3A	1994	170,000	8,980	1000.0	562°C
V 94.3A	1995	240,000	8,980	1410.0	562°C

Source: 1995 Handbook, Gas Turbine World, Southport, CT, 1995. With permission.

In the past two decades, several cycles which combine the gas turbine and the steam turbine have become available with good fuel utilization compared with other available power plants. The first ones in the 1970s had net plant efficiencies of about 40% with the most recent ones achieving net plant efficiencies of over 55%.

One way to utilize the energy available in gas turbine exhaust gases is as the energy source for a steam power plant. This is called a *combined cycle* and uses components from two systems which have independently proved themselves. The combined cycle power plant usually consists of one or more gas turbine engines exhausting into a heat-recovery steam generator (HRSG) where the energy is transferred from the exhaust gases to the water in the steam power plant.

The simplest arrangement for a combined cycle power plant is the single-pressure system as illustrated in Figure 8.15.1. It consists of the HRSG, a turbine, condenser, and pump. In this simplified configuration, it is assumed that no additional fuel is burned between the gas turbine and the HRSG.

The gases leaving the gas turbine engine enter the HRSG at state 4 and leave the HRSG at state 5. The water in the steam cycle portion of the combined cycle enters the HRSG economizer as a subcooler liquid (state a). The temperature of the water is increased in the economizer until it is a saturated liquid (state m). This is the point where the minimum temperature difference between the water in the steam cycle and the exhaust gases occurs and is called the *pinch point*. Typical pinch point values range from 10 to 30°C; the smaller the pinch point difference, the larger the required heat-transfer surface area. From state m to state v, the water is changed from a saturated liquid to a saturated vapor. From state v to state d, the water is superheated where it has its maximum temperature. A typical temperature-heat transfer diagram for this single-pressure steam turbine combined cycle is illustrated in Figure 8.15.2.

One should note that there is a significant difference between the desired feedwater conditions of a combined cycle power plant and a conventional steam power plant. In a combined cycle power plant, it is desirable to have the temperature of the water at as low a temperature as possible as it enters the HRSG to permit the maximum amount of energy to be transferred from the exhaust gases to the water

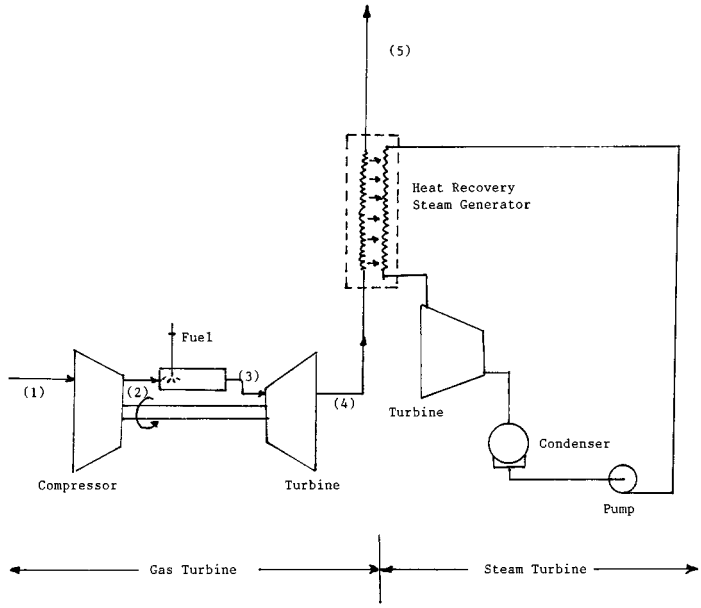


FIGURE 8.15.1 Combined cycle power plant with no supplementary firing.

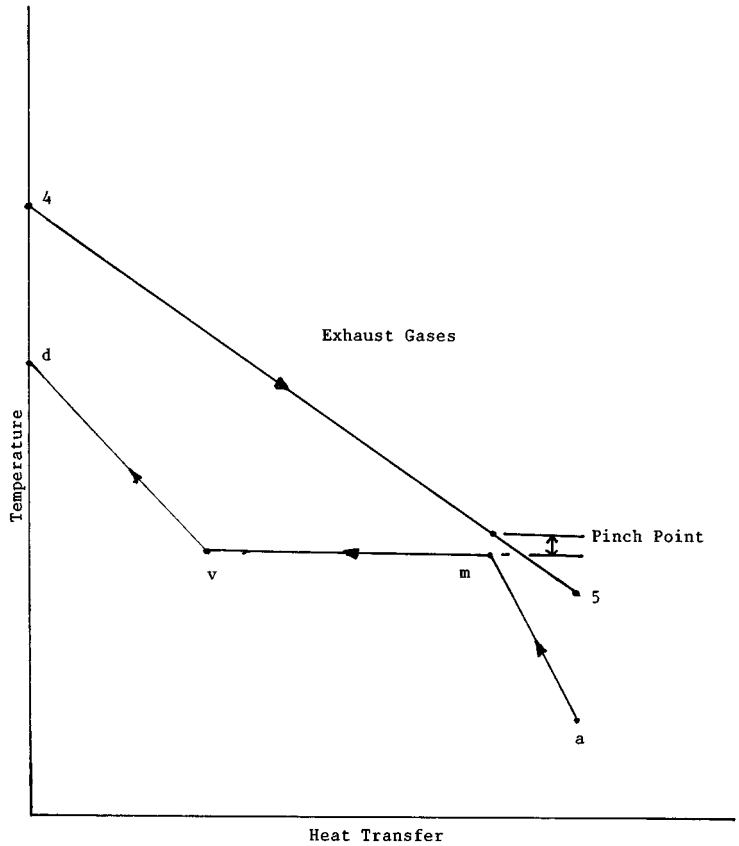


FIGURE 8.15.2 Typical temperature-heat transfer diagram for a single-pressure combined cycle power plant.

since the more energy transferred, the greater the amount of power produced by the steam turbine and the higher the combined cycle thermal efficiency or the lower the heat rate. In a conventional steam cycle power plant, the higher the feedwater temperature as it enters the steam generator (boiler), the higher the cycle thermal efficiency. Conventional steam cycle power plants achieve the higher feedwater temperature by the use of feedwater heaters.

Table 8.15.2 lists net plant output, heat rate, and net plant efficiency values for several combined cycles. One should observe for the units listed in Table 8.15.2 that

1. The size varies widely from a low of 22,800 kW to a high of 711,000 kW;
2. The heat rates vary from 7880 to 5885 Btu/kWhr. These compare with heat rate values for the gas turbines listed in Table 8.15.1 that vary from a high of 12,544 Btu/kWhr to a low of 8980 Btu/kWhr.
3. The net plant efficiency varies from a low of 42.8% to a high of 57.0%;
4. The fraction of the total net plant output from the steam turbine power plant in most cases ranges from 35 to 39%.

TABLE 8.15.2 Net Plant Output, Heat Rate, and Net Plant Efficiency for Several Combined Cycles

Manufacturer Model	Net Plant Output, kW	Heat Rate, Btu/kWhr	Net Plant Efficiency, %	Steam Turbine Power, kW
ABB				
KA 35-1	22,800	7,880	42.8	6,200
KA 8C-4	314,400	6,560	51.2	109,700
KA 11N-4	506,300	6,755	49.8	185,400
KA 24-2	501,800	5,885	57.0	182,800
GE Power Systems (60 Hz)				
S106B	59,200	7,020	48.6	22,600
S406B	240,000	6,930	49.3	93,600
S206FA	219,300	6,380	53.4	85,620
S207EC	357,700	6,415	53.2	134,400
S207FA	509,600	6,170	55.3	188,400
Siemens (KWU)				
GUD 3.64.3	275,000	6,550	52.1	98,000
GUD 3.84.2	486,000	6,560	52.0	186,000
GUD 3.94.2	711,000	6,525	52.3	269,000
GUD 2.84.3A	499,000	6,330	56.9	176,000

Source: 1995 Handbook, Gas Turbine World, Southport, CT, 1995. With permission.

One should keep in mind that only a few of the combined cycle plants currently available are listed in Table 8.15.2. The ones listed are intended to illustrate the range in net plant output and heat rate values and should not be interpreted as a list of all units currently available.

The amount of energy transferred in the HRSG is dependent on the steam pressure. This is illustrated in Figure 8.15.3. In this comparison, it is assumed in both cases that the steam exit temperature is the same and that the exhaust gas HRSG inlet temperature and flow rate are the same. One observes that the pressure of the water dictates the temperature at which evaporation occurs. The higher the evaporation temperature (and therefore steam pressure), the lower the mass flow rate of the steam, and therefore the lower the power generated by the steam power plant. It is obvious that the lower the steam pressure, the lower the flue gas temperature.

The output of an unfired single-pressure combined cycle power plant is determined by the gas turbine selected since this fixes the temperature and mass flow rate of the exhaust gases. One way to increase the steam cycle output is to introduce supplementary firing in the exhaust duct between the gas turbine exit and HRSG inlet. This is shown schematically in Figure 8.15.4. The advantages of adding supplementary firing are

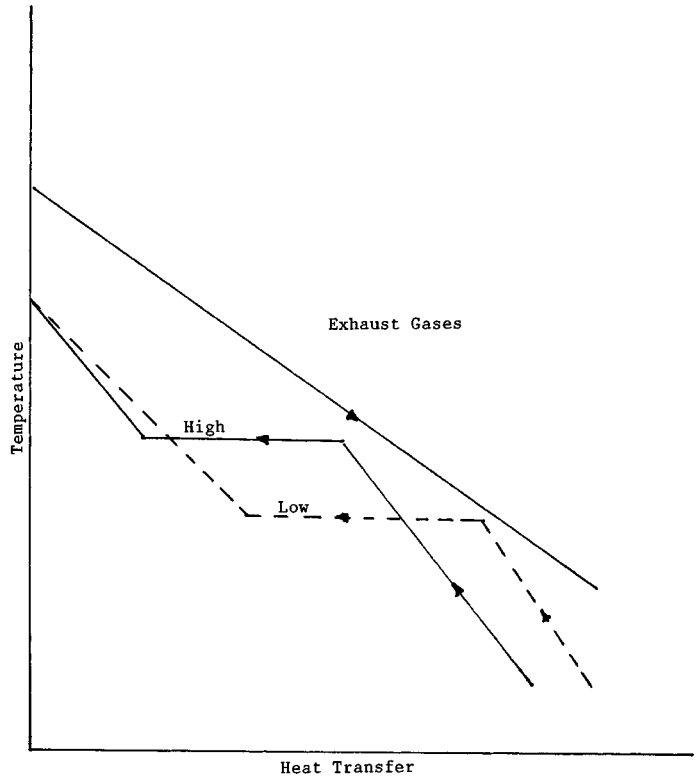


FIGURE 8.15.3 Effect of steam cycle pressure on energy transferred for a single-pressure combined cycle power plant.

1. The total output from the combined cycle will increase with a higher fraction of the output coming from the steam turbine cycle;
2. The temperature at the inlet to the HRSG can be controlled. This is important since the temperature and mass flow rate at the exit from the gas turbine are very dependent on the ambient temperature.

Figure 8.15.5 illustrates the effect on the temperature of the exhaust gases at the exit from the HRSG for unfired and supplementary fired combined cycles with the same pinch point temperature.

Combined cycle power plants with a single-pressure steam turbine do not maximize utilization of the energy in the exhaust gases. The ideal temperature-heat transfer diagram would be one where the temperature difference in the HRSG between the water (steam) and the exhaust gases is a constant.

One way to approach this constant temperature difference and improve the utilization of the energy in the exhaust gases is to use multipressures in the steam power plant cycle. Systems which have been used are

1. A dual pressure nonreheat cycle;
2. A dual-pressure reheat cycle; and
3. A triple-pressure reheat cycle.

Each system listed above has advantages and disadvantages. As new gas turbine engines enter the market with increased turbine inlet temperatures and component efficiencies, exhaust temperature from the gas turbine increases. This allows for higher superheated steam temperatures and improved combined cycle power plant efficiencies. This means that each design must be analyzed so that the selected design yields the most economical system.

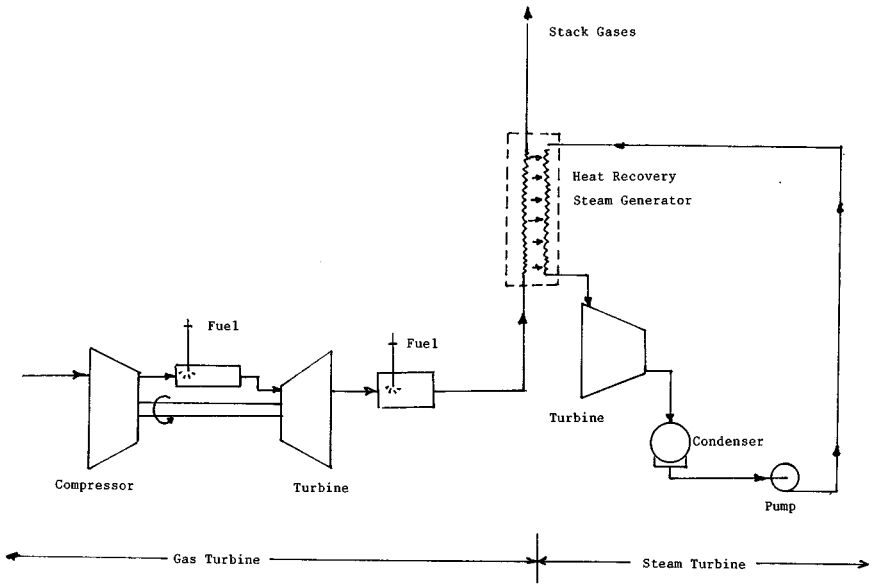


FIGURE 8.15.4 Combined cycle power plant with supplementary firing.

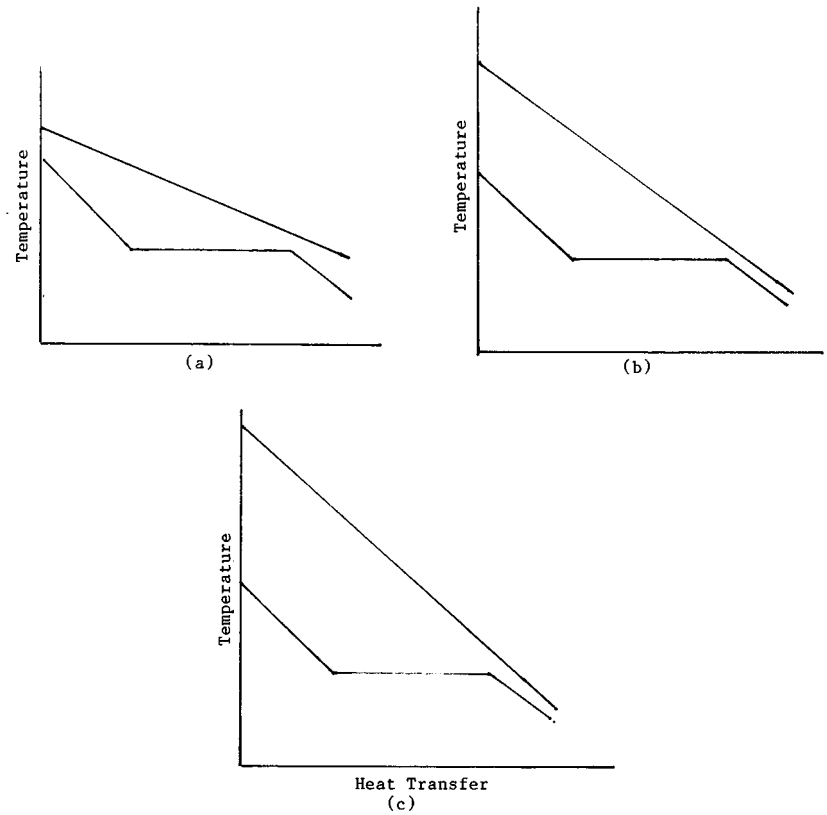


FIGURE 8.15.5 Effect of supplementary firing on exit temperature for same pinch point temperature difference. (a) No supplementary firing; (b), (c) supplementary firing.

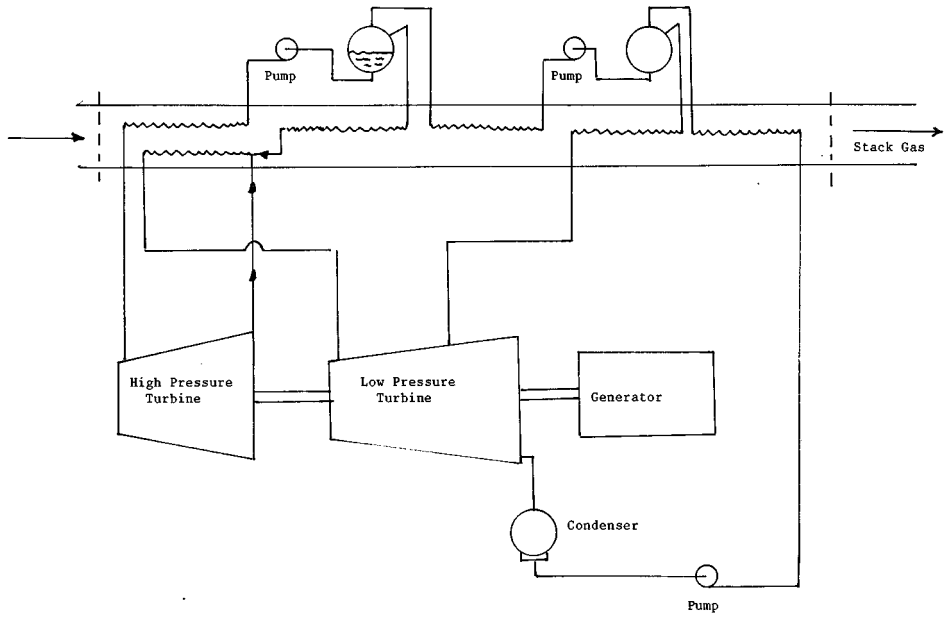


FIGURE 8.15.6 Schematic diagram for a three-pressure steam turbine combined cycle with superheating and reheat.

A simplified schematic diagram for a three-pressure combined cycle power plant is shown in Figure 8.15.6. This arrangement results in a high quality at the exit from the low-pressure steam turbine.

Reference

1995 Handbook, Gas Turbine World, Southport, CT, 1995.

Further Information

Bonzani, G. et al., Technical and economic optimization of a 450 MW combined cycle plant, in *1991 ASME Cogen-Turbo, 5th International Symposium and Exposition on Gas Turbines in Cogeneration, Repowering and Peak-Load Power Generation,* van der Linden, S. et al., Eds., pp. 131–143, ASME, New York, 1991.

Dechamps, P.J. et al., Advanced combined cycle alternatives with advanced gas turbines, in *ASME Cogen-Turbo Power '93, 7th Congress and Exposition on Gas Turbines in Cogeneration and Utility,* Holland, H.W. et al., Eds., pp. 387–396, ASME, New York, 1993.

Gyarmathy, G. and Ortmann, P., The off design of single- and dual-pressure steam cycles in CC plants, in *1991 ASME Cogen-Turbo, 5th International Symposium and Exposition on Gas Turbines in Cogeneration, Repowering and Peak-Load Power Generation,* van der Linden, S. et al., Eds., pp. 271–280, ASME, New York, 1991.

Horlock, J.H., *Combined Power Plants Including Combined Cycle Gas Turbine (CCGT) Plants,* Pergamon Press, New York, 1992.

Kehlhofer, W., *Combined-Cycle Gas & Steam Turbine Power Plants,* Fairmont Press, Englewood Cliffs, NJ, 1991.

Maurer, R., Destec's successes and plans for coal gasification combined cycle (CGCVC) power systems, in 1992 *ASME Cogen-Turbo, 6th International Conference in Cogeneration and Utility,* Cooke, D.H. et al., Eds., pp. 75–85, ASME, New York, 1992.

8.16 EMERGY Evaluation and Transformity

Howard T. Odum

EMERGY (spelled with an "m") evaluation is a method of energy analysis that puts all inputs and products on a common basis of what was previously required directly and indirectly to make each from one form of energy.

> EMERGY is the available energy of one kind previously used up directly and indirectly to make a service or product.

Its unit is the *emjoule,* defined to measure availability already used up. If average solar insolation at the earth's surface is chosen as the common base for the convenience of evaluating environmental and technological energy flows, then evaluations are made in units of solar emjoules, abbreviated sej.

In every energy transformation, available energy is used up to produce a smaller amount of energy of another kind. The EMERGY of one kind required directly and indirectly to make one unit of energy of another kind is defined as *transformity* (Odum, 1988).

> Solar transformity is the solar EMERGY of the inputs divided by the energy of the output.

$$\text{energy (J)} \times \text{transformity (sej/J)} = \text{EMERGY (sej)}$$

If all energy flows are expressed in solar EMERGY, then all kinds of energy may be compared according to their solar transformity. The more successive energy transformations there are, the higher the transformity. The higher the transformity, the more energy flows have converged in the process. Since many energy flows of one kind are usually required to support a smaller energy flow of another type, it is appropriate to describe the converging process as an *energy hierarchy.* The more transformations there are, the higher the transformity and the higher the position in a universal energy hierarchy. Figure 8.16.1 shows an environmental food chain with more energy flow but lower quality units (lower transformity) on the left and small total energy flow through high transformity units on the right. A land example is sunlight-grass-sheep-people. An aquatic series is sunlight-phytoplankton-zooplankton-small fish-large fish.

In environment, engineering, and economics the selective process of self organization generates higher quality energy capable of amplifying other processes. Thus observed transformitites are a measure of energy quality. Tables of solar transformity have been prepared based on previous EMERGY analysis (Odum, 1996). See sample in Table 8.16.1

Empower is the EMERGY used per unit time, for the United States about 8.5 E24 solar emjoules per year, a measure of the nation's processing of real wealth. By dividing the gross economic product ($6.5 trillion per year) the *EMERGY/money ratio* results (1.3 trillion solar emjoules per dollar). This is real wealth buying power of a dollar.

After an EMERGY evaluation is made, the solar EMERGY of each item can be divided by the *EMERGY*/money ratio to determine the *emdollars*, the buying power contributed by that item.

> *Emdollars are defined as the dollars of gross economic product due to that much real wealth measured as EMERGY.*

Because it puts all forms of available energy on a common basis, EMERGY may be the correct way to evaluate useful work. If real wealth comes from work, maximizing EMERGY production and use maximizes the real wealth of an economy. Selecting alternatives to maximize empower and emdollars is a new tool for engineering design and public policy.

Many energy analysis procedures add available energy as an exergy sum. However, EMERGY evaluation recognizes that each form of energy is different in its quality and the quantity of energy of other kinds required to make it. As shown in the EMERGY evaluation example (Table 8.16.2), each energy value is multiplied by its transformity so as to convert all items to solar EMERGY.

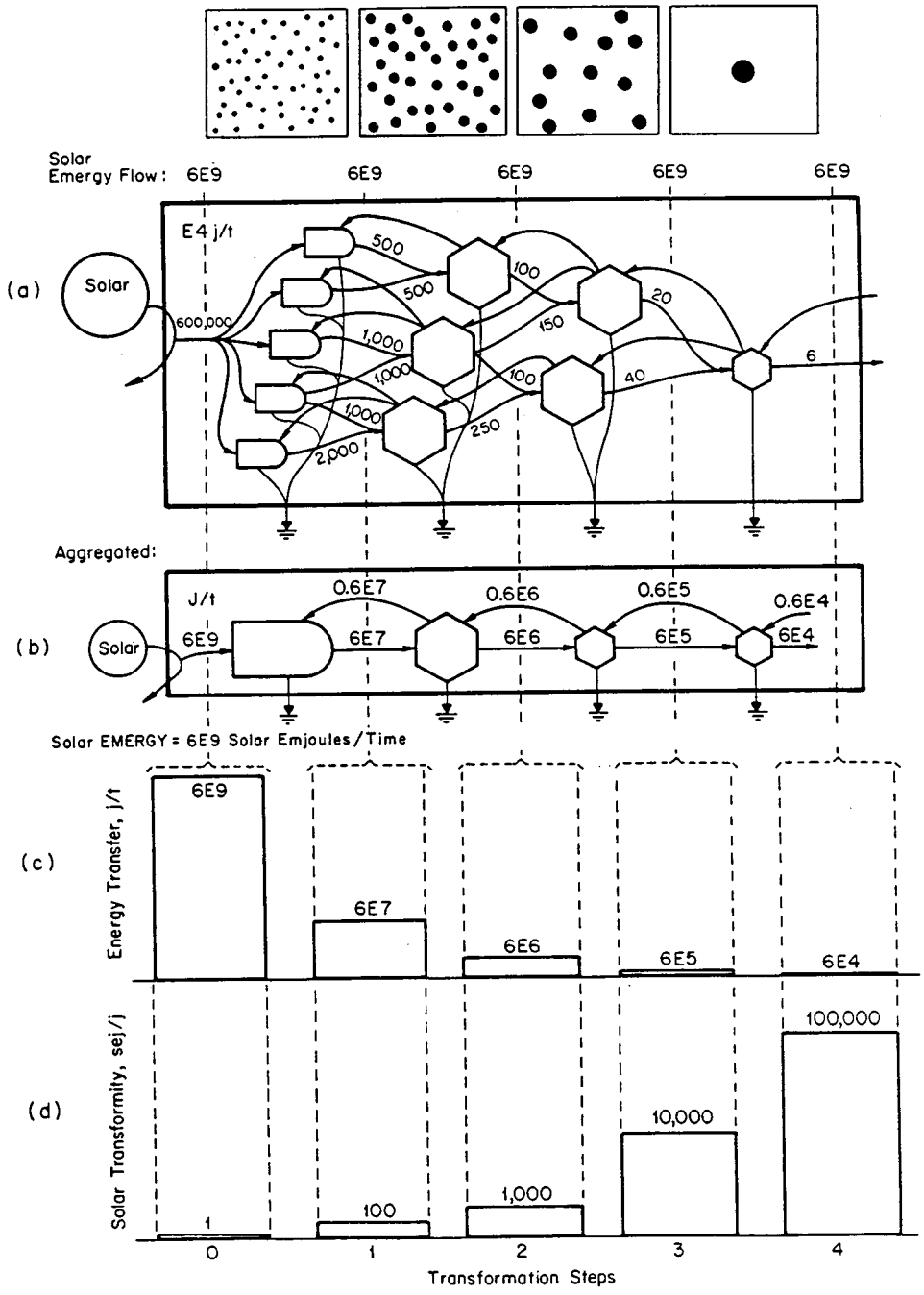


FIGURE 8.16.1 Energy transformation hierarchy: (a) spatial view of units, their sizes, and territories; (b) systems diagram of energy flow web; (c) web aggegated into an energy transformation chain; (d) energy flows between hierarchical levels; (e) transformitites of flows between levels.

TABLE 8.16.1 Typical Solar Transformities

Form of Energy	Solar Transformity sej/J
Solar insolation	1
Chemical energy in rain transpired over land#	18,000
Mined coal	40,000
Electric power	174,000
Human service, college graduate	343,000,000

sej/J = solar emjoules per Joule

Gibbs energy relative to sea water salt concentration in leaves, the main energy basis for land plants.

TABLE 8.16.2 EMERGY Evaluation of Lignite Processing in Texas*

Note	Item	Data Units/Day J, G, or $	Solar EMERGY/Unit sej/unit	Solar Empower E17 sej/day	Em$ 1995$ Thsd $/yr#
1	Diverted env. product.	7.10 E11 J	1.5 E4/J	0.11	8.5
2	Topsoil lost	5.04 E12 J	6.3 E4/J	3.18	244
3	Fuel used	6.38 E10 J	5.0 E4/J	0.032	2.5
4	Electricity used	3.11 E11 J	1.6 E5/J	0.50	38.5
5	Equipment support	13.8 E6 g	5.7 E9/g	0.79	60.8
6	Goods & service costs	2.8 E5 $	2.2 E12/$	6.2	477
				6.2	477
7	Sum of inputs			10.8	830
8	Lignite yield	2.0 E14 J	3.68 E4/J	73.6	5661

* Analysis of Big Brown Plant, Fairfield, Texas (Odum et al., 1987, revised).

Solar empower in Column 5 divided by 1.3 E12 sej/$ for U.S.A., 1995. Net EMERGY ratio = Yield/Inputs = 73.6/10.8 = 6.8

An Example of EMERGY Evaluation, Lignite

The evaluation procedure starts with an aggregated systems diagram to relate inputs, outputs, and main processes. This is used to identify line items in an evaluation table. An example is the analysis of lignite processing at the Big Brown Mine in Fairfield, Texas. Line items for the EMERGY evaluation in Table 8.16.2 were identified from the summarizing systems diagram in Figure 8.16.2. The net EMERGY ratio relates the EMERGY yield to the EMERGY required for the processing. The net EMERGY ratio of 6.8 (Table 8.16.2) means that 6.8 times more real wealth was contributed to the economy than required in the processing. The operation contributed 5.6 million emdollars per day to the economy as lignite yield, almost 12 times more than the economic value of $477,000 dollars per day.

Defining Terms

EMERGY: Available energy of one kind previously required directly and indirectly to make a product or service (unit: emjoules). Example: solar emjoules (abbreviation: sej).

Transformity: EMERGY per unit available energy (units: emjoules per joule). Example: solar emjoules per joule (abbreviation: sej/J).

Empower: EMERGY flow per unit time (units: emjoules per time). Example: solar emjoules per year (abbreviation: sej/yr).

Emdollars: EMERGY divided by EMERGY/money ratio. (abbreviation Em$).

Net EMERGY ratio: Yield EMERGY/EMERGY of purchased inputs.

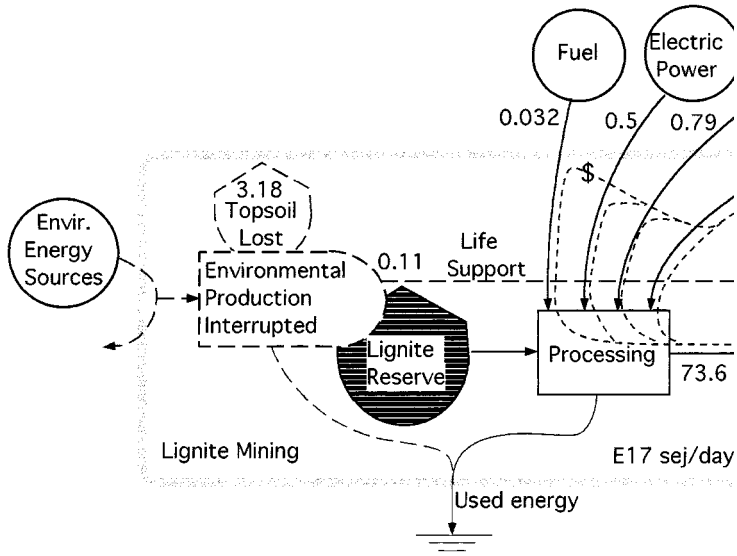


FIGURE 8.16.2 Energy systems diagram and emergy flows of a lignite mining operation evaluated in Table 8.16.2.

References

Odum, H.T., Odum, E.C., and Blissett, M. 1987. Ecology and Economy: "EMERGY" Analysis and Public Policy in Texas. Policy Research Project Report #78, Lyndon B. Johnson School of Public Affairs, The University of Texas, Austin, 178 pp.

Odum, H.T. 1988. Self organization, transformity, and information. *Science* 242, 1132–1139.

Odum, H.T. 1996. *Environmental Accounting, EMERGY and Decision Making*. John Wiley & Sons, New York, 276 pp.

Further Information

Hall, C.A.S., Ed. 1995. *Maximum Power.* University Press of Colorado, Niwot, CO, 393 pp.

Wang, S.K. and Lavan, Z. “Air-Conditioning and Refrigeration”
Mechanical Engineering Handbook
Ed. Frank Kreith
Boca Raton: CRC Press LLC, 1999

Air-Conditioning and Refrigeration

Shan K. Wang
Individual Consultant

Zalman Lavan
Professor Emeritus, Illinois Institute of Technology

9.1 Introduction

Air-Conditioning

Air-conditioning is a process that simultaneously conditions air; distributes it combined with the outdoor air to the conditioned space; and at the same time controls and maintains the required space's temperature, humidity, air movement, air cleanliness, sound level, and pressure differential within predetermined limits for the health and comfort of the occupants, for product processing, or both.

The acronym HVAC&R stands for heating, ventilating, air-conditioning, and refrigerating. The combination of these processes is equivalent to the functions performed by air-conditioning.

Because I-P units are widely used in the HVAC&R industry in the U.S., I-P units are used in this chapter. A table for converting I-P units to SI units is available in Appendix X of this handbook.

Air-Conditioning Systems

An *air-conditioning* or *HVAC&R system* consists of components and equipment arranged in sequential order to heat or cool, humidify or dehumidify, clean and purify, attenuate objectionable equipment noise, transport the conditioned outdoor air and recirculate air to the conditioned space, and control and maintain an indoor or enclosed environment at optimum energy use.

The types of buildings which the air-conditioning system serves can be classified as:

- Institutional buildings, such as hospitals and nursing homes
- Commercial buildings, such as offices, stores, and shopping centers

- Residential buildings, including single-family and multifamily low-rise buildings of three or fewer stories above grade
- Manufacturing buildings, which manufacture and store products

Types of Air-Conditioning Systems

In institutional, commercial, and residential buildings, air-conditioning systems are mainly for the occupants' health and comfort. They are often called *comfort air-conditioning systems*. In manufacturing buildings, air-conditioning systems are provided for product processing, or for the health and comfort of workers as well as processing, and are called *processing air-conditioning systems*.

Based on their size, construction, and operating characteristics, air-conditioning systems can be classified as the following.

Individual Room or Individual Systems. An individual air-conditioning system normally employs either a single, self-contained, packaged room air conditioner (installed in a window or through a wall) or separate indoor and outdoor units to serve an individual room, as shown in Figure 9.1.1. "Self-contained, packaged" means factory assembled in one package and ready for use.

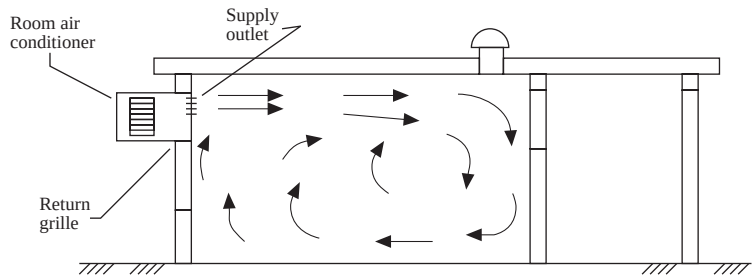


FIGURE 9.1.1 An individual room air-conditioning system.

Space-Conditioning Systems or Space Systems. These systems have their air-conditioning—cooling, heating, and filtration—performed predominantly in or above the conditioned space, as shown in Figure 9.1.2. Outdoor air is supplied by a separate outdoor ventilation system.

Unitary Packaged Systems or Packaged Systems. These systems are installed with either a single self-contained, factory-assembled packaged unit (PU) or two split units: an indoor air handler, normally with ductwork, and an outdoor condensing unit with refrigeration compressor(s) and condenser, as shown in Figure 9.1.3. In a packaged system, air is cooled mainly by direct expansion of refrigerant in coils called DX coils and heated by gas furnace, electric heating, or a heat pump effect, which is the reverse of a refrigeration cycle.

Central Hydronic or Central Systems. A central system uses chilled water or hot water from a central plant to cool and heat the air at the coils in an air handling unit (AHU) as shown in Figure 9.1.4. For energy transport, the heat capacity of water is about 3400 times greater than that of air. Central systems are built-up systems assembled and installed on the site.

Packaged systems are comprised of only air system, refrigeration, heating, and control systems. Both central and space-conditioning systems consist of the following.

Air Systems. An air system is also called an air handling system or the air side of an air-conditioning or HVAC&R system. Its function is to condition the air, distribute it, and control the indoor environment according to requirements. The primary equipment in an air system is an AHU or air handler; both of these include fan, coils, filters, dampers, humidifiers (optional), supply and return ductwork, supply outlets and return inlets, and controls.

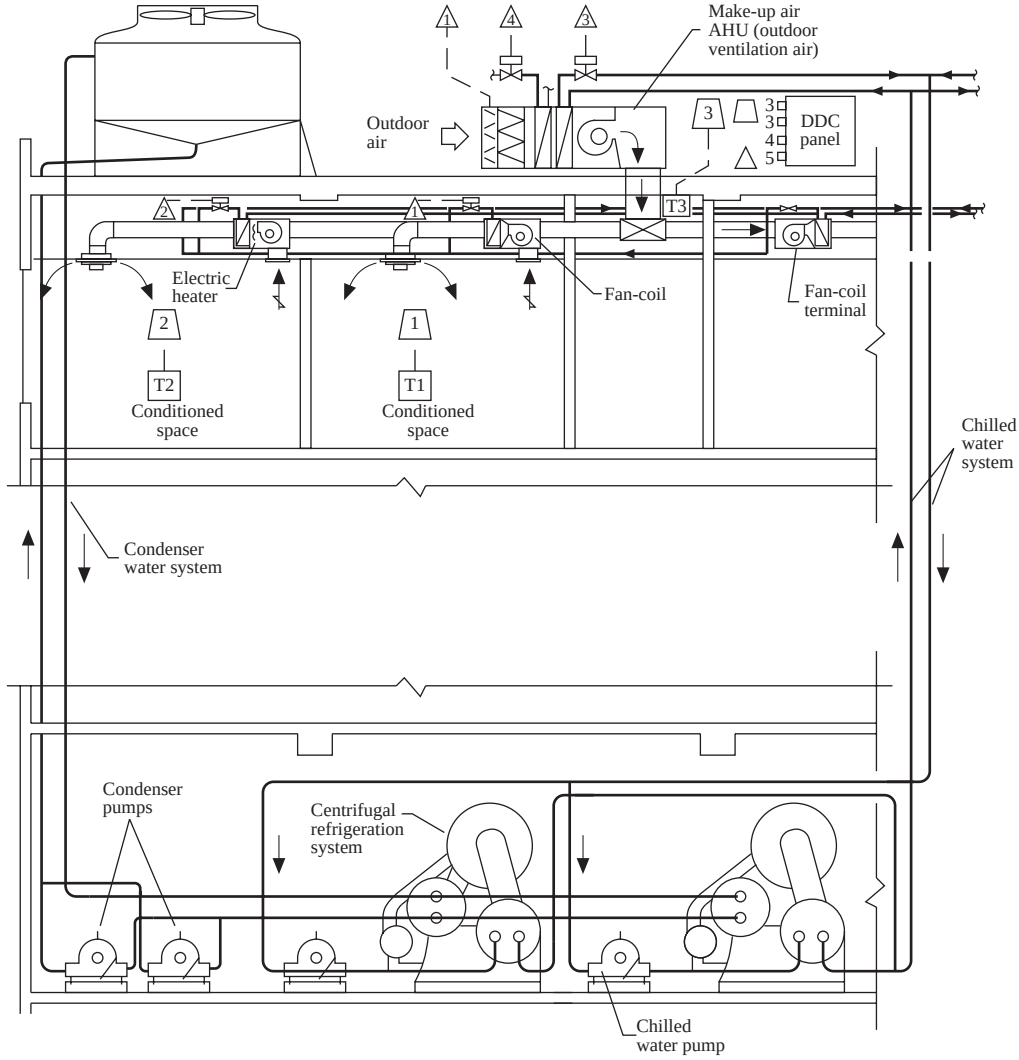


FIGURE 9.1.2 A space-conditioning air-conditioning system (fan-coil system).

Water Systems. These systems include chilled water, hot water, and condenser water systems. A water system consists of pumps, piping work, and accessories. The water system is sometimes called the water side of a central or space-conditioning system.

Central Plant Refrigeration and Heating Systems. The refrigeration system in the central plant of a central system is usually in the form of a chiller package with an outdoor condensing unit. The refrigeration system is also called the refrigeration side of a central system. A boiler and accessories make up the heating system in a central plant for a central system, and a direct-fired gas furnace is often the heating system in the air handler of a rooftop packaged system.

Control Systems. Control systems usually consist of sensors, a microprocessor-based direct digital controller (DDC), a control device, control elements, personal computer (PC), and communication network.

Based on Commercial Buildings Characteristics 1992, Energy Information Administration (EIA) of the Department of Energy of United States in 1992, for commercial buildings having a total floor area

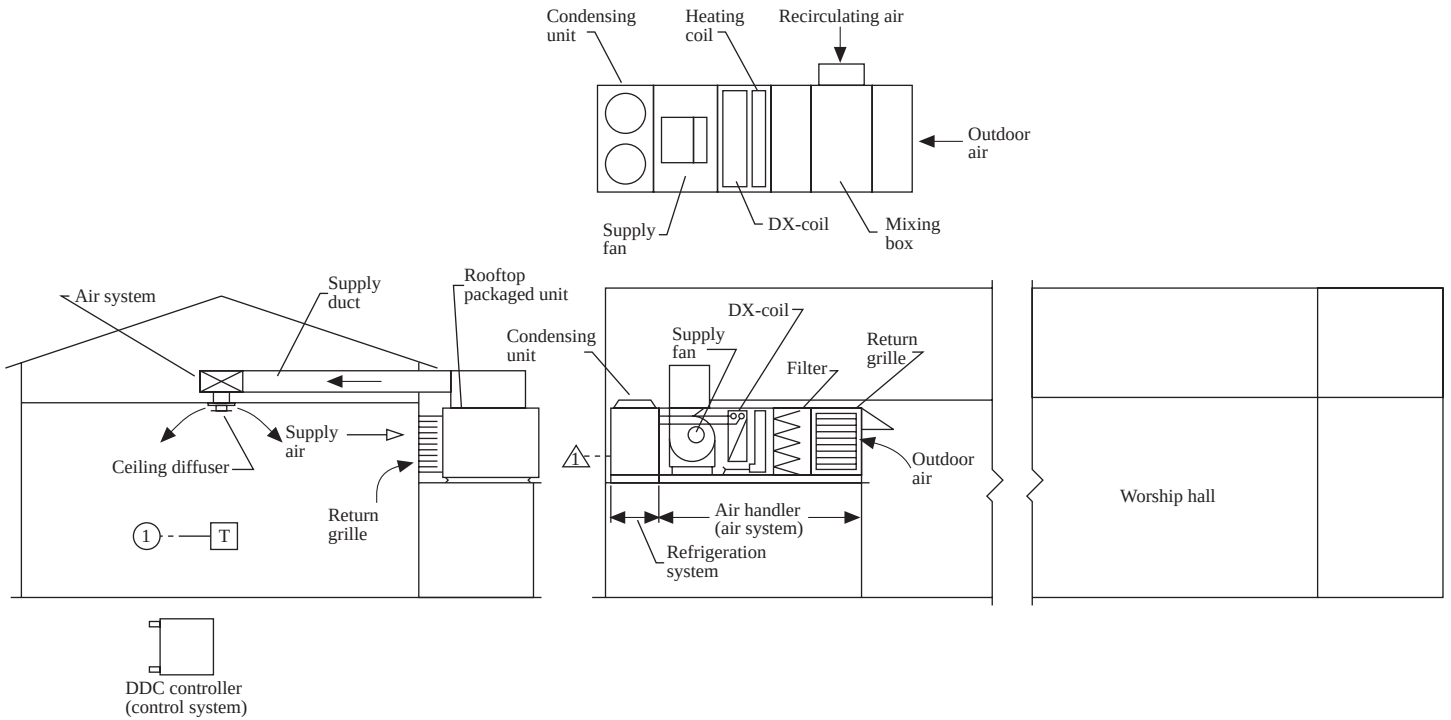


FIGURE 9.1.3 A packaged air-conditioning system.

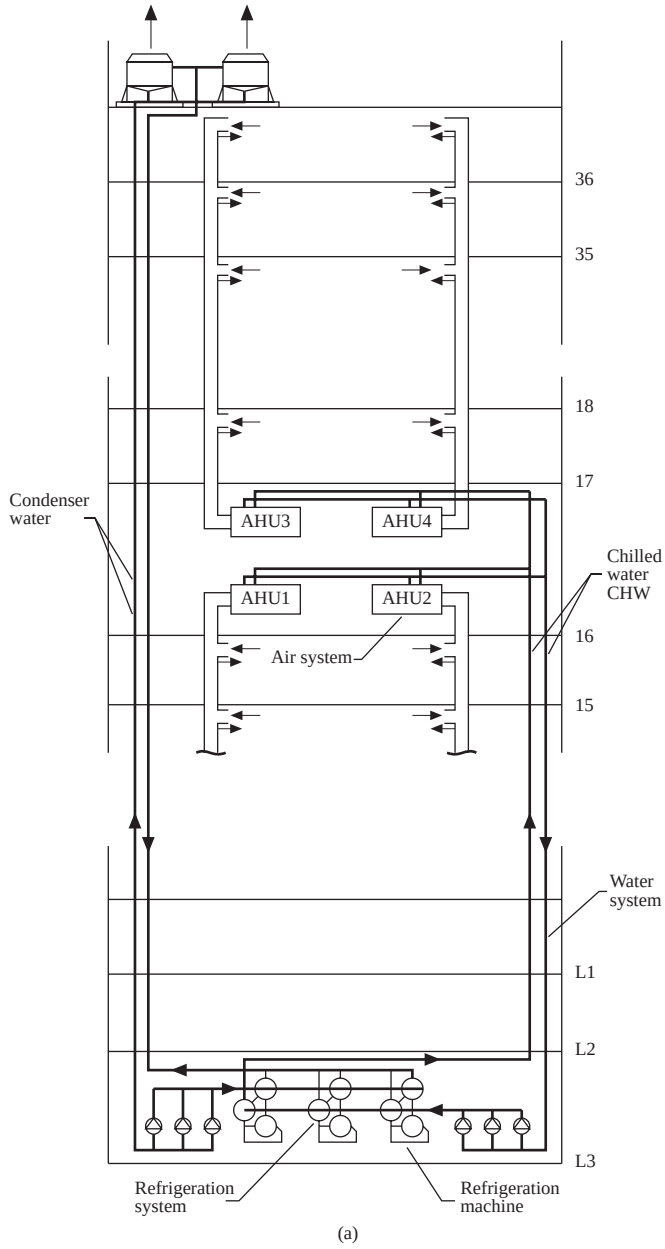


FIGURE 9.1.4a A central air-conditioning system: schematic diagram.

of 67,876 million ft², of which 57,041 million ft² or 84% is cooled and 61,996 million ft² or 91% is heated, the air-conditioning systems for cooling include:

Individual systems	19,239 million ft²	(25%)
Packaged systems	34,753 million ft²	(49%)
Central systems	14,048 million ft²	(26%)

Space-conditioning systems are included in central systems. Part of the cooled floor area has been counted for both individual and packaged systems. The sum of the floor areas for these three systems therefore exceeds the total cooled area of 57,041 million ft².

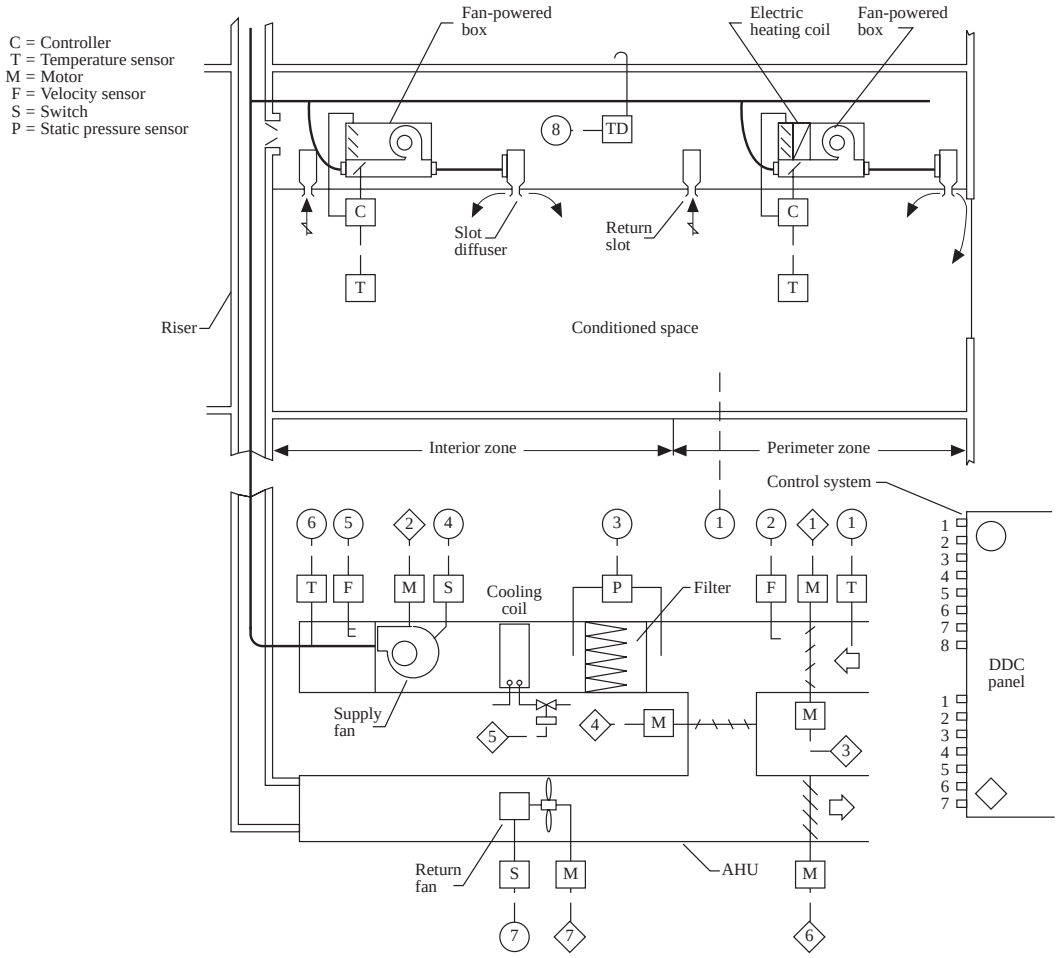


FIGURE 9.1.4b A central air-conditioning system: air and control systems for a typical floor.

Air-Conditioning Project Development and System Design

The goal of an air-conditioning/HVAC&R system is to provide a healthy and comfortable indoor environment with acceptable indoor air quality, while being energy efficient and cost effective.

ASHRAE Standard 62-1989 defines *acceptable indoor air quality* as "air in which there are no known contaminants at harmful concentrations as determined by cognizant authorities and with which a substantial majority (80% or more) of the people exposed do not express dissatisfaction."

The basic steps in the development and use of an air-conditioning project are design, installation, commissioning, operation, and maintenance. There are two types of air-conditioning projects: *design-bid* and *design-build*. A design-bid project separates the design (engineering consultant) and installation (contractors) responsibilities. In a design-build project, the design is also done by the installation contractor. A design-build project is usually a small project or a project having insufficient time to go through normal bidding procedures.

In the building construction industry, air-conditioning or HVAC&R is one of the *mechanical services*; these also include plumbing, fire protection, and escalators.

Air-conditioning design is a process of selecting the optimum system, subsystem, equipment, and components from various alternatives and preparing the drawings and specifications. Haines (1994) summarized this process in four phases: gather information, develop alternatives, evaluate alternatives,

and sell the best solution. Design determines the basic operating characteristics of a system. After an air-conditioning system is designed and constructed, it is difficult and expensive to change its basic characteristics.

The foundation of a successful project is teamwork and coordination between designer, contractor, and operator and between mechanical engineer, electrical engineer, facility operator, architect, and structural engineer.

Field experience is helpful to the designer. Before beginning the design process it is advisable to visit similar projects that have operated for more than 2 years and talk with the operator to investigate actual performance.

Mechanical Engineer's Responsibilities

The normal procedure in a design-bid construction project and the mechanical engineer's responsibilities are

1. Initiation of a project by owner or developer
2. Organizing a design team
3. Determining the design criteria and indoor environmental parameters
4. Calculation of cooling and heating loads
5. Selection of systems, subsystems, and their components
6. Preparation of schematic layouts; sizing of piping and ductwork
7. Preparation of contract documents: drawings and specifications
8. Competitive biddings by various contractors; evaluation of bids; negotiations and modifications
9. Advice on awarding of contract
10. Monitoring, supervision, and inspection of installation; reviewing shop drawings
11. Supervision of commissioning
12. Modification of drawings to the as-built condition; preparation of the operation and maintenance manual
13. Handing over to the property management for operation

Design Documents

Drawings and *specifications* are legal documents of a construction contract. The designer conveys the owner's or developer's requirements to the contractor through these documents. Drawings and specifications complement each other.

Drawings should clearly and completely show, define, and present the work. Adequate plan and sectional views should be drawn. More often, isometric drawings are used to show the flow diagrams for water or the supply, return, and exhaust air.

Specifications include the legal contract between the owner and the contractor, installer, or vendor and the technical specifications, which describe in detail what kind of material and equipment should be used and how they are to be installed.

Most projects now use a format developed by the Construction Specifications Institute (CSI) called the Masterformat for Specifications. It includes 16 divisions. The 15000 Mechanical division is divided into the following:

Section No.	Title
15050	Basic Mechanical Materials and Methods
15250	Mechanical Insulation
15300	Fire Protection
15400	Plumbing
15500	Heating, Ventilating, and Air-Conditioning
15550	Heat Generation
15650	Refrigeration
15750	Heat Transfer
15850	Air Handling

Section No.	Title
15880	Air Distribution
15950	Controls
15990	Testing, Adjusting, and Balancing

Each section includes general considerations, equipment and material, and field installation. Design criteria and selected indoor environmental parameters that indicate the performance of the HVAC&R system must be clearly specified in the general consideration of Section 15500.

There are two types of specifications: the performance specification, which depends mainly on the required performance criteria, and the or-equal specification, which specifies the wanted vendor. Specifications should be written in simple, direct, and clear language without repetition.

Computer-Aided Design and Drafting

With the wide acceptance of the PC and the availability of numerous types of engineering software, the use of *computer-aided drafting* (CAD) and *computer-aided design and drafting* (CADD) has increased greatly in recent years. According to the 1994 CADD Application and User Survey of design firms reported in *Engineering Systems* (1994[6]), "15% of the design firms now have a computer on every desk" and "Firms with high productivity reported that they perform 95% on CADD." Word processing software is widely used to prepare specifications.

Drafting software used to reproduce architectural drawings is the foundation of CADD. Automated CAD (AutoCAD) is the leading personal computer-based drafting tool software used in architectural and engineering design firms.

In "Software Review" by Amistadi (1993), duct design was the first HVAC&R application to be integrated with CAD.

- Carrier Corp. DuctLINK and Softdesk HVAC 12.0 are the two most widely used duct design software. Both of them convert the single-line duct layout drawn with CAD to two-dimensional (2D) double-line drawings with fittings, terminals, and diffusers.
- Tags and schedules of HVAC&R equipment, ductwork, and duct fittings can be produced as well.
- DuctLINK and Softdesk can also interface with architectural, electrical, and plumbing drawings through AutoCAD software.

Software for piping system design and analysis can also be integrated with CAD. The software developed at the University of Kentucky, KYCAD/KYPIPE, is intended for the design and diagnosis of large water piping systems, has extensive hydraulic modeling capacities, and is the most widely used. Softdesk AdCADD Piping is relative new software; it is intended for drafting in 2D and 3D, linking to AutoCAD through design information databases.

Currently, software for CADD for air-conditioning and HVAC&R falls into two categories: engineering and product. The engineering category includes CAD (AutoCAD integrated with duct and piping system), load calculations and energy analysis, etc. The most widely used software for load calculations and energy analysis is Department of Energy DOE-2.1D, Trane Company's TRACE 600, and Carrier Corporation's softwares for load calculation, E20-II Loads.

Product categories include selection, configuration, performance, price, and maintenance schedule. Product manufacturers provide software including data and CAD drawings for their specific product.

Codes and Standards

Codes are federal, state, or city laws that require the designer to perform the design without violating people's (including occupants and the public) safety and welfare. Federal and local codes must be followed. The designer should be thoroughly familiar with relevant codes. HVAC&R design codes are definitive concerning structural and electrical safety, fire prevention and protection (particularly for gas- or oil-fired systems), environmental concerns, indoor air quality, and energy conservation.

Conformance with *ASHRAE Standards* is voluntary. However, for design criteria or performance that has not been covered in the codes, whether the ASHRAE Standard is followed or violated is the vital criterion, as was the case in a recent indoor air quality lawsuit against a designer and contractor.

For the purpose of performing an effective, energy-efficient, safe, and cost-effective air-conditioning system design, the following ASHRAE Standards should be referred to from time to time:

- ASHRAE/IES Standard 90.1-1989, Energy Efficient Design of New Buildings Except New Low-Rise Residential Buildings
- ANSI/ASHRAE Standard 62-1989, Ventilation for Acceptable Indoor Air Quality
- ANSI/ASHRAE Standard 55-1992, Thermal Environmental Conditions for Human Occupancy
- ASHRAE Standard 15-1992, Safety Code for Mechanical Refrigeration

9.2 Psychrometrics

Moist Air

Above the surface of the earth is a layer of air called the *atmosphere*, or *atmospheric air*. The lower atmosphere, or homosphere, is composed of moist air, that is, a mixture of dry air and water vapor.

Psychrometrics is the science of studying the thermodynamic properties of moist air. It is widely used to illustrate and analyze the change in properties and the thermal characteristics of the air-conditioning process and cycles.

The composition of dry air varies slightly at different geographic locations and from time to time. The approximate composition of dry air by volume is nitrogen, 79.08%; oxygen, 20.95%; argon, 0.93%; carbon dioxide, 0.03%; other gases (e.g., neon, sulfur dioxide), 0.01%.

The amount of water vapor contained in the moist air within the temperature range 0 to 100°F changes from 0.05 to 3% by mass. The variation of water vapor has a critical influence on the characteristics of moist air.

The equation of state for an ideal gas that describes the relationship between its thermodynamic properties covered in Chapter 2 is

$$pv = RT_R \tag{9.2.1}$$

or

$$pV = mRT_R \tag{9.2.2}$$

where p = pressure of the gas, psf (1 psf = 144 psi)
v = specific volume of the gas, ft^3/lb
R = gas constant, $ftlb_f/lb_m$ °R
T_R = absolute temperature of the gas, °R
V = volume of the gas, ft^3
m = mass of the gas, lb

The most exact calculation of the thermodynamic properties of moist air is based on the formulations recommended by Hyland and Wexler (1983) of the U.S. National Bureau of Standards. The psychrometric charts and tables developed by ASHRAE are calculated and plotted from these formulations. According to Nelson et al. (1986), at a temperature between 0 and 100°F, enthalpy and specific volume calculations using ideal gas Equations (9.2.1) and (9.2.2) show a maximum deviation of 0.5% from the results of Hyland and Wexler's exact formulations. Therefore, ideal gas equations are used in the development and calculation of psychrometric formulations in this handbook.

Although air contaminants may seriously affect human health, they have little effect on the thermodynamic properties of moist air. For thermal analysis, moist air may be treated as a binary mixture of dry air and water vapor.

Applying Dalton's law to moist air:

$$p_{at} = p_a + p_w \tag{9.2.3}$$

where p_{at} = atmospheric pressure of the moist air, psia
p_a = partial pressure of dry air, psia
p_w = partial pressure of water vapor, psia

Dalton's law is summarized from the experimental results and is more accurate at low gas pressure. Dalton's law can also be extended, as the Gibbs-Dalton law, to describe the relationship of internal energy, enthalpy, and entropy of the gaseous constituents in a mixture.

Humidity and Enthalpy

The *humidity ratio* of moist air, w, in lb/lb is defined as the ratio of the mass of the water vapor, m_w to the mass of dry air, m_a, or

$$w = m_w / m_a = 0.62198 p_w / \left(p_{at} - p_w\right) \tag{9.2.4}$$

The *relative humidity* of moist air, φ, or RH, is defined as the ratio of the mole fraction of water vapor, x_w, to the mole fraction of saturated moist air at the same temperature and pressure, x_{ws}. Using the ideal gas equations, this relationship can be expressed as:

$$\varphi = x_w / x_{ws}\big|_{T,p} = p_w / p_{ws}\big|_{T,p} \tag{9.2.5}$$

and

$$x_w = n_w / \left(n_a + n_w\right); \quad x_{ws} = n_{ws} / \left(n_a + n_{ws}\right)$$

$$x_a + x_w = 1 \tag{9.2.6}$$

where p_{ws} = pressure of saturated water vapor, psia
T = temperature, °F
n_a, n_w, n_{ws} = number of moles of dry air, water vapor, and saturated water vapor, mol

Degree of saturation μ is defined as the ratio of the humidity ratio of moist air, w, to the humidity ratio of saturated moist air, w_s, at the same temperature and pressure:

$$\mu = w / w_s\big|_{T,p} \tag{9.2.7}$$

The difference between φ and μ is small, usually less than 2%.

At constant pressure, the difference in specific enthalpy of an ideal gas, in Btu/lb, is $\Delta h = c_p \Delta T$. Here c_p represents the specific heat at constant pressure, in Btu/lb. For simplicity, the following assumptions are made during the calculation of the *enthalpy* of moist air:

1. At 0°F, the enthalpy of dry air is equal to zero.
2. All water vapor is vaporized at 0°F.
3. The enthalpy of saturated water vapor at 0°F is 1061 Btu/lb.
4. The unit of the enthalpy of the moist air is Btu per pound of dry air and the associated water vapor, or Btu/lb.

Then, within the temperature range 0 to 100°F, the enthalpy of the moist air can be calculated as:

$$\begin{aligned} h &= c_{pd} T + w\left(h_{g0} + c_{ps} T\right) \\ &= 0.240 T + w(1061 + 0.444 T) \end{aligned} \tag{9.2.8}$$

where c_{pd}, c_{ps} = specific heat of dry air and water vapor at constant pressure, Btu/lb°F. Their mean values can be taken as 0.240 and 0.444 Btu/lb°F, respectively.
h_{g0} = specific enthalpy of saturated water vapor at 0°F.

Moist Volume, Density, Specific Heat, and Dew Point

The specific *moist volume v*, in ft^3/lb, is defined as the volume of the mixture of dry air and the associated water vapor when the mass of the dry air is exactly 1 lb:

$$v = V/m_a \tag{9.2.9}$$

where V = total volume of the moist air, ft^3. Since moist air, dry air, and water vapor occupy the same volume,

$$v = R_a T_R / p_{at}(1 + 1.6078w) \tag{9.2.10}$$

where R_a = gas constant for dry air.

Moist air density, often called *air density* ρ, in lb/ft^3, is defined as the ratio of the mass of dry air to the total volume of the mixture, or the reciprocal of the moist volume:

$$\rho = m_a/V = 1/v \tag{9.2.11}$$

The *sensible heat of moist air* is the thermal energy associated with the change of air temperature between two state points. In Equation (9.2.8), $(c_{pd} + wc_{ps})T$ indicates the sensible heat of moist air, which depends on its temperature T above the datum 0°F. *Latent heat of moist air*, often represented by wh_{fg0}, is the thermal energy associated with the change of state of water vapor. Both of them are in Btu/lb. Within the temperature range 0 to 100°F, if the average humidity ratio w is taken as 0.0075 lb/lb, the *specific heat of moist air* c_{pa} can be calculated as:

$$c_{pa} = c_{pd} + wc_{ps} = 0.240 + 0.0075 \times 0.444 = 0.243 \text{ Btu/lb °F} \tag{9.2.12}$$

The *dew point temperature* T_{dew}, in °F, is the temperature of saturated moist air of the moist air sample having the same humidity ratio at the same atmospheric pressure. Two moist air samples of similar dew points T_{dew} at the same atmospheric pressure have the same humidity ratio w and the same partial pressure of water vapor p_w.

Thermodynamic Wet Bulb Temperature and Wet Bulb Temperature

The *thermodynamic wet bulb temperature* of moist air, T^*, is equal to the saturated state of a moist air sample at the end of a constant-pressure, ideal adiabatic saturation process:

$$h_1 + \left(w_s^* - w_1\right)h_w^* = h_s^* \tag{9.2.13}$$

where h_1, h_s^* = enthalpy of moist air at the initial state and enthalpy of saturated air at the end of the constant-pressure, ideal adiabatic saturation process, Btu/lb

w_1, w_s^* = humidity ratio of moist air at the initial state and humidity ratio of saturated air at the end of the constant-pressure, ideal adiabatic saturation process, lb/lb

h_w^* = enthalpy of water added to the adiabatic saturation process at temperature T^*, Btu/lb

An *ideal adiabatic saturation process* is a hypothetical process in which moist air at initial temperature T_1, humidity ratio w_1, enthalpy h_1, and pressure p flows over a water surface of infinite length through a well-insulated channel. Liquid water is therefore evaporated into water vapor at the expense of the sensible heat of the moist air. The result is an increase of humidity ratio and a drop of temperature until the moist air is saturated at the thermodynamic wet bulb temperature T^* during the end of the ideal adiabatic saturation process.

The thermodynamic wet bulb temperature T^* is a unique fictitious property of moist air that depends only on its initial properties, T_1, w_1, or h_1.

A sling-type *psychrometer*, as shown in Figure 9.2.1, is an instrument that determines the temperature, relative humidity, and thus the state of the moist air by measuring its dry bulb and wet bulb temperatures. It consists of two mercury-in-glass thermometers. The sensing bulb of one of them is dry and is called the dry bulb. Another sensing bulb is wrapped with a piece of cotton wick, one end of which dips into a water tube. This wetted sensing bulb is called the wet bulb and the temperature measured by it is called the *wet bulb temperature* T'.

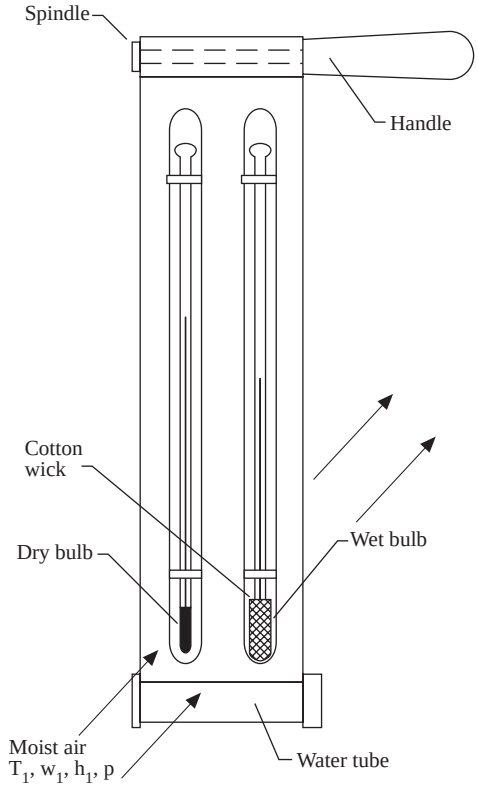


FIGURE 9.2.1 A sling psychrometer.

When unsaturated moist air flows over the surface of the wetted cotton wick, liquid water evaporates from its surface. As it absorbs sensible heat, mainly from the surrounding air, the wet bulb temperature drops. The difference between the dry and wet bulb temperatures is called *wet bulb depression* $(T - T')$. Turning the handle forces the surrounding air to flow over the dry and wet bulbs at an air velocity between 300 to 600 fpm. Distilled water must be used to wet the cotton wick.

At steady state, if heat conduction along the thermometer stems is neglected and the temperature of the wetted cotton wick is equal to the wet bulb temperature of the moist air, as the sensible heat transfer from the surrounding moist air to the cotton wick exactly equals the latent heat required for evaporation, the heat and mass transfer per unit area of the wet bulb surface can be evaluated as:

$$h_c(T - T') + h_r(T_{ra} - T) = h_d(w'_s - w_1) \tag{9.2.14}$$

where h_c, h_r = mean conductive and radiative heat transfer coefficient, Btu/hr ft²°F
h_d = mean convective mass transfer coefficient, lb/hr ft²

T = temperature of undisturbed moist air at a distance from the wet bulb, °F
T_{ra} = mean radiant temperature (covered later), °F
w_1, w'_s = humidity ratio of the moist air and the saturated film at the interface of cotton wick and surrounding air, lb/lb
h'_{fg} = latent heat of vaporization at the wet bulb temperature, Btu/lb

The humidity ratio of the moist air is given by:

$$w_1 = w'_s - K'(T - T')\left(1 + \left\{h_r(T_{ra} - T')/\left[h_c(T - T')\right]\right\}\right)$$

$$K' = c_{pa}\,\mathrm{Le}^{0.6667}/h'_{fg} \tag{9.2.15}$$

where K' = wet bulb constant, which for a sling psychrometer = 0.00218 1/°F
Le = Lewis number

The wet bulb temperature T' depends not only on its initial state but also on the rate of heat and mass transfer at the wet bulb. Therefore, the thermodynamic wet bulb temperature is used in ASHRAE psychrometric charts.

According to Threlkeld (1970), for a sling psychrometer whose wet bulb diameter is 1 in. and for air flowing at a velocity of 400 fpm over the wet bulb, if the dry bulb temperature is 90°F and the measured wet bulb temperature is 70°F, the difference between the measured wet bulb and the thermodynamic wet bulb $(T' - T^*)/(T^* - T')$ is less than 1%.

Psychrometric Charts

A *psychrometric chart* is a graphical presentation of the thermodynamic properties of moist air and various air-conditioning processes and air-conditioning cycles. A psychrometric chart also helps in calculating and analyzing the work and energy transfer of various air-conditioning processes and cycles.

Psychrometric charts currently use two kinds of basic coordinates:

1. *h*-*w* charts. In *h*-*w* charts, enthalpy *h*, representing energy, and humidity ratio *w*, representing mass, are the basic coordinates. Psychrometric charts published by ASHRAE and the Charted Institution of Building Services Engineering (CIBSE) are *h*-*w* charts.
2. *T*-*w* charts. In *T*-*w* charts, temperature *T* and humidity ratio *w* are basic coordinates. Psychrometric charts published by Carrier Corporation, the Trane Company, etc. are *T*-*w* charts.

Figure 9.2.2 shows an abridged ASHRAE psychrometric chart. In the ASHRAE chart:

- A normal temperature chart has a temperature range of 32 to 120°F, a high-temperature chart 60 to 250°F, and a low-temperature chart –40 to 50°F. Since enthalpy is the basic coordinate, temperature lines are not parallel to each other. Only the 120°F line is truly vertical.
- Thermodynamic properties of moist air are affected by atmospheric pressure. The standard atmospheric pressure is 29.92 in. Hg at sea level. ASHRAE also published charts for high altitudes of 5000 ft, 24.89 in. Hg, and 7500 ft, 22.65 in. Hg. Both of them are in the normal temperature range.
- Enthalpy *h*-lines incline downward to the right-hand side (negative slope) at an angle of 23.5° to the horizontal line and have a range of 12 to 54 Btu/lb.
- Humidity ratio *w*-lines are horizontal lines. They range from 0 to 0.28 lb/lb.
- Relative humidity φ-lines are curves of relative humidity 10%, 20%, ... 90% and a saturation curve. A saturation curve is a curve of the locus of state points of saturated moist air, that is, φ = 100%. On a saturation curve, temperature *T*, thermodynamic wet temperature bulb T^*, and dew point temperature T_{dew} have the same value.

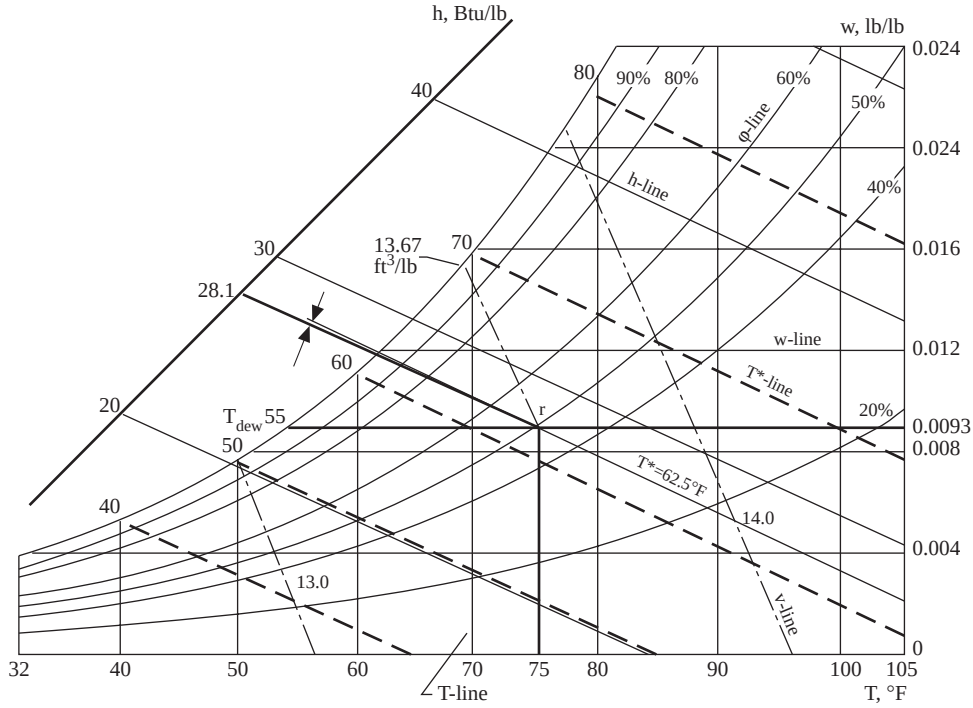


FIGURE 9.2.2 The abridged ASHRAE psychrometric chart and the determination of properties as in Example 9.2.1.

- Thermodynamic wet bulb T^*-lines have a negative slope slightly greater than that of the h-lines. A T^*-line meets the T-line of the same magnitude on the saturation curve.
- Moist volume v-lines have a far greater negative slope than h-lines and T^*-lines. The moist volume ranges from 12.5 to 15 ft^3/lb.

Moist air has seven independent thermodynamic properties or property groups: h, T, φ, T^*, p_{at}, $\rho - v$, and $w - p_w - T_{dew}$. When p_{at} is given, any additional two of the independent properties determine the state of moist air on the psychrometric chart and the remaining properties.

Software using AutoCAD to construct the psychrometric chart and calculate the thermodynamic properties of moist air is available. It can also be linked to the load calculation and energy programs to analyze the characteristics of air-conditioning cycles.

Refer to Wang's *Handbook of Air Conditioning and Refrigeration* (1993) and *ASHRAE Handbook, Fundamentals* (1993) for details of psychrometric charts and psychrometric tables that list thermodynamic properties of moist air.

Example 9.2.1

An air-conditioned room at sea level has an indoor design temperature of 75°F and a relative humidity of 50%. Determine the humidity ratio, enthalpy, density, dew point, and thermodynamic wet bulb temperature of the indoor air at design condition.

Solution

1. Since the air-conditioned room is at sea level, a psychrometric chart of standard atmospheric pressure of 14.697 psi should be used to find the required properties.
2. Plot the state point of the room air at design condition r on the psychrometric chart. First, find the room temperature 75°F on the horizontal temperature scale. Draw a line parallel to the 75°F

temperature line. This line meets the relative humidity curve of 50% at point *r*, which denotes the state point of room air as shown in Figure 9.2.2.

3. Draw a horizontal line toward the humidity ratio scale from point *r*. This line meets the ordinate and thus determines the room air humidity ratio $\varphi_r = 0.0093$ lb/lb.
4. Draw a line from point *r* parallel to the enthalpy line. This line determines the enthalpy of room air on the enthalpy scale, $h_r = 28.1$ Btu/lb.
5. Draw a line through point *r* parallel to the moist volume line. The perpendicular scale of this line indicates $v_r = 13.67$ ft^3/lb.
6. Draw a horizontal line to the left from point *r*. This line meets the saturation curve and determines the dew point temperature, $T_{dew} = 55$°F.
7. Draw a line through point *r* parallel to the thermodynamic wet bulb line. The perpendicular scale to this line indicates that the thermodynamic wet bulb temperature $T^* = 62.5$°F.

9.3 Air-Conditioning Processes and Cycles

Air-Conditioning Processes

An *air-conditioning process* describes the change in thermodynamic properties of moist air between the initial and final stages of conditioning as well as the corresponding energy and mass transfers between the moist air and a medium, such as water, refrigerant, absorbent or adsorbent, or moist air itself. The energy balance and conservation of mass are the two principles used for the analysis and the calculation of the thermodynamic properties of the moist air.

Generally, for a single air-conditioning process, heat transfer or mass transfer is positive. However, for calculations that involve several air-conditioning processes, heat supplied to the moist air is taken as positive and heat rejected is negative.

The *sensible heat ratio* (SHR) of an air-conditioning process is defined as the ratio of the change in absolute value of sensible heat to the change in absolute value of total heat, both in Btu/hr:

$$\text{SHR} = \left|q_{\text{sen}}\right| \Big/ \left|q_{\text{total}}\right| = \left|q_{\text{sen}}\right| \Big/ \left(\left|q_{\text{sen}}\right| + \left|q_1\right|\right) \tag{9.3.1}$$

For any air-conditioning process, the sensible heat change

$$q_{\text{sen}} = 60\,\mathring{V}_s\,\rho_s c_{\text{pa}}\left(T_2 - T_1\right) = 60\,\mathring{m}_a\,c_{\text{pa}}\left(T_2 - T_1\right) \tag{9.3.2}$$

where $\mathring{V}_s$ = volume flow rate of supply air, cfm
ρ_s = density of supply air, lb/ft^3
T_2, T_1 = moist air temperature at final and initial states of an air-conditioning process, °F

and the mass flow rate of supply air

$$\mathring{m}_s = \mathring{V}_s\,\rho_s \tag{9.3.3}$$

The latent heat change is

$$q_1 \approx 60\,\mathring{V}_s\,\rho_s\left(w_2 - w_1\right)h_{\text{fg.58}} = 1060 \times 60\,\mathring{V}_s\,\rho_s\left(w_2 - w_1\right) \tag{9.3.4}$$

where w_2,w_1 = humidity ratio at final and initial states of an air-conditioning process, lb/lb.

In Equation (9.3.4), $h_{\text{fg58}} \approx 1060$ Btu/lb represents the latent heat of vaporization or condensation of water at an estimated temperature of 58°F, where vaporization or condensation occurs in an air-handling unit or packaged unit. Therefore

$$\text{SHR} = \mathring{m}_a\,c_{\text{pa}}\left(T_2 - T_1\right) \Big/ \left[\mathring{m}_a\,c_{\text{pa}}\left(T_2 - T_1\right) + \mathring{m}_a\left(w_2 - w_1\right)h_{\text{fg.58}}\right] \tag{9.3.5}$$

Space Conditioning, Sensible Cooling, and Sensible Heating Processes

In a *space conditioning process*, heat and moisture are absorbed by the supply air at state *s* and then removed from the conditioned space at the state of space air *r* during summer, as shown by line *sr* in Figure 9.3.1, or heat or moisture is supplied to the space to compensate for the transmission and infiltration losses through the building envelope as shown by line *s′r′*. Both processes are aimed at maintaining a desirable space temperature and relative humidity.

The space cooling load q_{rc}, in Btu/hr, can be calculated as:

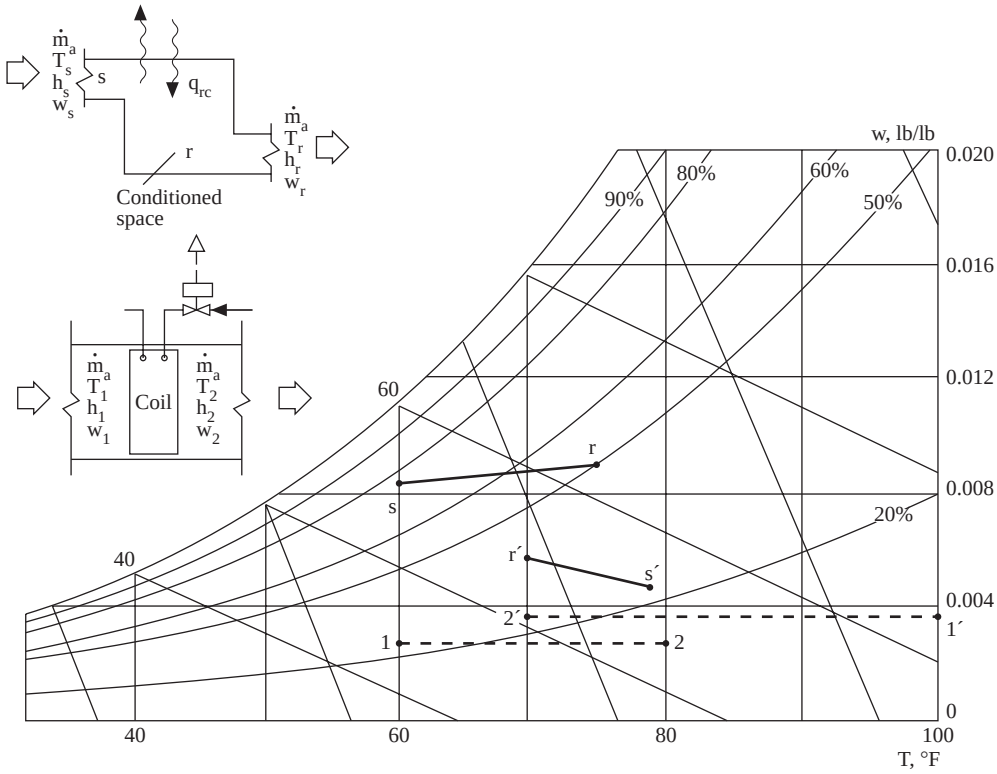


FIGURE 9.3.1 Supply conditioning, sensible heating, and sensible cooling processes.

$$q_{rc} = 60\dot{m}_a\left(h_r - h_s\right) = 60\mathring{V}_s\rho_s\left(h_r - h_s\right) \tag{9.3.6}$$

where h_r, h_s = enthalpy of space air and supply air, Btu/lb.

The space sensible cooling load q_{rs}, in Btu/hr, can be calculated from Equation (9.3.2) and the space latent load q_{rl}, in Btu/hr, from Equation (9.3.1). In Equation (9.3.4), T_2 should be replaced by T_r and T_1 by T_s. Also in Equation (9.3.1), w_2 should be replaced by w_r and w_1 by w_s. The space heating load q_{rh} is always a sensible load, in Btu/hr, and can be calculated as:

$$q_{rh} = 60\mathring{m}_a c_{pa}\left(T_s - T_r\right) = 60\mathring{V}_s\rho_s c_{pa}\left(T_s - T_r\right) \tag{9.3.7}$$

where T_s, T_r = temperature of supply and space air, °F.

A *sensible heating process* adds heat to the moist air in order to increase its temperature; its humidity ratio remains constant, as shown by line 12 in Figure 9.3.1. A sensible heating process occurs when moist air flows over a heating coil. Heat is transferred from the hot water inside the tubes to the moist air. The rate of heat transfer from the hot water to the colder moist air is often called the heating coil load q_{rh}, in Btu/hr, and is calculated from Equation (9.3.2).

A *sensible cooling process* removes heat from the moist air, resulting in a drop of its temperature; its humidity ratio remains constant, as shown by line 1′2′ in Figure 9.3.1. The sensible cooling process occurs when moist air flows through a cooling coil containing chilled water at a temperature equal to or greater than the dew point of the entering moist air. The sensible cooling load can also be calculated from Equation (9.3.2). T_2 is replaced by T_1 and T_1 by $T_{2'}$.

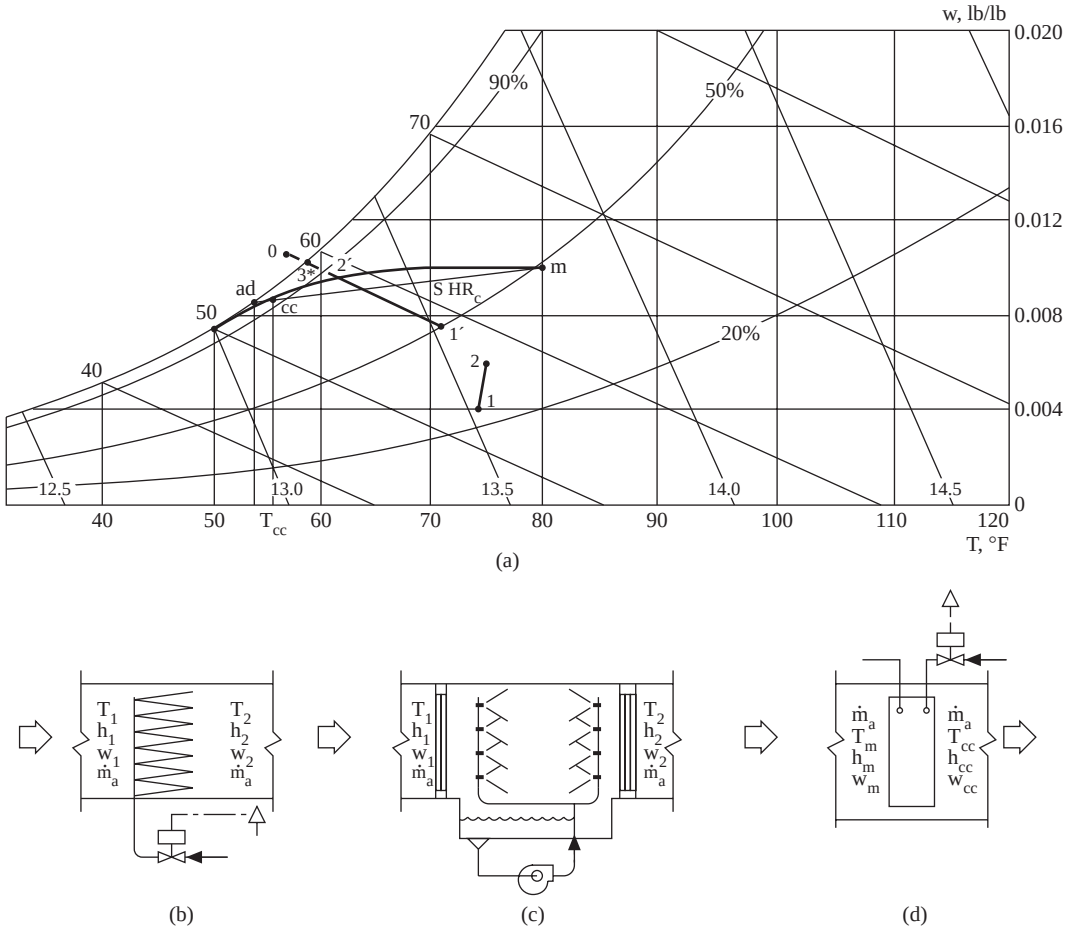


FIGURE 9.3.2 Humidifying and cooling and dehumidifying processes: (a) process on psychrometric chart, (b) steam humidifier, (c) air washer, and (d) water cooling or DX coil.

Humidifying and Cooling and Dehumidifying Processes

In a *humidifying process*, water vapor is added to moist air and increases the humidity ratio of the moist air entering the humidifier if the moist air is not saturated. Large-scale humidification of moist air is usually performed by steam injection, evaporation from a water spray, atomizing water, a wetted medium, or submerged heating elements. Some details of their construction and characteristics are covered in a later section. Dry steam in a steam injection humidifying process is often supplied from the main steam line to a grid-type humidifier and injected into the moist air directly through small holes at a pressure slightly above atmospheric, as shown by line 12 in Figure 9.3.2(a) and (b). The humidifying capacity $\mathring{m}_{hu}$, in lb/min, is given by:

$$\mathring{m}_{hu} = \mathring{V}_s \rho_s \left(w_{hl} - w_{he}\right) \tag{9.3.8}$$

where w_{hl}, w_{he} = humidity ratio of moist air leaving and entering the humidifier, lb/lb. The slight inclination at the top of line 12 is due to the high temperature of the steam. The increase in temperature of the moist air due to steam injection can be calculated as:

$$\left(T_2 - T_1\right) = w_{sm} c_{ps} T_s \Big/ \left(c_{pd} + w_{12} c_{ps}\right) \tag{9.3.9}$$

where T_2, T_1 = temperature of moist air at initial and final states, °F
w_{sm} = ratio of mass flow rate of injected steam to moist air, $\mathring{m}_s / \mathring{m}_a$
T_s = temperature of injected steam, °F
w_{12} = average humidity ratio of moist air, lb/lb

An *air washer* is a device that sprays water into moist air in order to humidify, to cool and dehumidify, and to clean the air, as shown in Figure 9.3.2(c). When moist air flows through an air washer, the moist air is humidified and approaches saturation. This actual adiabatic saturation process approximately follows the thermodynamic wet bulb line on the psychrometric chart as shown by line 1′2′. The humidity ratio of the moist air is increased while its temperature is reduced. The cooling effect of this adiabatic saturation process is called *evaporative cooling*.

Oversaturation occurs when the amount of water present in the moist air w_{os}, in lb/lb, exceeds the saturated humidity ratio at thermodynamic wet bulb temperature w_s^*, as shown in Figure 9.3.2(a). When moist air leaves the air washer, atomizing humidifier, or centrifugal humidifier after humidification, it often contains unevaporated water droplets at state point 2′, w_w, in lb/lb. Because of the fan power heat gain, duct heat gain, and other heat gains providing the latent heat of vaporization, some evaporation takes place due to the heat transfer to the water drops, and the humidity ratio increases further. Such evaporation of oversaturated drops is often a process with an increase of both humidity ratio and enthalpy of moist air. Oversaturation can be expressed as:

$$w_{os} = w_o - w_s^* = \left(w_{2'} + w_w\right) - w_s^* \tag{9.3.10}$$

where $w_{2'}$ = humidity ratio at state point 2′, lb/lb
w_o = sum of $w_{2'}$ and w_w, lb/lb

The magnitude of w_w depends mainly on the construction of the humidifier and water eliminator, if any. For an air washer, w_w may vary from 0.0002 to 0.001 lb/lb. For a pulverizing fan without an eliminator, w_w may be up to 0.00135 lb/lb.

Cooling and Dehumidifying Process

In a cooling and dehumidifying process, both the humidity ratio and temperature of moist air decrease. Some water vapor is condensed in the form of liquid water, called a *condensate*. This process is shown by curve m cc on the psychrometric chart in Figure 9.3.2(a). Here m represents the entering mixture of outdoor and recirculating air and cc the conditioned air leaving the cooling coil.

Three types of heat exchangers are used in a cooling and dehumidifying process: (1) water cooling coil as shown in Figure 9.3.2(d); (2) direct expansion DX coil, where refrigerant evaporates directly inside the coil's tubes; and (3) air washer, in which chilled water spraying contacts condition air directly.

The temperature of chilled water entering the cooling coil or air washer T_{we}, in °F, determines whether it is a sensible cooling or a cooling and dehumidifying process. If T_{we} is smaller than the dew point of the entering air T_{ae}'' in the air washer, or T_{we} makes the outer surface of the water cooling coil $T_{s.t} < T_{ae}''$, it is a cooling and dehumidifying process. If $T_{we} \geq T_{ae}''$, or $T_{s.t} \geq T_{ae}''$, sensible cooling occurs. The cooling coil's load or the cooling capacity of the air washer q_{cc}, in Btu/hr, is

$$q_{cc} = 60 \mathring{V}_s \rho_s \left(h_{ae} - h_{cc}\right) - 60 \mathring{m}_w h_w \tag{9.3.11a}$$

where h_{ae}, h_{cc} = enthalpy of moist air entering and leaving the coil or washer, Btu/lb
$\mathring{m}_w$ = mass flow rate of the condensate, lb/min
h_w = enthalpy of the condensate, Btu/lb

Since the thermal energy of the condensate is small compared with q_{cc}, in practical calculations the term $60\dot{m}_w h_w$ is often neglected, and

$$q_{cc} = 60\dot{V}_s \rho_s (h_{ae} - h_{cc}) \tag{9.3.11b}$$

The sensible heat ratio of the cooling and dehumidifying process SHR_c can be calculated from

$$SHR_c = q_{cs}/q_{cc} \tag{9.3.12}$$

where q_{cs} = sensible heat removed during the cooling and dehumidifying process, Btu/hr. SHR_c is shown by the slope of the straight line joining points m and cc.

The relative humidity of moist air leaving the water cooling coil or DX coil depends mainly on the outer surface area of the coil including pipe and fins. For coils with ten or more fins per inch, if the entering moist air is around 80°F dry bulb and 68°F wet bulb, the relative humidity of air leaving the coil (off-coil) may be estimated as:

Four-row coil	90 to 95%
Six-row and eight-row coils	96 to 98%

Two-Stream Mixing Process and Bypass Mixing Process

For a *two-stream adiabatic mixing process*, two moist air streams, 1 and 2, are mixed together adiabatically and a mixture *m* is formed in a mixing chamber as shown by line 1 m1 2 in Figure 9.3.3. Since the AHU or PU is well insulated, the heat transfer between the mixing chamber and ambient air is small and is usually neglected. Based on the principle of heat balance and conservation of mass:

$$\begin{aligned} \dot{m}_1 h_1 + \dot{m}_2 h_2 &= \dot{m}_m h_m \\ \dot{m}_1 w_1 + \dot{m}_2 w_2 &= \dot{m}_m w_m \\ \dot{m}_1 T_1 + \dot{m}_2 T_2 &= \dot{m}_m T_m \end{aligned} \tag{9.3.13}$$

In Equation (9.3.13), $\dot{m}$ represents the mass flow rate of air, lb/min; *h* the enthalpy, in Btu/lb; w the humidity ratio, in lb/lb; and *T* the temperature, in °F. Subscripts 1 and 2 indicate air streams 1 and 2 and m the mixture; also,

$$\begin{aligned} \dot{m}_1/\dot{m}_m &= (h_2 - h_m)/(h_2 - h_1) = (w_2 - w_m)/(w_2 - w_1) \\ &= (\text{line segment m1 2})/(\text{line segment 12}) \end{aligned} \tag{9.3.14}$$

Similarly,

$$\begin{aligned} \dot{m}_2/\dot{m}_m &= (h_m - h_1)/(h_2 - h_1) = (w_m - w_1)/(w_2 - w_1) \\ &= (\text{line segment 1 m1})/(\text{line segment 12}) \end{aligned} \tag{9.3.15}$$

Mixing point *m* must lie on the line that joins points 1 and 2 as shown in Figure 9.3.3.

If the differences between the density of air streams 1 and 2 and the density of the mixture are neglected,

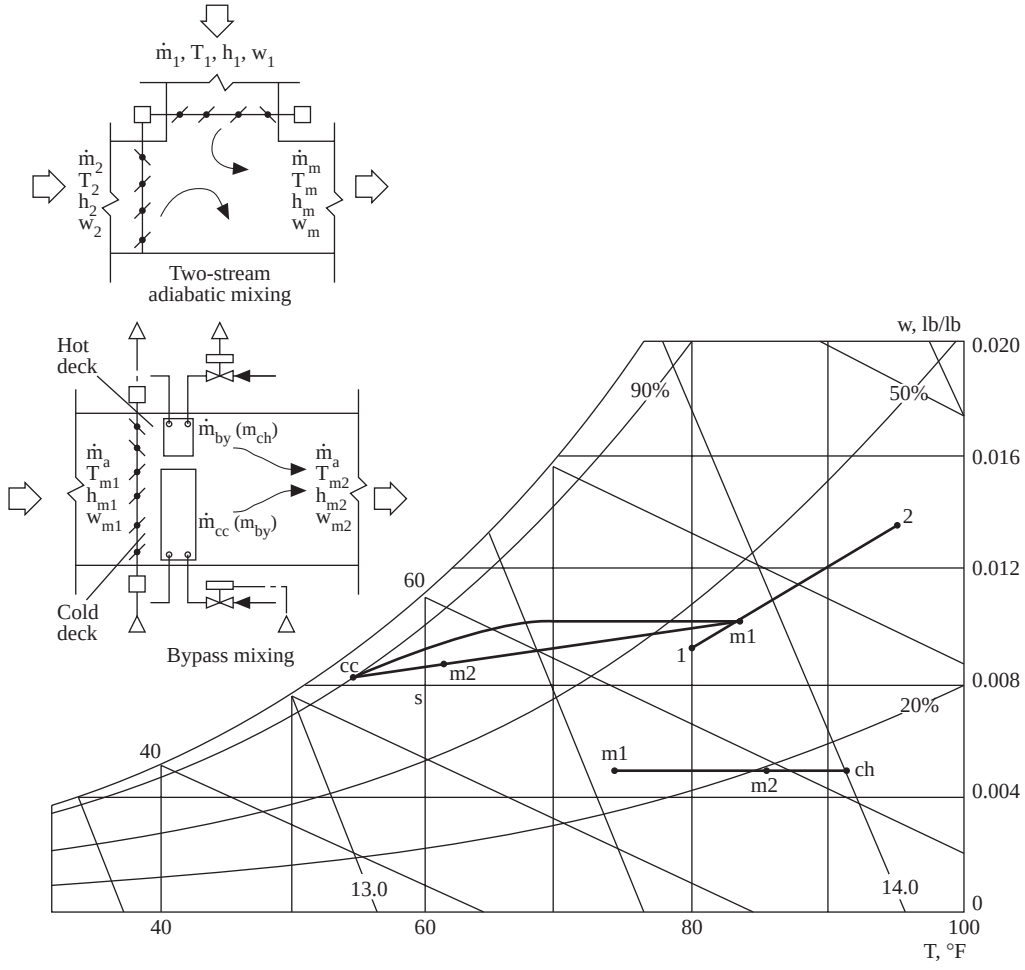


FIGURE 9.3.3 Mixing processes.

$$\mathring{V}_1 h_1 + \mathring{V}_2 h_2 = \mathring{V}_m h_m$$

$$\mathring{V}_1 w_1 + \mathring{V}_2 w_2 = \mathring{V}_m w_m \qquad (9.3.16)$$

$$\mathring{V}_1 T_1 + \mathring{V}_2 T_2 = \mathring{V}_m T_m$$

$$\mathring{V}_1 + \mathring{V}_2 = \mathring{V}_m \qquad (9.3.17)$$

In a *bypass mixing process*, a conditioned air stream is mixed with a bypass air stream that is not conditioned. The cold conditioned air is denoted by subscript cc, the heated air ch, and the bypass air by.

Equations (9.3.14) and (9.3.17) can still be used but subscript 1 should be replaced by cc or ch and subscript 2 by "by" (bypass).

Let $K_{cc} = \mathring{m}_{cc}/\mathring{m}_m$ and $K_{ch} = \mathring{m}_{ch}/\mathring{m}_m$; then the cooling coil's load q_{cc} and heating coil's load q_{ch}, both in Btu/hr, for a bypass mixing process are

$$q_{cc} = K_{cc} \dot{V}_s \rho_s (h_m - h_{cc})$$
$$q_{ch} = K_{ch} \dot{V}_s \rho_s (h_2 - h_1) \quad (9.3.18)$$

In Equation (9.3.18), subscript s denotes the supply air and m the mixture air stream.

Air-Conditioning Cycle and Operating Modes

An *air-conditioning cycle* comprises several air-conditioning processes that are connected in a sequential order. An air-conditioning cycle determines the operating performance of the air system in an air-conditioning system. The *working substance* to condition air may be chilled or hot water, refrigerant, desiccant, etc.

Each type of air system has its own air-conditioning cycle. Psychrometric analysis of an air-conditioning cycle is an important tool in determining its operating characteristics and the state of moist air at various system components, including the volume flow rate of supply air, the coil's load, and the humidifying and dehumidifying capacity.

According to the cycle performance, air-conditioning cycles can be grouped into two categories:

- *Open cycle*, in which the moist air at its end state does not resume its original state. An air-conditioning cycle with all outdoor air is an open cycle.
- *Closed cycle*, in which moist air resumes its original state at its end state. An air-conditioning cycle that conditions the mixture of recirculating and outdoor air, supplies it, recirculates part of the return air, and mixes it again with outdoor air is a closed cycle.

Based on the outdoor weather and indoor operating conditions, the operating modes of air-conditioning cycles can be classified as:

- *Summer mode*: when outdoor and indoor operating parameters are in summer conditions.
- *Winter mode*: when outdoor and indoor operating parameters are in winter conditions.
- *Air economizer mode*: when all outdoor air or an amount of outdoor air that exceeds the minimum amount of outdoor air required for the occupants is taken into the AHU or PU for cooling. The air economizer mode saves energy use for refrigeration.

Continuous modes operate 24 hr a day and 7 days a week. Examples are systems that serve hospital wards and refrigerated warehouses. An *intermittently operated mode* usually shuts down once or several times within a 24-hr operating cycle. Such systems serve offices, class rooms, retail stores, etc. The 24-hr day-and-night cycle of an intermittently operated system can again be divided into:

1. *Cool-down or warm-up period.* When the space is not occupied and the space air temperature is higher or lower than the predetermined value, the space air should be cooled down or warmed up before the space is occupied.
2. *Conditioning period*. The air-conditioning system is operated during the occupied period to maintain the required indoor environment.
3. *Nighttime shut-down period.* The air system or terminal is shut down or only partly operating to maintain a set-back temperature.

Summer, winter, air economizer, and continuously operating modes consist of *full-load* (design load) and part-load operations. *Part load* occurs when the system load is less than the design load. The capacity of the equipment is selected to meet summer and winter system design loads as well as system loads in all operating modes.

Basic Air-Conditioning Cycle — Summer Mode

A *basic air-conditioning system* is a packaged system of supply air at a constant volume flow rate, serving a single zone, equipped with only a single supply/return duct. A *single zone* is a conditioned space for which a single controller is used to maintain a unique indoor operating parameter, probably indoor temperature. A *basic air-conditioning cycle* is the operating cycle of a basic air-conditioning system. Figure 9.1.3 shows a basic air-conditioning system. Figure 9.3.4 shows the basic air-conditioning cycle of this system. In summer mode at design load, recirculating air from the conditioned space, a worship hall, enters the packaged unit through the return grill at point ru. It is mixed with the required minimum amount of outdoor air at point o for acceptable indoor air quality and energy saving. The mixture *m* is then cooled and dehumidified to point cc at the DX coil, and the conditioned air is supplied to the hall through the supply fan, supply duct, and ceiling diffuser. Supply air then absorbs the sensible and latent load from the space, becoming the space air r. Recirculating air enters the packaged unit again and forms a closed cycle. *Return air* is the air returned from the space. Part of the return air is exhausted to balance the outdoor air intake and infiltration. The remaining part is the *recirculating air* that enters the PU or AHU.

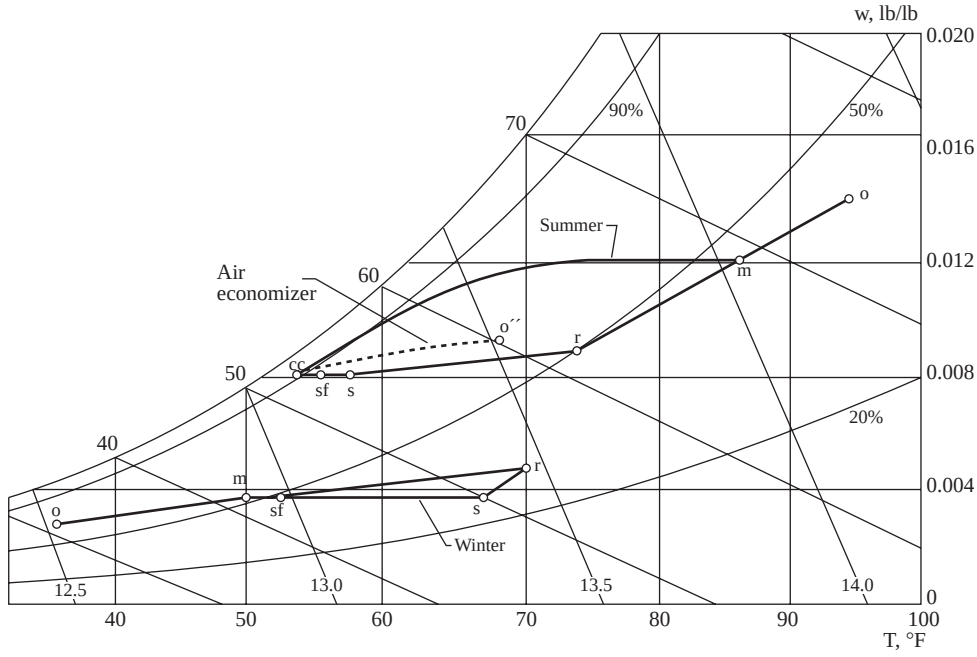


FIGURE 9.3.4 Basic air-conditioning cycle — summer, winter, and air economizer mode.

The summer mode operating cycle consists of the following processes:

1. Sensible heating process, represented by line r ru, due to the return system gain $q_{r.s}$, in Btu/hr, when recirculating air flows through the return duct, ceiling plenum, and return fan, if any. In this packaged system, the return system heat gain is small and neglected.
2. Adiabatic mixing process of recirculating air at point ru and outdoor air at point o in the mixing box, represented by line ru m o.
3. Cooling and dehumidifying process m cc at the DX coil whose coil load determines the cooling capacity of the system calculated from Equation (9.3.11).
4. Sensible heating process related to the supply system heat gain $q_{s.s}$, in Btu/hr, represented by line cc sf s. $q_{s.s}$ consists of the fan power heat gain q_{sf}, line cc sf, and duct heat gain q_{sd}, line sf s, that is:

$$q_{s.s} = q_{sf} + q_{sd} = \mathring{V}_s \rho_s c_{pa} \Delta T_{s.s} \tag{9.3.19}$$

It is more convenient to use the temperature rise of the supply system $\Delta T_{s.s}$ in psychrometric analysis.

5. Supply conditioning process line sr.

Design Supply Volume Flow Rate

Design supply volume flow rate and cooling and heating capacities are primary characteristics of an air-conditioning system. Design supply volume flow rate is used to determine the size of fans, grills, outlets, air-handling units, and packaged units. For most comfort systems and many processing air-conditioning systems, *design supply volume flow rate* $\mathring{V}_{s.d}$, in cfm, is calculated on the basis of the capacity to remove the space cooling load at summer design conditions to maintain a required space temperature T_r:

$$\mathring{V}_{s.d} = q_{rc.d} \Big/ \left[60\rho_s \left(h_r - h_s\right)\right] = q_{rs.d} \Big/ \left[60\rho_s c_{pa} \left(T_r - T_s\right)\right] \tag{9.3.20}$$

where $q_{rc.d}$, $q_{rs.d}$ = design space cooling load and design sensible cooling load, Btu/hr. In Equation (9.3.20), the greater the $q_{rs.d}$, the higher $\mathring{V}_s$ will be. Specific heat c_{pa} is usually considered constant. Air density ρ_s may vary with the various types of air systems used. A greater ρ_s means a smaller $\mathring{V}_{s.d}$ for a given supply mass flow rate. For a given $q_{rs.d}$, the supply temperature differential $\Delta T_s = (T_r - T_s)$ is an important parameter that affects $\mathring{V}_{s.d}$. Conventionally, a 15 to 20°F ΔT_s is used for comfort air-conditioning systems. Recently, a 28 to 34°F ΔT_s has been adopted for cold air distribution in ice-storage central systems. When ΔT_s has a nearly twofold increase, there is a considerable reduction in $\mathring{V}_s$ and fan energy use and saving in investment on ducts, terminals, and outlets.

The summer cooling load is often greater than the winter heating load, and this is why q_{rc} or $q_{rs.d}$ is used to determine $\mathring{V}_{s.d}$, except in locations where the outdoor climate is very cold.

Sometimes the supply volume flow rate may be determined from the following requirements:

- To dilute the concentration of air contaminants in the conditioned space C_i, in mg/m^3, the design supply volume flow rate is

$$\mathring{V}_{s.d} = 2118 \mathring{m}_{par} \Big/ \left(C_1 - C_s\right) \tag{9.3.21}$$

where C_s = concentration of air contaminants in supply air, mg/m^3
$\mathring{m}_{par}$ = rate of contaminant generation in the space, mg/sec

- To maintain a required space relative humidity φ_r and a humidity ratio w_r at a specific temperature, the design supply volume flow rate is

$$\mathring{V}_{s.d} = q_{rl.d} \Big/ \left[60\rho_s \left(w_r - w_s\right) h_{fg.58}\right] \tag{9.3.22}$$

where $q_{rl.d}$ = design space latent load, Btu/hr.

- To provide a required air velocity v_r, in fpm, within the working area of a clean room, the supply volume flow rate is given by

$$\mathring{V}_{s.d} = A_r v_r \tag{9.3.23a}$$

where A_r = cross-sectional area perpendicular to the air flow in the working area, ft^2.

- To exceed the outdoor air requirement for acceptable air quality for occupants, the supply volume flow rate must be equal to or greater than

$$\mathring{V}_s \geq n\mathring{V}_{oc} \tag{9.3.23b}$$

where n = number of occupants
$\mathring{V}_{oc}$ = outdoor air requirement per person, cfm/person

- To exceed the sum of the volume flow rate of exhaust air $\mathring{V}_{ex}$ and the exfiltrated or relief air $\mathring{V}_{ef}$, both in cfm,

$$\mathring{V}_s \geq \mathring{V}_{ex} + \mathring{V}_{ef} \tag{9.3.24}$$

The design supply volume flow rate should be the largest of any of the foregoing requirements.

Rated Supply Volume Flow Rate

For an air system at atmospheric pressure, since the required mass flow rate of the supply air is a function of air density and remains constant along the air flow,

$$\begin{aligned} \mathring{m}_a &= \mathring{V}_{cc}\rho_{cc} = \mathring{V}_s\rho_s = \mathring{V}_{sf}\rho_{sf} \\ \mathring{V}_{sf} &= \mathring{V}_s\rho_s/\rho_{sf} \end{aligned} \tag{9.3.25}$$

where $\mathring{V}_{sf}$ = volume flow rate at supply fan outlet, cfm
ρ_{sf} = air density at supply fan outlet, lb/ft^3

A supply fan is rated at *standard air conditions*, that is, dry air at a temperature of 70°F, an atmospheric pressure of 29.92 in. Hg (14.697 psia), and an air density of 0.075 lb/ft^3. However, a fan is a constant-volume machine at a given fan size and speed; that is, $\mathring{V}_{sf} = \mathring{V}_{sf.r}$. Here $\mathring{V}_{sf.r}$ represents the rated volume flow rate of a fan at standard air conditions. Therefore,

$$\mathring{V}_{sf.r} = \mathring{V}_{sf} = q_{rs.d}/\left[60\rho_{sf}c_{pa}\left(T_r - T_s\right)\right] \tag{9.3.26}$$

- For conditioned air leaving the cooling coil at T_{cc} = 55°F with a relative humidity of 92% and T_{sf} of 57°F, $\rho_{sf.r} = 1/v_{sf} = 1/13.20 = 0.0758$ lb/ft^3. From Equation (9.3.26):

$$\mathring{V}_{sf.r} = q_{rs.d}/\left[60\times 0.0758\times 0.243\left(T_r - T_s\right)\right] = q_{rs.d}/\left[1.1\left(T_r - T_s\right)\right] \tag{9.3.26a}$$

Equation (9.3.26a) is widely used in calculating the supply volume flow rate of comfort air-conditioning systems.

- For cold air distribution, T_{cc} = 40°F and φ_{cc} = 98%, if T_{sf} = 42°F, then v_{sf} = 12.80 ft^3/lb, and the rated supply volume flow rate:

$$\mathring{V}_{sf.r} = 12.80q_{rs.d}/\left[60\times 0.243\left(T_r - T_s\right)\right] = q_{rs.d}/\left[1.14\left(T_r - T_s\right)\right] \tag{9.3.26b}$$

- For a blow-through fan in which the fan is located upstream of the coil, if T_{sf} = 82°F and φ_{sf} = 43%, then v_{sf} = 13.87 ft^3/lb, and the rated supply volume flow rate:

$$\mathring{V}_{sf.r} = 13.87 q_{rs.d} / \left[60 \times 0.243 \left(T_r - T_s\right)\right] = q_{rs.d} / \left[1.05 \left(T_r - T_s\right)\right] \quad (9.3.26c)$$

Effect of the Altitude

The higher the altitude, the lower the atmospheric pressure and the air density. In order to provide the required mass flow rate of supply air, a greater $\mathring{V}_{sf.r}$ is needed. For an air temperature of 70°F:

$$\mathring{V}_{x.ft} = \mathring{V}_{sf.r} \left(p_{sea} / p_{x.ft}\right) = \mathring{V}_{sf.r} \left(\rho_{sea} / \rho_{x.ft}\right) \quad (9.3.27)$$

where $\mathring{V}_{x.ft}$ = supply volume flow rate at an altitude of *x* ft, cfm
p_{sea}, $p_{x.ft}$ = atmospheric pressure at sea level and an altitude of *x* ft, psia
ρ_{sea}, $\rho_{x.ft}$ = air density at sea level and an altitude of *x* ft, psia

Following are the pressure or air density ratios at various altitudes. At 2000 ft above sea level, the rated supply volume flow rate $\mathring{V}_{r.2000} = \mathring{V}_{sf.r}$ ($p_{sea}/p_{x.ft}$) = 1.076 $\mathring{V}_{sf.r}$ cfm instead of $\mathring{V}_{sf.r}$ cfm at sea level.

Altitude, ft	p_{at}, psia	ρ, lb/ft^3	($p_{sea}/p_{x,ft}$)
0	14.697	0.075	1.000
1000	14.19	0.0722	1.039
2000	13.58	0.0697	1.076
3000	13.20	0.0672	1.116
5000	12.23	0.0625	1.200

Off-Coil and Supply Air Temperature

For a given design indoor air temperature T_r, space sensible cooling load q_{rs}, and supply system heat gain $q_{s.s}$, a lower air off-coil temperature T_{cc} as well as supply temperature T_s means a greater supply temperature differential ΔT_s and a lower space relative humidity φ_r and vice versa. A greater ΔT_s decreases the supply volume flow rate $\mathring{V}_s$ and then the fan and terminal sizes, duct sizes, and fan energy use. The result is a lower investment and energy cost.

A lower T_{cc} and a greater ΔT_s require a lower chilled water temperature entering the coil T_{we}, a lower evaporating temperature T_{ev} in the DX coil or refrigerating plant, and therefore a greater power input to the refrigerating compressors. When an air-conditioning system serves a conditioned space of a single zone, optimum T_{cc}, T_s, and T_{we} can be selected. For a conditioned space of multizones, T_{cc}, T_s, and T_{we} should be selected to satisfy the lowest requirement. In practice, T_s and T_{we} are often determined according to previous experience with similar projects.

In general, the temperature rise due to the supply fan power system heat gain q_{sf} can be taken as 1 to 3°F depending on the fan total pressure. The temperature rise due to the supply duct system heat gain at design flow can be estimated as 1°F/100 ft insulated main duct length based on 1-in. thickness of duct insulation.

Outside Surface Condensation

The outside surface temperature of the ducts, terminals, and supply outlets T_{sur} in the ceiling plenum in contact with the return air should not be lower than the dew point of the space air T_r'', in °F. The temperature rise due to the fan power heat gain is about 2°F. According to Dorgan (1988), the temperature difference between the conditioned air inside the terminal and the outside surface of the terminal with insulation wrap is about 3°F. For a space air temperature of 75°F and a relative humidity of 50%, its dew point temperature is 55°F. If the outside surface temperature $T_s = (T_{cc} + 2 + 3) \le 55$°F, condensation may occur on the outside surface of the terminal. Three methods are often used to prevent condensation:

1. Increase the thickness of the insulation layer on the outside surface.
2. Adopt a supply outlet that induces more space air.

3. Equip with a terminal that mixes the supply air with the space air or air from the ceiling plenum.

During the cool-down period, due to the high dew point temperature of the plenum air when the air system is started, the supply air temperature must be controlled to prevent condensation.

Example 9.3.1

The worship hall of a church uses a package system with a basic air system. The summer space sensible cooling load is 75,000 Btu/hr with a latent load of 15,000 Btu/hr. Other design data for summer are as follows:

Outdoor summer design temperature: dry bulb 95°F and wet bulb 75°F
Summer indoor temperature: 75°F with a space relative humidity of 50%:
Temperature rise: fan power 2°F
supply duct 2°F
Relative humidity of air leaving cooling coil: 93%
Outdoor air requirement: 1800 cfm

Determine the

1. Temperature of supply air at summer design conditions
2. Rated volume flow rate of the supply fan
3. Cooling coil load
4. Possibility of condensation at the outside surface of the insulated branch duct to the supply outlet

Solution

1. From Equation 9.3.1 the sensible heat ratio of the space conditioning line is

$$\mathrm{SHR}_s = |q_{rs}| / (|q_{rs}| + |q_{rl}|) = 60{,}000/(60{,}000 + 15{,}000) = 0.8$$

   On the psychrometric chart, from given T_r = 75°F and φ_r = 50%, plot space point r. Draw a space conditioning line sr from point r with SHR_s = 0.8.
   Since $\Delta T_{s.s}$ = 2 + 2 = 4°F, move line segment cc s (4°F) up and down until point s lies on line sr and point cc lies on the φ_{cc} = 93% line. The state points s and cc are then determined as shown in Figure 9.3.4:

   T_s = 57.5°F, φ_s = 82%, and w_s = 0.0082 lb/lb
   T_{cc} = 53.5°F, φ_{cc} = 93%, h_{cc} = 21.8 Btu/lb, and w_{cc} = 0.0082 lb/lb

2. Since T_{sf} = 53.5 + 2 = 55.5°F and w_{sf} = 0.0082 lb/lb, $\rho_{sf} = 1/v_{sf}$ = 1/13.15 = 0.076 lb/ft^3. From Equation 9.4.2, the required rated supply volume flow rate is

$$\dot{V}_{sf.r} = q_{rs.d} / [60 \rho_{sf} c_{pa} (T_r - T_s)]$$

$$= 60{,}000 / [60 \times 0.076 \times 0.243 (75 - 57.5)] = 3094 \text{ cfm}$$

3. Plot outdoor air state point o on the psychrometric chart from given dry bulb 95°F and wet bulb 75°F. Connect line ro. Neglect the density differences between points r, m, and o; then

$$\mathrm{rm}/\mathrm{ro} = 1800/3094 = 0.58$$

   From the psychrometric chart, the length of line ro is 2.438 in. As shown in Figure 9.3.4, point m is then determined as:

$$T_m = 86.7°F, \qquad h_m = 35 \text{ Btu/lb}$$

From Equation (9.3.11), the cooling coil load is

$$q_{cc} = 60\dot{V}_s\rho_s\left(h_m - h_{cc}\right) = 60 \times 3094 \times 0.076(35 - 21.8) = 186{,}234 \text{ Btu/lb}$$

4. From the psychrometric chart, since the dew point of the space air T_r'' = 55°F and is equal to that of the plenum air, the outside surface temperature of the branch duct T_s = 53.5 + 2 + 3 = 58°F which is higher than T_r'' = 55°F. Condensation will not occur at the outside surface of the branch duct.

Basic Air-Conditioning Cycle — Winter Mode

When the basic air-conditioning systems are operated in winter mode, their air-conditioning cycles can be classified into the following four categories:

Cold Air Supply without Space Humidity Control. In winter, for a fully occupied worship hall, if the heat loss is less than the space sensible cooling load, a cold air supply is required to offset the space sensible cooling load and maintain a desirable indoor environment as shown by the lower cycle in Figure 9.3.4. Usually, a humidifier is not used.

The winter cycle of a cold air supply without humidity control consists of the following air-conditioning processes:

1. Adiabatic mixing process of outdoor air and recirculating air o m r.
2. Sensible heating process due to supply fan power heat gain m sf. Because of the smaller temperature difference between the air in the ceiling plenum and the supply air inside the supply duct, heat transfer through duct wall in winter can be neglected.
3. Supply conditioning line sr.

For a winter-mode basic air-conditioning cycle with a cold air supply without space humidity control, the space relative humidity depends on the space latent load, the humidity ratio of the outdoor air, and the amount of outdoor air intake. In order to determine the space humidity ratio w_r, in lb/lb, and the space relative humidity φ_r, in %, Equations (9.3.15) and (9.3.22) should be used to give the following relationships:

$$\begin{aligned} \left(w_r - w_m\right)/\left(w_r - w_o\right) &= \dot{V}_o/\dot{V}_s \\ \left(w_r - w_s\right) &= q_{rl}/\left(60\dot{V}_s\rho_s h_{fg.58}\right) \\ w_s &= w_m \end{aligned} \tag{9.3.28}$$

For a cold air supply, if there is a high space sensible cooling load, the amount of outdoor air must be sufficient, and the mixture must be cold enough to satisfy the following relationships:

$$\begin{aligned} \left(T_r - T_s\right) &= q_{rs}/\left(60\dot{V}_s\rho_s c_{pa}\right) \\ \left(T_r - T_s\right)/\left(T_r - T_o\right) &= \dot{V}_o/\dot{V}_s \end{aligned} \tag{9.3.29}$$

The heating coil load for heating of the outdoor air can be calculated using Equation (9.3.7).

Example 9.3.2

For the same packaged air-conditioning system using a basic air system to serve the worship hall in a church as in Example 9.3.1, the space heating load at winter design condition is 10,000 Btu/hr and the latent load is 12,000 Btu/hr. Other winter design data are as follows:

Winter outdoor design temperature	35°F
Winter outdoor design humidity ratio	0.00035 lb/lb
Winter indoor design temperature	70°F
Temperature rise due to supply fan heat gain	2°F
Outdoor air requirement	1800 cfm

Determine (1) the space relative humidity at winter design temperature and (2) the heating coil load.

Solution

1. Assume that the supply air density $\rho_{sf} = 1/v_{sf} = 1/13.0 = 0.0769$ lb/ft^3, and the mass flow rate of the supply air is the same as in summer mode. Then from Equation 9.3.28 the humidity ratio difference is

$$\left(w_r - w_s\right) = q_{rl} \Big/ \left(60\,\mathring{V}_{sf.r}\,\rho_{sf}\,h_{fg.58}\right) = 12{,}000/(60 \times 3094 \times 0.0769 \times 1060) = 0.00079 \text{ lb/lb}$$

   From Equation 9.3.29, the supply air temperature differential is

$$\left(T_r - T_s\right) = q_{rs.d} \Big/ \left(60\,\mathring{V}_{sf.r}\,\rho_{sf}\,c_{pa}\right) = 10{,}000/(60 \times 3094 \times 0.0769 \times 0.243) = 2.88°\text{F}$$

   Since $\mathring{V}_o / \mathring{V}_s$ = 1800/3094 = 0.58 and $w_s = w_m$,

$$\left(w_r - w_s\right)/\left(w_r - w_o\right) = 0.00079/\left(w_r - w_o\right) = \mathring{V}_o \Big/ \mathring{V}_s = 0.58$$

$$\left(w_r - w_o\right) = 0.00079/0.58 = 0.00136 \text{ lb/lb}$$

   And from given information,

$$w_r = 0.00136 + w_o = 0.00136 + 0.0035 = 0.00486 \text{ lb/lb}$$

   From the psychrometric chart, for T_r = 70°F and w_r = 0.00486 lb/lb, point r can be plotted, and φ_r is about 32% (see Figure 9.3.4).
2. Since mr/or = 0.58, point m can be determined, and from the psychrometric chart T_m = 50.0°F. As T_s = 70 – 2.88 = 67.12°F and $T_{sf} = T_m + 2 = 50.0 + 2 = 52.0$°F, from Equation 9.3.7 the heating coil's load is

$$q_{ch} = 60\,\mathring{V}_s\,\rho_s c_{pa}\left(T_s - T_{sf}\right) = 60 \times 3094 \times 0.0769 \times 0.243(67.12 - 52.0) = 52{,}451 \text{ Btu/hr}$$

Warm Air Supply without Space Humidity Control

When the sum of space heat losses is greater than the sum of the internal heat gains in winter, a warm air supply is needed. For many comfort systems such as those in offices and stores, in locations where winter is not very cold, humidification is usually not necessary. The basic air-conditioning cycle for a warm air supply without space humidity control is shown in Figure 9.3.5(a). This cycle is similar to the

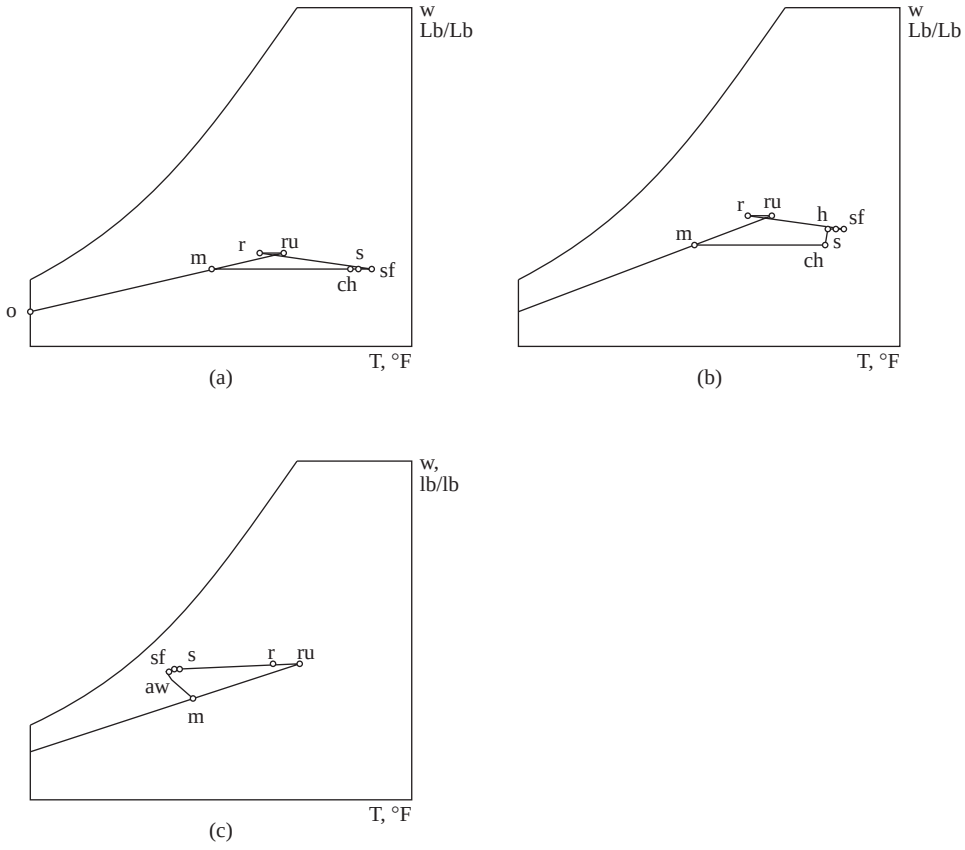


FIGURE 9.3.5 Basic air-conditioning cycle — winter modes: (a) warm air supply without space humidity control, (b) cold air supply without space humidity control, and (c) cold air supply with space humidity control. ch = air leaving heating coil, h = air leaving humidifer, and aw = air leaving air washer.

winter mode cycle of a cold air supply without space humidity control shown in Figure 9.3.4 except that the supply air temperature is higher than space temperature, that is, $T_s > T_r$. To prevent stratification, with the warm supply air staying at a higher level, $(T_s - T_r) > 20°F$ is not recommended.

Warm Air Supply with Space Humidity Control

This operating cycle (see Figure 9.3.5[b]) is often used for hospitals, nurseries, etc. or in locations where winter is very cold. The state point of supply air must be determined first by drawing a space conditioning line with known SHR_s and then from the calculated supply temperature differential ΔT_s. The difference in humidity ratio $(w_s - w_{ch})$ is the water vapor must be added at the humidifier. Humidifying capacity can be calculated from Equation 9.3.8.

Cold Air Supply with Space Humidity Control

This operating cycle (shown in Figure 9.3.5[c]) is widely used in industrial applications such as textile mills where a cold air supply is needed to remove machine load in winter and maintains the space relative humidity required for the manufacturing process. An outdoor air and recirculating air mixture is often used for the required cold air supply. An air washer is adopted for winter humidification.

Air Economizer Mode

In the *air economizer* mode, as shown by the middle dotted line cycle o″-cc-sf-s-r in Figure 9.3.4, all outdoor air or an outdoor air-recirculating air mixture is used to reduce the refrigeration capacity and improve the indoor air quality during spring, fall, or winter.

When all outdoor air is admitted, it is an open cycle. Outdoor air is cooled and often dehumidified to point cc. After absorbing fan and duct heat gains, it is supplied to the conditioned space. Space air is exhausted entirely through openings, relief dampers, or relief/exhaust fans to the outside. An all-outdoor air-operating mode before the space is occupied is often called an *air purge* operation, used to dilute space air contaminants.

Cool-Down and Warm-Up Modes

In summer, when an air system is shut down during an unoccupied period at night, the space temperature and relative humidity often tend to increase because of infiltration of hot and humid air and heat transfer through the building envelope. The air system is usually started before the space is occupied in cool-down mode to cool the space air until the space temperature falls within predetermined limits.

In winter, the air system is also started before the occupied period to warm up the space air to compensate for the nighttime space temperature setback to 55 to 60°F for energy saving or the drop of space temperature due to heat loss and infiltration.

If dilution of indoor air contaminants is not necessary, only recirculating space air is used during cool-down or warm-up periods in order to save energy.

9.4 Refrigerants and Refrigeration Cycles

Refrigeration and Refrigeration Systems

Refrigeration is the cooling effect of the process of extracting heat from a lower temperature heat source, a substance or cooling medium, and transferring it to a higher temperature heat sink, probably atmospheric air and surface water, to maintain the temperature of the heat source below that of the surroundings.

A *refrigeration system* is a combination of components, equipment, and piping, connected in a sequential order to produce the refrigeration effect. Refrigeration systems that provide cooling for air-conditioning are classified mainly into the following categories:

1. *Vapor compression systems.* In these systems, a compressor(s) compresses the refrigerant to a higher pressure and temperature from an evaporated vapor at low pressure and temperature. The compressed refrigerant is condensed into liquid form by releasing the latent heat of condensation to the condenser water. Liquid refrigerant is then throttled to a low-pressure, low-temperature vapor, producing the refrigeration effect during evaporation. Vapor compression is often called *mechanical refrigeration*, that is, refrigeration by mechanical compression.
2. *Absorption systems*. In an absorption system, the refrigeration effect is produced by means of thermal energy input. After liquid refrigerant produces refrigeration during evaporation at very low pressure, the vapor is absorbed by an aqueous absorbent. The solution is heated by a direct-fired gas furnace or waste heat, and the refrigerant is again vaporized and then condensed into liquid form. The liquid refrigerant is throttled to a very low pressure and is ready to produce the refrigeration effect again.
3. *Gas expansion systems.* In an air or other gas expansion system, air or gas is compressed to a high pressure by compressors. It is then cooled by surface water or atmospheric air and expanded to a low pressure. Because the temperature of air or gas decreases during expansion, a refrigeration effect is produced.

Refrigerants, Cooling Mediums, and Absorbents

A *refrigerant* is a primary working fluid used to produce refrigeration in a refrigeration system. All refrigerants extract heat at low temperature and low pressure during evaporation and reject heat at high temperature and pressure during condensation.

A *cooling medium* is a working fluid cooled by the refrigerant during evaporation to transport refrigeration from a central plant to remote cooling equipment and terminals. In a large, centralized air-conditioning system, it is more economical to pump the cooling medium to the remote locations where cooling is required. Chilled water and brine are cooling media. They are often called secondary refrigerants to distinguish them from the primary refrigerants.

A *liquid absorbent* is a working fluid used to absorb the vaporized refrigerant (water) after evaporation in an absorption refrigeration system. The solution that contains the absorbed vapor is then heated. The refrigerant vaporizes, and the solution is restored to its original concentration to absorb water vapor again.

A numbering system for refrigerants was developed for hydrocarbons and halocarbons. According to ANSI/ASHRAE Standard 34-1992, the first digit is the number of unsaturated carbon–carbon bonds in the compound. This digit is omitted if the number is zero. The second digit is the number of carbon atoms minus one. This is also omitted if the number is zero. The third digit denotes the number of hydrogen atoms plus one. The last digit indicates the number of fluorine atoms. For example, the chemical formula for refrigerant R-123 is $CHCl_2CF_3$. In this compound:

No unsaturated carbon–carbon bonds, first digit is 0
There are two carbon atoms, second digit is $2 - 1 = 1$
There is one hydrogen atom, third digit is $1 + 1 = 2$

There are three fluorine atoms, last digit is 3

To compare the relative ozone depletion of various refrigerants, an index called the *ozone depletion potential* (ODP) has been introduced. ODP is defined as the ratio of the rate of ozone depletion of 1 lb of any halocarbon to that of 1 lb of refrigerant R-11. For R-11, ODP = 1.

Similar to the ODP, halocarbon global warming potential (HGWP) is an index used to compare the global warming effect of a halocarbon refrigerant with the effect of refrigerant R-11.

Classification of Refrigerants

Nontoxic and nonflammable synthetic chemical compounds called *halogenated hydrocarbons*, or simply *halocarbons*, were used almost exclusively in vapor compression refrigeration systems for comfort air-conditioning until 1986. Because chlorofluorocarbons (CFCs) cause ozone depletion and global warming, they must be replaced. A classification of refrigerants based on ozone depletion follows (see Table 9.4.1):

Hydrofluorocarbons (HFCs)

HFCs contain only hydrogen, fluorine, and carbon atoms and cause no ozone depletion. HFCs group include R-134a, R-32, R-125, and R-245ca.

HFC's Azeotropic Blends or Simply HFC's Azeotropic

An azeotropic is a mixture of multiple components of volatilities (refrigerants) that evaporate and condense as a single substance and do not change in volumetric composition or saturation temperature when they evaporate or condense at constant pressure. HFC's azeotropics are blends of refrigerant with HFCs. ASHRAE assigned numbers between 500 and 599 for azeotropic. HFC's azeotropic R-507, a blend of R-125/R-143, is the commonly used refrigerant for low-temperature vapor compression refrigeration systems.

HFC's Near Azeotropic

A near azeotropic is a mixture of refrigerants whose characteristics are near those of an azeotropic. Because the change in volumetric composition or saturation temperature is rather small for a near azeotropic, such as, 1 to 2°F, it is thus named. ASHRAE assigned numbers between 400 and 499 for zeotropic. R-404A (R-125/R-134a/R-143a) and R-407B (R-32/R-125/R134a) are HFC's near azeotropic. R-32 is flammable; therefore, its composition is usually less than 30% in the mixture. HFC's near azeotropic are widely used for vapor compression refrigeration systems.

Zeotropic or nonazeotropic, including near azeotropic, shows a change in composition due to the difference between liquid and vapor phases, leaks, and the difference between charge and circulation. A shift in composition causes the change in evaporating and condensing temperature/pressure. The difference in dew point and bubble point during evaporation and condensation is called glide, expressed in °F. Near azeotropic has a smaller glide than zeotropic. The midpoint between the dew point and bubble point is often taken as the evaporating and condensing temperature for refrigerant blends.

Hydrochlorofluorocarbons (HCFCs) and Their Zeotropics

HCFCs contain hydrogen, chlorine, fluorine, and carbon atoms and are not fully halogenated. HCFCs have a much shorter lifetime in the atmosphere (in decades) than CFCs and cause far less ozone depletion (ODP 0.02 to 0.1). R-22, R-123, R-124, etc. are HCFCs. HCFCs are the most widely used refrigerants today.

HCFC's near azeotropic and *HCFC's zeotropic* are blends of HCFCs with HFCs. They are transitional or interim refrigerants and are scheduled for a restriction in production starting in 2004.

Inorganic Compounds

These compounds include refrigerants used before 1931, like ammonia R-717, water R-718, and air R-729. They are still in use because they do not deplete the ozone layer. Because ammonia is toxic and

TABLE 9.4.1 Properties of Commonly Used Refrigerants 40°F Evaporating and 100°F Condensin

		Chemical Formula	Molecular Mass	Ozone Depletion Potential (ODP)	Global Warming Potential (HGWP)	Evaporating Pressure, psia	Condensing Pressure, psia	Compression Ratio	Refrigeration Effect, Btu/lb
Hydrofluorocarbons HFCs									
R-32	Difluoromethane	CH_2F_2	52.02	0.0	0.14	135.6	340.2	2.51	
R-125	Pentafluoroethane	CHF_2CF_3	120.03	0.0	0.84	111.9	276.2	2.47	37.1
R-134a	Tetrafluoroethane	CF_3CH_2F	102.03	0.0	0.26	49.7	138.8	2.79	65.2
R-143a	Trifluoroethane	CH_3CF_3	84.0	0.0					
R-152a	Difluoroethane	CH_3CHF_2	66.05	0.0		44.8	124.3	2.77	
R-245ca	Pentafluoropropane	$CF_3CF_2CH_3$	134.1	0.0					
HFC's azeotropics									
R-507	R-125/R-143 (45/55)			0.0	0.98				
HFC's near azeotropic									
R-404A	R-125/R-143a (44/52/4)			0.0	0.94				
R-407A	R-32/R-125/R-134a (20/40/40)			0.0	0.49				
R-407C	R-32/R-125/R-134a (23/25/52)			0.0	0.70				
Hydrochlorofluorocarbons HCFCs and their azeotropics									
R-22	Chlorodifluoromethane	$CHCIF_2$	86.48	0.05	0.40	82.09	201.5	2.46	69.0
R-123	Dichlorotrifluoroethane	$CHCl_2CF_3$	152.93	0.02	0.02	5.8	20.8	3.59	62.9
R-124	Chlorotetrafluoroethane	$CHFClCF_3$	136.47	0.02		27.9	80.92	2.90	5.21
HCFC's near azeotropics									
R-402A	R-22/R-125/R-290 (38/60/2)			0.02	0.63				
HCFC's azeotropics									
R-401A	R-22/R-124/R-152a (53/34/13)			0.37	0.22				
R-401B	R-22/R-124/R-152a (61/28/11)			0.04	0.24				

TABLE 9.4.1 Properties of Commonly Used Refrigerants 40°F Evaporating and 100°F Condensin (continued)

		Chemical Formula	Molecular Mass	Ozone Depletion Potential (ODP)	Global Warming Potential (HGWP)	Evaporating Pressure, psia	Condensing Pressure, psia	Compression Ratio	Refrigeration Effect, Btu/lb
Inorganic compounds									
R-717	Ammonia	NH_3	17.03	0	0	71.95	206.81	2.87	467.4
R-718	Water	H_2O	18.02	0					
R-729	Air		28.97	0					
Chlorofluorocarbons CFCs, halons BFCs and their azeotropic									
R-11	Trichlorofluoromethane	CCl_3F	137.38	1.00	1.00	6.92	23.06	3.33	68.5
R-12	Dichlorodifluoromethane	CCl_2F_2	120.93	1.00	3.20	50.98	129.19	2.53	50.5
R-13B1	Bromotrifluoromethane	$CBrF_3$	148.93	10					
R-113	Trichlorotrifluoroethane	CCl_2FCClF_2	187.39	0.80	1.4	2.64	10.21	3.87	54.1
R-114	Dichlorotetrafluoroethane	CCl_2FCF_3	170.94	1.00	3.9	14.88	45.11	3.03	42.5
R-500	R-12/R-152a (73.8/26.2)		99.31			59.87	152.77	2.55	60.5
R-502	R-22/R-115 (48.8/51.2)		111.63	0.283	4.10				

TABLE 9.4.1 Properties of Commonly Used Refrigerants 40°F Evaporating and 100°F Condensing (continued)

	Replacement of	Trade Name	Specific Volume of Vapor ft^3/lb	Compresssor Displacement cfm/ton	Power Consumption hp/ton	Critical Temperature °F	Discharge Temperature °F	Flammability	Safety
Hydrofluorocarbons HFCs									
R-32			0.63			173.1			
R-125			0.33			150.9	103	Nonflammable	A1
R134a	R-12		0.95			213.9		Nonflammable	A1
R143a									
R-152a			1.64			235.9		Lower flammable	A2
R-245ca									
HFC's azeotropics									
R-507	R-502	Genetron AZ-50							
HFC's near azeotropic									
R-404A	R-22	SUVA HP-62							A1/A1[a]
R-407A	R-22	KLEA 60							A1/A1[a]
R-407C	R-22	KLEA 66							A1/A1[a]
Hydrochlorofluorocarbons HCFC's and their azeotropics									
R-22			0.66	1.91	0.696	204.8	127	Nonflammable	A1
R-123	R-11		5.88	18.87	0.663	362.6		Nonflammable	B1
R-124			1.30	5.06	0.698	252.5			
HCFC's near azeotropics									
R-402A	R-502	SUVA HP-80							A1/A1[a]
HCFC's azeotropics									
R-401A	R-12	MP 39							A1/A1[a]
R-401B	R-12	MP 66							A1/A1[a]
Inorganic compounds									
R-717			3.98	1.70	0.653	271.4	207	Lower flammability	B2
R-718								Nonflammable	
R-729								Nonflammable	

TABLE 9.4.1 Properties of Commonly Used Refrigerants 40°F Evaporating and 100°F Condensing (continued)

	Replacement of	Trade Name	Specific Volume of Vapor ft³/lb	Compresssor Displacement cfm/ton	Power Consumption hp/ton	Critical Temperature °F	Discharge Temperature °F	Flammability	Safety
Chlorofluorocarbons CFCs, halons BFCs, and their azeotropics									
R-11			5.43	15.86	0.636	388.4	104	Nonflammable	A1
R-12			5.79	3.08	0.689	233.6	100	Nonflammable	A1
R-13B1			0.21			152.6	103	Nonflammable	A1
R-113			10.71	39.55	0.71	417.4	86	Nonflammable	A1
R-114			2.03	9.57	0.738	294.3	86	Nonflammable	A1
R-500	R-12/R-152a (73.8/26.2)		0.79	3.62	0.692	221.9	105	Nonflammable	A1
R-502	R-22/R-115 (48.8/51.2)						98	Nonflammable	A1

Source: Adapted with permission from *ASHRAE Handbooks 1993 Fundamentals.* Also from refrigerant manufacturers.

[a] First classification is that safety classification of the formulated composition. The second is the worst case of fractionation.

flammable, it is used in industrial applications. Inorganic compounds are assigned numbers between 700 and 799 by ASHRAE.

Chlorofluorocarbons, Halons, and Their Azeotropic

CFCs contain only chlorine, fluorine, and carbon atoms. CFCs have an atmospheric lifetime of centuries and cause ozone depletion (ODP from 0.6 to 1). R-11, R-12, R-113, R-114, R-115… are all CFCs.

Halons or BFCs contain bromide, fluorine, and carbon atoms. R-13B1 and R-12B1 are BFCs. They cause very high ozone depletion (ODP for R-13B1 = 10). Until 1995, R-13B1 was used for very low temperature vapor compression refrigeration systems.

Phaseout of CFCs, BFCs, HCFCs, and Their Blends

On September 16, 1987, the European Economic Community and 24 nations, including the United States, signed a document called the Montreal Protocol. It is an agreement to restrict the production and consumption of CFCs and BFCs in the 1990s because of ozone depletion.

The Clean Air Act amendments, signed into law in the United States on November 15, 1990, concern two important issues: the phaseout of CFCs and the prohibition of deliberate venting of CFCs and HCFCs.

In February 1992, President Bush called for an accelerated ban of CFCs in the United States. In late November 1992, representatives of 93 nations meeting in Copenhagen agreed to phase out CFCs beginning January 1, 1996. Restriction on the use of HCFCs will start in 2004, with a complete phaseout by 2030.

In the earlier 1990s, R-11 was widely used for centrifugal chillers, R-12 for small and medium-size vapor compression systems, R-22 for all vapor compression systems, and CFC/HCFC blend R-502 for low-temperature vapor compression systems. Because of the phaseout of CFCs and BFCs before 1996 and HCFCs in the early years of the next century, alternative refrigerants have been developed to replace them:

- R-123 (an HCFC of ODP = 0.02) to replace R-11 is a short-term replacement that causes a slight reduction in capacity and efficiency. R-245ca (ODP = 0) may be the long-term alternative to R-11.
- R-134a (an HFC with ODP = 0) to replace R-12 in broad applications. R-134a is not miscible with mineral oil; therefore, a synthetic lubricant of polyolester is used.
- R-404A (R-125/R-134a/143a) and R-407C (R-32/R-125/R-134a) are both HFCs near azeotropic of ODP = 0. They are long-term alternatives to R-22. For R-407C, the composition of R-32 in the mixture is usually less than 30% so that the blend will not be flammable. R-407C has a drop of only 1 to 2% in capacity compared with R-22.
- R-507 (R-125/R-143a), an HFC's azeotropic with ODP = 0, is a long-term alternative to R-502. Synthetic polyolester lubricant oil will be used for R-507. There is no major performance difference between R-507 and R-502. R-402A (R-22/R-125/R-290), an HCFC's near azeotropic, is a short-term immediate replacement, and drop-in of R-502 requires minimum change of existing equipment except for reset of a higher condensing pressure.

Required Properties of Refrigerants

A refrigerant should not cause ozone depletion. A low global warming potential is required. Additional considerations for refrigerant selection are

1. *Safety*, including toxicity and flammability. ANSI/ASHRAE Standard 34-1992 classifies the *toxicity* of refrigerants as Class A and Class B. Class A refrigerants are of low toxicity. No toxicity was identified when their time-weighted average concentration was less than or equal to 400 ppm, to which workers can be exposed for an 8-hr workday and 40-hr work week without adverse effect. Class B refrigerants are of higher toxicity and produce evidence of toxicity.

ANSI/ASHRAE Standard 34-1982 classifies the *flammability* of refrigerants as Class 1, no flame propagation; Class 2, lower flammability; and Class 3, higher flammability.

The safety classification of refrigerants is based on the combination of toxicity and flammability: A1, A2, A3, B1, B2, and B3. R-134a and R-22 are in the A1 group, lower toxicity and nonflammable; R-123 in the B1 group, higher toxicity and nonflammable; and R-717 (ammonia) in the B2 group, higher toxicity and lower flammability.

2. *Effectiveness of refrigeration cycle*. High effectiveness is a desired property. The power consumed per ton of refrigeration produced, hp/ton or kW/ton, is an index for this assessment. Table 9.4.1 gives values for an ideal single-stage vapor compression cycle.
3. *Oil miscibility.* Refrigerant should be miscible with mineral lubricant oil because a mixture of refrigerant and oil helps to lubricate pistons and discharge valves, bearings, and other moving parts of a compressor. Oil should also be returned from the condenser and evaporator for continuous lubrication. R-22 is partially miscible. R-134a is hardly miscible with mineral oil; therefore, synthetic lubricant of polyolester will be used.
4. *Compressor displacement.* Compressor displacement per ton of refrigeration produced, in cfm/ton, directly affects the size of the positive displacement compressor and therefore its compactness. Ammonia R-717 requires the lowest compressor displacement (1.70 cfm/ton) and R-22 the second lowest (1.91 cfm/ton).
5. Desired properties:
   - Evaporating pressure p_{ev} should be higher than atmospheric. Then noncondensable gas will not leak into the system.
   - Lower condensing pressure for lighter construction of compressor, condenser, piping, etc.
   - A high thermal conductivity and therefore a high heat transfer coefficient in the evaporator and condenser.
   - Dielectric constant should be compatible with air when the refrigerant is in direct contact with motor windings in hermetic compressors.
   - An inert refrigerant that does not react chemically with material will avoid corrosion, erosion, or damage to system components. Halocarbons are compatible with all containment materials except magnesium alloys. Ammonia, in the presence of moisture, is corrosive to copper and brass.
   - Refrigerant leakage can be easily detected. Halide torch, electronic detector, and bubble detection are often used.

Ideal Single-Stage Vapor Compression Cycle

Refrigeration Process

A refrigeration process shows the change of the thermodynamic properties of the refrigerant and the energy and work transfer between the refrigerant and surroundings.

Energy and work transfer is expressed in British thermal units per hour, or Btu/hr. Another unit in wide use is ton of refrigeration, or ton. A ton = 12,000 Btu/hr of heat removed; i.e., 1 ton of ice melting in 24 hr = 12,000 Btu/hr.

Refrigeration Cycles

When a refrigerant undergoes a series of processes like evaporation, compression, condensation, throttling, and expansion, absorbing heat from a low-temperature source and rejecting it to a higher temperature sink, it is said to have undergone a refrigeration cycle. If its final state is equal to its initial state, it is a *closed cycle*; if the final state does not equal the initial state, it is an *open cycle*. Vapor compression refrigeration cycles can be classified as single stage, multistage, compound, and cascade cycles.

A *pressure-enthalpy diagram* or *p-h* diagram is often used to calculate the energy transfer and to analyze the performance of a refrigeration cycle, as shown in Figure 9.4.1. In a *p-h* diagram, pressure *p*, in psia or psig logarithmic scale, is the ordinate, and enthalpy *h*, in Btu/lb, is the abscissa. The saturated liquid and saturated vapor line encloses a two-phase region in which vapor and liquid coexist. The two-phase region separates the subcooling liquid and superheated vapor regions. The constant-temperature

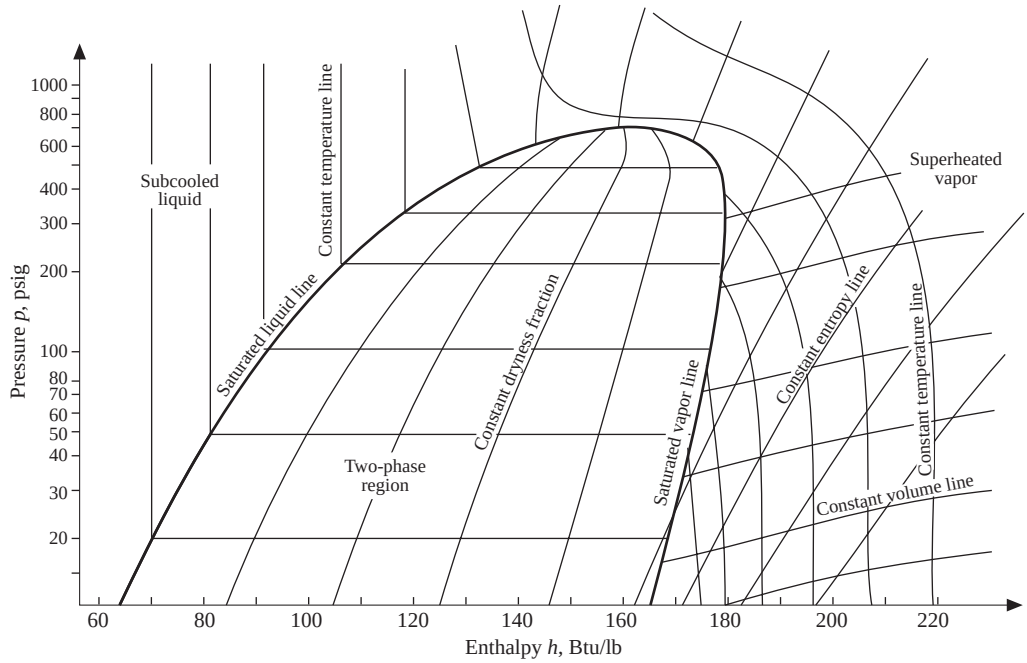


FIGURE 9.4.1 Skeleton of pressure-enthalpy diagram for R-22.

line is nearly vertical in the subcooling region, horizontal in the two-phase region, and curved down sharply in the superheated region.

In the two-phase region, a given saturated pressure determines the saturated temperature and vice versa. The constant-entropy line is curved upward to the right-hand side in the superheated region. Each kind of refrigerant has its own *p-h* diagram.

Refrigeration Processes in an Ideal Single-Stage Cycle

An *ideal cycle* has isentropic compression, and pressure losses in the pipeline, valves, and other components are neglected. All refrigeration cycles covered in this section are ideal. Single stage means a single stage of compression.

There are four refrigeration processes in an ideal single-stage vapor compression cycle, as shown in Figure 9.4.2(a) and (b):

1. Isothermal evaporation process 4–1 — The refrigerant evaporates completely in the evaporator and produces refrigeration effect q_{rf}, in Btu/lb:

$$q_{rf} = (h_1 - h_4) \tag{9.4.1}$$

   where h_1,h_4 = enthalpy of refrigerant at state points 1 and 4, respectively, Btu/lb.
2. Isentropic compression process 1–2 — Vapor refrigerant is extracted by the compressor and compressed isentropically from point 1 to 2. The work input to the compressor W_{in}, in Btu/lb, is

$$W_{in} = (h_2 - h_1) \tag{9.4.2}$$

   where h_2 = enthalpy of refrigerant at state point 2, Btu/lb.
   The greater the difference in temperature/pressure between the condensing pressure p_{con} and evaporating pressure p_{ev}, the higher will be the work input to the compressor.

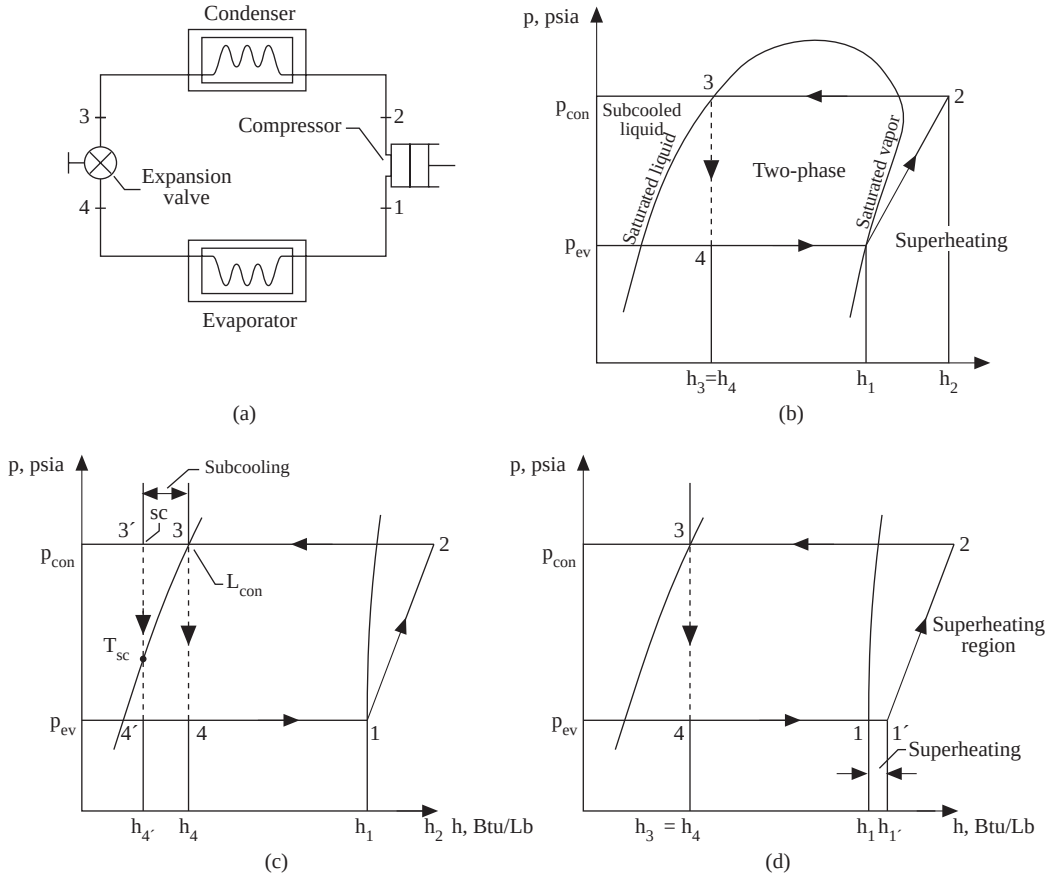


FIGURE 9.4.2 A single-stage ideal vapor compression refrigeration cycle: (a) schematic diagram, (b) *p-h* diagram, (c) subcooling, and (d) superheating.

3. Isothermal condensation process 2–3 — Hot gaseous refrigerant discharged from the compressor is condensed in the condenser into liquid, and the latent heat of condensation is rejected to the condenser water or ambient air. The heat rejection during condensation, q_{2-3}, in Btu/lb, is

$$-q_{2-3} = \left(h_2 - h_3\right) \tag{9.4.3}$$

where h_3 = enthalpy of refrigerant at state point 3, Btu/lb.

4. Throttling process 3–4 — Liquid refrigerant flows through a throttling device (e.g., an expansion valve, a capillary tube, or orifices) and its pressure is reduced to the evaporating pressure. A portion of the liquid flashes into vapor and enters the evaporator. This is the only irreversible process in the ideal cycle, usually represented by a dotted line. For a throttling process, assuming that the heat gain from the surroundings is negligible:

$$h_3 = h_4 \tag{9.4.4}$$

The mass flow rate of refrigerant $\mathring{m}_r$, in lb/min, is

$$\mathring{m}_r = q_{rc}/60q_{rf} \tag{9.4.5}$$

where q_{rc} = refrigeration capacity of the system, Btu/hr.
The ideal single-stage vapor compression refrigeration cycle on a *p-h* diagram is divided into two pressure regions: high pressure (p_{con}) and low pressure (p_{ev}).

Coefficient of Performance of Refrigeration Cycle

The *coefficient of performance* (COP) is a dimensionless index used to indicate the performance of a thermodynamic cycle or thermal system. The magnitude of COP can be greater than 1.

- If a *refrigerator* is used to produce a refrigeration effect, COP_{ref} is

$$COP_{ref} = q_{rf} / W_{in} \tag{9.4.6}$$

- If a *heat pump* is used to produce a useful heating effect, its performance denoted by COP_{hp} is

$$COP_{hp} = q_{2-3} / W_{in} \tag{9.4.7}$$

- For a heat recovery system when both refrigeration and heating effects are produced, the COP_{hr} is denoted by the ratio of the sum of the absolute values of q_{rf} and $q_{2\text{-}3}$ to the work input, or

$$COP_{hr} = \left(\left|q_{rf}\right| + \left|q_{2-3}\right|\right) / W_{in} \tag{9.4.8}$$

Subcooling and Superheating

Condensed liquid is often subcooled to a temperature lower than the saturated temperature corresponding to the condensing pressure p_{con}, in psia or psig, as shown in Figure 9.4.2(c). *Subcooling* increases the refrigeration effect to $q_{rf.sc}$ as shown in Figure 9.4.2(c):

$$q_{rf.sc} = \left(h_{4'} - h_1\right) > \left(h_4 - h_1\right) \tag{9.4.9}$$

The enthalpy of subcooled liquid refrigerant h_{sc} approximately equals the enthalpy of the saturated liquid refrigerant at subcooled temperature $h_{s.sc}$, both in Btu/lb:

$$h_{sc} = h_{3'} = h_{4'} = h_{1.con} - c_{pr}\left(T_{con} - T_{sc}\right) \approx h_{s.sc} \tag{9.4.10}$$

where $h_{3'}$, $h_{4'}$ = enthalpy of liquid refrigerant at state points 3′ and 4′ respectively, Btu/lb
$h_{l.con}$ = enthalpy of saturated liquid at condensing temperature, Btu/lb
c_{pr} = specific heat of liquid refrigerant at constant pressure, Btu/lb °F
T_{con} = condensing temperature or saturated temperature of liquid refrigerant at condensing pressure, °F
T_{sc} = temperature of subcooled liquid refrigerant, °F

The purpose of *superheating* is to prevent liquid refrigerant flooding back into the compressor and causing slugging damage as shown in Figure 9.4.2(d). The degree of superheating depends mainly on the types of refrigerant feed, construction of the suction line, and type of compressor. The state point of vapor refrigerant after superheating of an ideal system must be at the evaporating pressure with a specific degree of superheat and can be plotted on a *p-h* diagram for various refrigerants.

Refrigeration Cycle of Two-Stage Compound Systems with a Flash Cooler

A *multistage system* employs more than one compression stage. Multistage vapor compression systems are classified as compound systems and cascade systems. A *compound system* consists of two or more

compression stages connected in series. It may have one high-stage compressor (higher pressure) and one low-stage compressor (lower pressure), several compressors connected in series, or two or more impellers connected internally in series and driven by the same motor.

The *compression ratio* R_{com} is defined as the ratio of discharge pressure p_{dis} to the suction pressure at the compressor inlet p_{suc}:

$$R_{com} = p_{dis} / p_{suc} \tag{9.4.11}$$

Compared to a single-stage system, a multistage has a smaller compression ratio and higher compression efficiency for each stage of compression, greater refrigeration effect, lower discharge temperature at the high-stage compressor, and greater flexibility. At the same time, a multistage system has a higher initial cost and more complicated construction.

The pressure between the discharge pressure of the high-stage compressor and the suction pressure of the low-stage compressor of a multistage system is called *interstage pressure* p_i, in psia. Interstage pressure for a two-stage system is usually determined so that the compression ratios are nearly equal between two stages for a higher COP. Then the interstage pressure is

$$p_i = \sqrt{\left(p_{con} p_{ev}\right)} \tag{9.4.12}$$

where p_{con}, p_{ev} = condensing and evaporating pressures, psia.

For a multistage system of n stages, then, the compression ratio of each stage is

$$R_{com} = \left(p_{con} / p_{suc}\right)^{1/n} \tag{9.4.13}$$

Figure 9.4.3(a) shows a schematic diagram and Figure 9.4.3(b) the refrigeration cycle of a two-stage compound system with a flash cooler. A *flash cooler*, sometimes called an economizer, is used to subcool the liquid refrigerant to the saturated temperature corresponding to the interstage pressure by vaporizing a portion of the liquid refrigerant in the flash cooler.

Based on the principle of heat balance, the fraction of evaporated refrigerant, x, or quality of the mixture in the flash cooler is

$$x = \left(h_{5'} - h_8\right) / \left(h_7 - h_8\right) \tag{9.4.14}$$

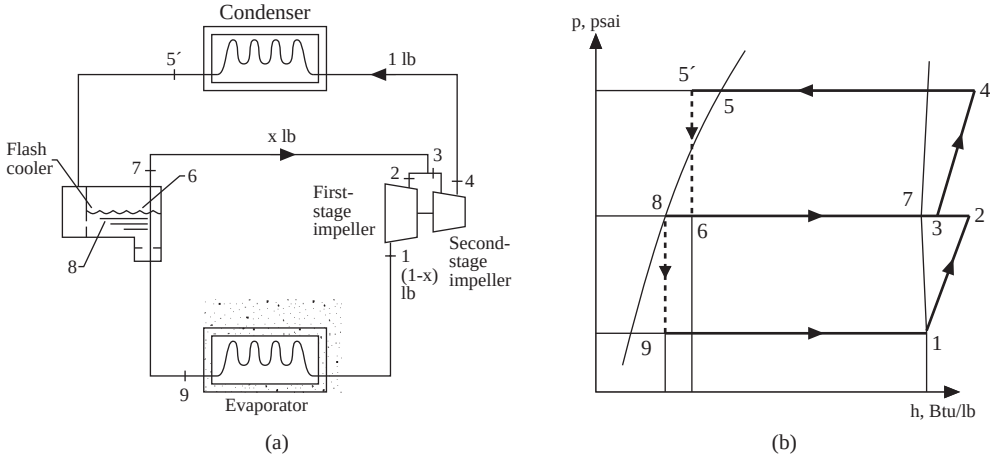


FIGURE 9.4.3 Two-stage compound system with a flash cooler: (a) schematic diagram and (b) refrigeration cycle.

where $h_{5'}$, h_7, h_8 = enthalpy of the refrigerant at state points 5′, 7, and 8, respectively, Btu/lb. The coefficient of performance of the refrigeration cycle of a two-stage compound system with a flash cooler, COP_{ref}, is given as

$$COP_{ref} = q_{rf}/W_{in} = (1-x)(h_1 - h_9)/[(1-x)(h_2 - h_1) + (h_4 - h_3)] \quad (9.4.15)$$

where h_1, h_2, h_3, h_4, h_9 = enthalpy of refrigerant at state points 1, 2, 3, 4, and 9, respectively, Btu/lb. The mass flow rate of refrigerant flowing through the condenser, $\mathring{m}_r$, in lb/min, can be calculated as

$$\mathring{m}_r = q_{rc}/60q_{rf} \quad (9.4.16)$$

Because a portion of liquid refrigerant is flashed into vapor in the flash cooler and goes directly to the second-stage impeller inlet, less refrigerant is compressed in the first-stage impeller. In addition, the liquid refrigerant in the flash cooler is cooled to the saturated temperature corresponding to the interstage temperature before entering the evaporator, which significantly increases the refrigeration effect of this compound system. Two-stage compound systems with flash coolers are widely used in large central air-conditioning systems.

Cascade System Characteristics

A *cascade system* consists of two independently operated single-stage refrigeration systems: a lower system that maintains a lower evaporating temperature and produces a refrigeration effect and a higher system that operates at a higher evaporating temperature as shown in Figure 9.4.4(a) and (b). These two separate systems are connected by a *cascade condenser* in which the heat released by the condenser in the lower system is extracted by the evaporator in the higher system.

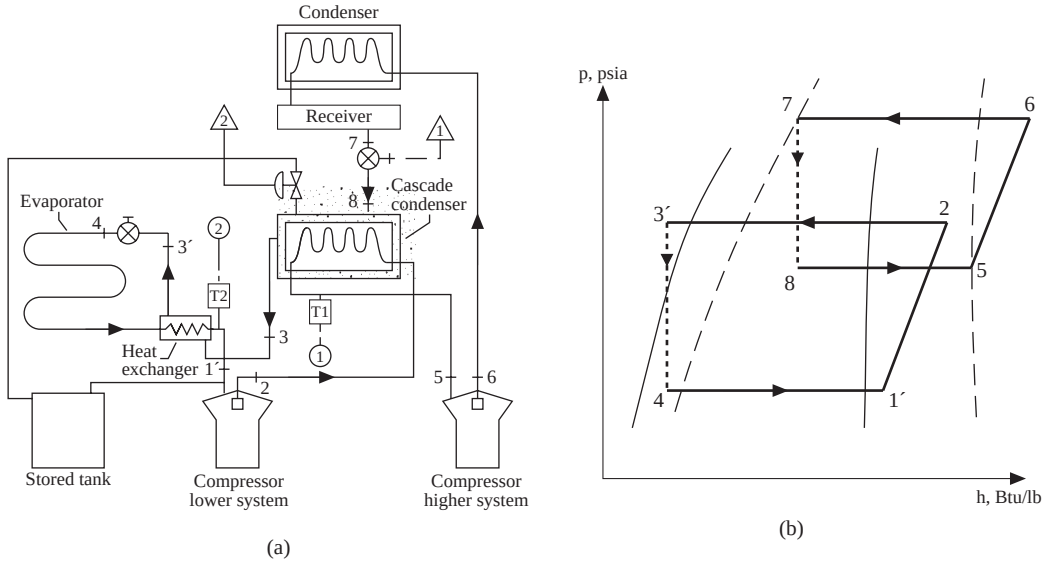


FIGURE 9.4.4 Cascade system: (a) schematic diagram and (b) refrigeration cycle.

A heat exchanger is often used between the liquid refrigerant from the condenser and the vapor refrigerant leaving the evaporator of the lower system. When the system is shut down in summer, a relief valve connected to a stored tank should be used to relieve the higher pressure of refrigerant at the higher storage temperature.

The main advantages of a cascade system compared with a compound system are that different refrigerants, oils, and equipment can be used for the lower and higher systems. Disadvantages are the overlap of the condensing temperature of the lower system and the evaporating temperature of the higher system because of the heat transfer in the cascade condenser and a more complicated system.

The refrigeration effect q_{rf} of the cascade system is

$$q_{rf} = \left(h_1 - h_4\right) \tag{9.4.17}$$

where h_1, h_4 = enthalpy of the refrigerant leaving and entering the evaporator of the lower system, Btu/lb. The total work input to the compressors in both higher and lower systems W_{in}, in Btu/lb, can be calculated as

$$W_{in} = \left(h_2 - h_{1'}\right) + \mathring{m}_h\left(h_6 - h_5\right) \Big/ \mathring{m}_1 \tag{9.4.18}$$

where h_2 = enthalpy of refrigerant discharged from the compressor of the lower system
$h_{1'}$ = enthalpy of the vapor refrigerant leaving the heat exchanger
h_6, h_5 = enthalpy of the refrigerant leaving and entering the high-stage compressor
$\mathring{m}_h, \mathring{m}_1$ = mass flow rate of the refrigerant of the higher and lower systems, respectively

The coefficient of performance of a cascade system COP_{ref} is

$$COP_{ref} = q_{rf} / W_{in} = \mathring{m}_1\left(h_1 - h_4\right) \Big/ \left[\mathring{m}_1\left(h_2 - h_{1'}\right) + \mathring{m}_h\left(h_6 - h_5\right)\right] \tag{9.4.19}$$

9.5 Outdoor Design Conditions and Indoor Design Criteria

Outdoor Design Conditions

In principle, the capacity of air-conditioning equipment should be selected to offset or compensate for the space load so that indoor design criteria can be maintained if the outdoor weather does not exceed the design values. Outdoor and indoor design conditions are used to calculate the design space loads. In energy use calculations, hour-by-hour outdoor climate data of a design day should be adopted instead of summer and winter design values.

ASHRAE Handbook 1993 Fundamentals (Chapter 24 and 27) and *Wang's Handbook of Air Conditioning and Refrigeration* (Chapter 7) both list tables of climate conditions for the U.S. and Canada based on the data from the National Climate Data Center (NCDC), U.S. Air Force, U.S. Navy, and Canadian Atmospheric Environment Service. In these tables:

- *Summer design dry bulb temperature* in a specific location $T_{o.s}$, in °F, is the rounded higher integral number of the statistically determined summer outdoor design dry bulb temperature $T_{o.ss}$ so that the average number of hours of occurrence of outdoor dry bulb temperature T_o higher than $T_{o.ss}$ during June, July, August, and September is less than 1, 2.5, or 5% of the total number of hours in these summer months (2928 hr). The data are an average of 15 years. An occurrence of less than 2.5% of 2928 hr of summer months, that is, $0.025 \times 2928 = 73$ hr, is most widely used.
- *Summer outdoor mean coincident wet bulb temperature* $T'_{o.s}$, in °F, is the mean of all the wet bulb temperatures at the specific summer outdoor design dry bulb temperature $T_{o.s}$ during the summer months.
- *Summer outdoor 2.5% design wet bulb temperature* is the design wet bulb temperature that has an average annual occurrence of $T'_o > T'_{o.s}$ less than 73 hr. This design value is often used for evaporative cooling design.
- *Mean daily range*, in °F, is the difference between the average daily maximum and the average daily minimum temperature during the warmest month.
- In *ASHRAE Handbook 1993 Fundamentals, solar heat gain factors* (SHGFs), in Btu/h.ft^2, are the average solar heat gain per hour during cloudless days through double-strength sheet (DSA) glass. The *maximum SHGFs* are the maximum values of SHGFs on the 21st of each month for a specific latitude.
- *Winter outdoor design dry bulb temperature* $T_{o.w}$, in °F, is the rounded lower integral value of the statically determined winter outdoor design temperature $T_{o.ws}$, so that the annual average number of hours of occurrence of outdoor temperature $T_o > T_{o.ws}$ is equal to or exceeds 99%, or 97.5% of the total number of hours in December, January, and February (2160 hr).

A *degree day* is the difference between a base temperature and the mean daily outdoor air temperature $T_{o.m}$ for any one day, in °F. The total numbers of heating degree days HDD65 and cooling degree days CDD65 referring to a base temperature of 65°F per annum are

$$\text{HDD65} = \sum_{n=1} \left(65 - T_{o.m}\right)$$
$$\text{CDD65} = \sum_{m=1} \left(T_{o.m} - 65\right) \qquad (9.5.1)$$

where n = number of days for which $T_{o.m} < 65$°F
m = number of days for which $T_{o.m} > 65$°F

Indoor Design Criteria and Thermal Comfort

Indoor design criteria, such as space temperature, humidity, and air cleanliness, specify the requirements for the indoor environment as well as the quality of an air-conditioning or HVAC&R project.

The human body requires energy for physical and mental activity. This energy comes from the oxidation of food. The rate of heat release from the oxidation process is called the *metabolic rate*, expressed in met (1 met = 18.46 Btu/h.ft^2). The metabolic rate depends mainly on the intensity of the physical activity of the human body. Heat is released from the human body by two means: *sensible heat exchange* and *evaporative heat loss*. Experience and experiments all show that there is thermal comfort only under these conditions:

- Heat transfer from the human body to the surrounding environment causes a steady state of thermal equilibrium; that is, there is no heat storage in the body core and skin surface.
- Evaporative loss or regulatory sweating is maintained at a low level.

The physiological and environmental factors that affect the thermal comfort of the occupants in an air-conditioned space are mainly:

1. Metabolic rate M determines the amount of heat that must be released from the human body.
2. *Indoor air temperature* T_r and *mean radiant temperature* T_{rad}, both in °F. The *operating temperature* T_o is the weighted sum of T_r and T_{rad}. T_{rad} is defined as the temperature of a uniform black enclosure in which the surrounded occupant would have the same radiative heat exchange as in an actual indoor environment. T_r affects both the sensible heat exchange and evaporative losses, and T_{rad} affects only sensible heat exchange. In many indoor environments, $T_{rad} \approx T_r$.
3. Relative humidity of the indoor air φ_r, in %, which is the primary factor that influences evaporative heat loss.
4. Air velocity of the indoor air v_r, in fpm, which affects the heat transfer coefficients and therefore the sensible heat exchange and evaporative loss.
5. *Clothing insulation* R_{cl}, in clo (1 clo = 0.88 h.ft^2.°F/Btu), affects the sensible heat loss. Clothing insulation for occupants is typically 0.6 clo in summer and 0.8 to 1.2 clo in winter.

Indoor Temperature, Relative Humidity, and Air Velocity

For comfort air-conditioning systems, according to ANSI/ASHRAE Standard 55-1981 and ASHRAE/IES Standard 90.1-1989, the following indoor design temperatures and air velocities apply for conditioned spaces where the occupant's activity level is 1.2 met, indoor space relative humidity is 50% (in summer only), and $T_r = T_{rad}$:

	Clothing insulation (clo)	**Indoor temperature (°F)**	**Air velocity (fpm)**
Winter	0.8–0.9	69–74	<30
Summer	0.5–0.6	75–78	<50

If a suit jacket is the clothing during summer for occupants, the summer indoor design temperature should be dropped to 74 to 75°F.

Regarding the indoor humidity:

1. Many comfort air-conditioning systems are not equipped with humidifiers. Winter indoor relative humidity should not be specified in such circumstances.
2. When comfort air-conditioning systems are installed with humidifiers, ASHRAE/IES Standard 90.1-1989 requires that the humidity control prevent "the use of fossil fuel or electricity to produce humidity in excess of 30% ... or to reduce relative humidity below 60%."

3. Indoor relative humidity should not exceed 75% to avoid increasing bacterial and viral populations.
4. For air-conditioning systems that use flow rate control in the water cooling coil, space indoor relative humidity may be substantially higher in part load than at full load.

Therefore, for comfort air-conditioning systems, the recommended indoor relative humidities, in %, are

	Tolerable range	**Preferred value**
Summer	30–65	40–50
Winter		
With humidifier		25–30
Without humidifier		Not specified

In surgical rooms or similar health care facilities, the indoor relative humidity is often maintained at 40 to 60% year round.

Indoor Air Quality and Outdoor Ventilation Air Requirements

According to the National Institute for Occcupational Safety and Health (NIOSH), 1989, the causes of indoor air quality complaints in buildings are inadequate outdoor ventilation air, 53%; indoor contaminants, 15%; outdoor contaminants, 10%; microbial contaminants, 5%; construction and furnishings, 4%; unknown and others, 13%. For space served by air-conditioning systems using low- and medium-efficiency air filters, according to the U.S. Environmental Protection Agency (EPA) and Consumer Product Safety Commission (CPSC) publication "A Guide to Indoor Air Quality" (1988) and the field investigations reported by Bayer and Black (1988), *indoor air contaminants* may include some of the following:

1. *Total particulate concentration.* This concentration comprises particles from building materials, combustion products, mineral fibers, and synthetic fibers. In February 1989, the EPA specified the allowable indoor concentration of particles of 10 μm and less in diameter (which penetrate deeply into lungs) as:

   50 $\mu g/m^3$ (0.000022 grain/ft^3), 1 year
   150 $\mu g/m^3$ (0.000066 grain/ft^3), 24 hr

   In these specifications, "1 year" means maximum allowable exposure per day over the course of a year.
2. *Formaldehyde and organic gases*. Formaldehyde is a colorless, pungent-smelling gas. It comes from pressed wood products, building materials, and combustion. Formaldehyde causes eye, nose, and throat irritation as well as coughing, fatigue, and allergic reactions. Formaldehyde may also cause cancer. Other organic gases come from building materials, carpeting, furnishings, cleaning materials, etc.
3. *Radon*. Radon, a colorless and odorless gas, is released by the decay of uranium from the soil and rock beneath buildings, well water, and building materials. Radon and its decay products travel through pores in soil and rock and infiltrate into buildings along the cracks and other openings in the basement slab and walls. Radon at high levels causes lung cancer. The EPA believes that levels in most homes can be reduced to 4 pCi/l (picocuries per liter) of air. The estimated national average is 1.5 pCi/l, and levels as high as 200 pCi/l have been found in houses.
4. *Biologicals*. These include bacteria, fungi, mold and mildew, viruses, and pollen. They may come from wet and moist walls, carpet furnishings, and poorly maintained dirty air-conditioning systems and may be transmitted by people. Some biological contaminants cause allergic reactions, and some transmit infectious diseases.

5. *Combustion products*. These include environmental tobacco smoke, nitrogen dioxide, and carbon monoxide. *Environmental tobacco* smoke from cigarettes is a discomfort factor to other persons who do not smoke, especially children. Nicotine and other tobacco smoke components cause lung cancer, heart disease, and many other diseases. *Nitrogen dioxide* and *carbon monoxide* are both combustion products from unvented kerosene and gas space heaters, woodstoves, and fireplaces. Nitrogen dioxide (NO_2) causes eye, nose, and throat irritation; may impair lung function; and increases respiratory infections. Carbon monoxide (CO) causes fatigue at low concentrations; impaired vision, headache, and confusion at higher concentrations; and is fatal at very high concentrations. Houses without gas heaters and gas stoves may have CO levels varying from 0.5 to 5 parts per million (ppm).
6. *Human bioeffluents*. These include the emissions from breath including carbon dioxide exhaled from the lungs, body odors from sweating, and gases emitted as flatus.

There are three basic means of reducing the concentration of indoor air contaminants and improving indoor air quality: (1) eliminate or reduce the source of air pollution, (2) enhance the efficiency of air filtration, and (3) increase the ventilation (outdoor) air intake. Dilution of the concentrations of indoor contaminants by outdoor ventilation air is often the simple and cheapest way to improve indoor air quality. The amount of outdoor air required for metabolic oxidation is rather small.

Abridged outdoor air requirements listed in ANSI/ASHRAE Standard 62-1989 are as follows:

Applications	cfm/person
Hotels, conference rooms, offices	20
Retail stores	0.2–0.3 cfm/ft^2
Classrooms, theaters, auditoriums	15
Hospital patient rooms	25

These requirements are based on the analysis of dilution of CO_2 as the representative human bioeffluent to an allowable indoor concentration of 1000 ppm. Field measurements of daily maximum CO_2 levels in office buildings reported by Persily (1993) show that most of them were within the range 400 to 820 ppm. The quality of outdoor air must meet the EPA's National Primary and Secondary Ambient Air Quality Standards, some of which is listed below:

	Long-term concentration			**Short-term concentration**		
Pollutants	**μg/m^3**	**ppm**	**Exposure**	**μg/m^3**	**ppm**	**Exposure**
Particulate	75		1 year	260		24 hr
SO_2	80	0.03	1 year	365	0.14	24 hr
CO				40,000	35	1 hr
				10,000	9	8 hr
NO_2	100	0.055	1 year			
Lead	1.5		3 months			

Here exposure means average period of exposure.

If unusual contaminants or unusually strong sources of contaminants are introduced into the space, or recirculated air is to replace part of the outdoor air supply for occupants, then acceptable indoor air quality is achieved by controlling known and specific contaminants. This is called an indoor air quality procedure. Refer to ANSI/ASHRAE Standard 62-1989 for details.

Clean Rooms

Electronic, pharmaceutical, and aerospace industries and operating rooms in hospitals all need strict control of air cleanliness during manufacturing and operations. According to ASHRAE Handbook 1991 HVAC Applications, clean rooms can be classified as follows based on the particle count per ft^3:

Class	Particle size	
	0.5 μm and larger	5 μm and larger
	Particle count per ft^3 not to exceed	
1	1	0
10	10	0
100	100	
1000	1000	
10,000	10,000	65
100,000	100,000	700

For clean rooms, space temperature is often maintained at 72 ± 2°F and space humidity at 45 ± 5%. Here, ±2°F and ±5% are allowable tolerances. Federal Standard 209B specifies that the ventilation (outdoor air) rate should be 5 to 20% of the supply air.

Space Pressure Differential

Most air-conditioning systems are designed to maintain a slightly higher pressure than the surroundings, a positive pressure, to prevent or reduce infiltration and untreated air entering the space directly. For laboratories, restrooms, or workshops where toxic, hazardous, or objectional gases or contaminants are produced, a slightly lower pressure than the surroundings, a negative pressure, should be maintained to prevent or reduce the diffusion of these contaminants' exfiltrate to the surrounding area.

For comfort air-conditioning systems, the recommended pressure differential between the indoor and outdoor air is 0.02 to 0.05 in in. WG. WG indicates the pressure at the bottom of a top-opened water column of specific inches of height; 1 in. WG = 0.03612 psig.

For clean rooms, Federal Standard No. 209B, Clean Rooms and Work Stations Requirements (1973), specifies that the minimum positive pressure between the clean room and any adjacent area with lower cleanliness requirements should be 0.05 in. WG with all entryways closed. When the entryways are open, an outward flow of air is to be maintained to prevent migration of contaminants into the clean room. In comfort systems, the space pressure differential is usually not specified in the design documents.

Sound Levels

Noise is any unwanted sound. In air-conditioning systems, noise should be attenuated or masked with another less objectionable sound.

Sound power is the capability to radiate power from a sound source exited by an energy input. The intensity of sound power is the output from a sound source expressed in watts (W). Due to the wide variation of sound output at a range of 10^{20} to 1, it is more convenient to use a logarithmic expression to define a *sound power level* L_w, in dB:

$$L_w = 10\log\left(w/10^{-12}\ \text{W}\right)\ \text{re 1 pW} \tag{9.5.2}$$

where w = sound source power output, W.

The human ear and microphones are sound pressure sensitive. Similarly to the sound power level, the *sound pressure level* L_p, in dB, is defined as:

$$L_p = 20\log\left(p/2\times10^{-5}\ \text{Pa}\right)\ \text{re 20 μPa} \tag{9.5.3}$$

where p = sound pressure, Pa.

The sound power level of any sound source is a fixed output. It cannot be measured directly; it can only be calculated from the measured sound pressure level. The sound pressure level at any one point is affected by the distance from the source and the characteristics of the surroundings.

Human ears can hear frequencies from 20 Hz to 20 kHz. For convenience in analysis, sound frequencies are often subdivided into eight octave bands. An *octave* is a frequency band in which the frequency of the upper limit of the octave is double the frequency of the lower limit. An octave band is represented by its center frequency, such as 63, 125, 250, 500, 1000, 2000, 4000, and 8000 Hz. On 1000 Hz the octave band has a higher limit of 1400 Hz and a lower limit of 710 Hz. Human ears do not respond in the same way to low frequencies as to high frequencies.

The object of noise control in an air conditioned space is to provide background sound low enough that it does not interfere with the acoustical requirements of the occupants. The distribution of background sound should be balanced over a broad range of frequencies, that is, without whistle, hum, rumble, and beats.

The most widely used criteria for sound control are the noise criteria NC curves. The shape of NC curves is similar to the equal-loudness contour representing the response of the human ear. NC curves also intend to indicate the permissible sound pressure level of broad-band noise at various octave bands rated by a single NC curve. NC curves are practical and widely used.

Other criteria used are room criteria RC curves and A-weighted sound level, dBA. RC curves are similar to NC curves except that the shape of the RC curves is a close approximation to a balanced, bland-sounding spectrum. The A-weighted sound level is a single value and simulates the response of the human ear to sound at low sound pressure levels.

The following are abridged indoor design criteria, NC or RC range, listed in *ASHRAE Handbook 1987 Systems and Applications*:

Type of area	Recommended NC or RC range (dB)
Hotel guest rooms	30–35
Office	
Private	30–35
Conference	25–30
Open	30–35
Computer equipment	40–45
Hospital, private	25–30
Churches	25–30
Movie theaters	30–35

For industrial factories, if the machine noise in a period of 8 hr exceeds 90 dBA, Occupational Safety and Health Administration Standard Part 1910.95 requires the occupants to use personal protection equipment. If the period is shorter, the dBA level can be slightly higher. Refer to this standard for details.

9.6 Load Calculations

Space Loads

Space, Room, and Zone

Space indicates a volume or a site without partitions, or a partitioned room or a group of rooms. A *room* is an enclosed or partitioned space that is considered as a single load. An air-conditioned room does not always have an individual zone control system. A *zone* is a space of a single room or group of rooms having similar loads and operating characteristics. An air-conditioned zone is always installed with an individual control system. A typical floor in a building may be treated as a single zone space, or a *multizone space* of perimeter, interior, east, west, south, north, ... zones.

Space and equipment loads can be classified as:

1. *Space heat gain* q_e, in Btu/hr, is the rate of heat transfer entering a conditioned space from an external heat source or heat releases to the conditioned space from an internal source. The rate of sensible heat entering the space is called *sensible heat gain* q_{es}, whereas the rate of latent heat entering the space is called *latent heat gain* q_{el}. In most load calculations, the time interval is often 1 hr, and therefore q_e, q_{es}, and q_{el} are all expressed in Btu/hr.
2. *Space cooling load* or simply *cooling load* q_{rc}, also in Btu/hr, is the rate at which heat must be removed from a conditioned space to maintain a constant space temperature and an acceptable relative humidity. The sensible heat removed is called *sensible cooling load* q_{rs}, and the latent heat removed is called *latent cooling load* q_{rl}, both in Btu/hr.
3. *Space heat extraction rate* q_{ex}, in Btu/hr, is the rate at which heat is removed from the conditioned space. When the space air temperature is constant, $q_{ex} = q_{rc}$.
4. *Space heating load* q_{rh}, in Btu/hr, is the rate at which heat must be added to the conditioned space to maintain a constant temperature.
5. *Coil load* q_c, in Btu/hr, is the rate of heat transfer at the coil. The cooling *coil load* q_{cc} is the rate of heat removal from the conditioned air by the chilled water or refrigerant inside the coil. The *heating coil load* q_{ch} is the rate of heat energy addition to the conditioned air by the hot water, steam, or electric elements inside the coil.
6. *Refrigeration load* q_{rl}, in Btu/hr, is the rate at which heat is extracted by the evaporated refrigerant at the evaporator. For packaged systems using a DX coil, $q_{rl} = q_{cc}$. For central systems:

$$q_{rl} = q_{cc} + q_{pi} + q_{pu} + q_{s.t} \tag{9.6.1}$$

where q_{pi} = chilled water piping heat gain, Btu/hr
q_{pu} = pump power heat gain, Btu/hr
$q_{s.t}$ = storage tank heat gain, if any, Btu/hr

Heat gains q_{pi} and q_{pu} are usually about 5 to 10% of the cooling coil load q_{cc}.

Convective Heat and Radiative Heat

Heat enters a space and transfer to the space air from either an external source or an internal source is mainly in the form of *convective heat* and *radiative heat* transfer.

Consider radiative heat transfer, such as solar radiation striking the outer surface of a concrete slab as shown in Figure 9.6.1(a) and (b). Most of the radiative heat is absorbed by the slab. Only a small fraction is reflected. After the heat is absorbed, the outer surface temperature of the slab rises. If the slab and space air are in thermal equilibrium before the absorption of radiative heat, heat is convected from the outer surface of the slab to the space air as well as radiated to other surfaces. At the same time, heat is conducted from the outer surface to the inner part of the slab and stored there when the temperature of the inner part of the slab is lower than that of its outer surface. Heat convected from the outer surface of the concrete slab to the space air within a time interval forms the sensible cooling load.

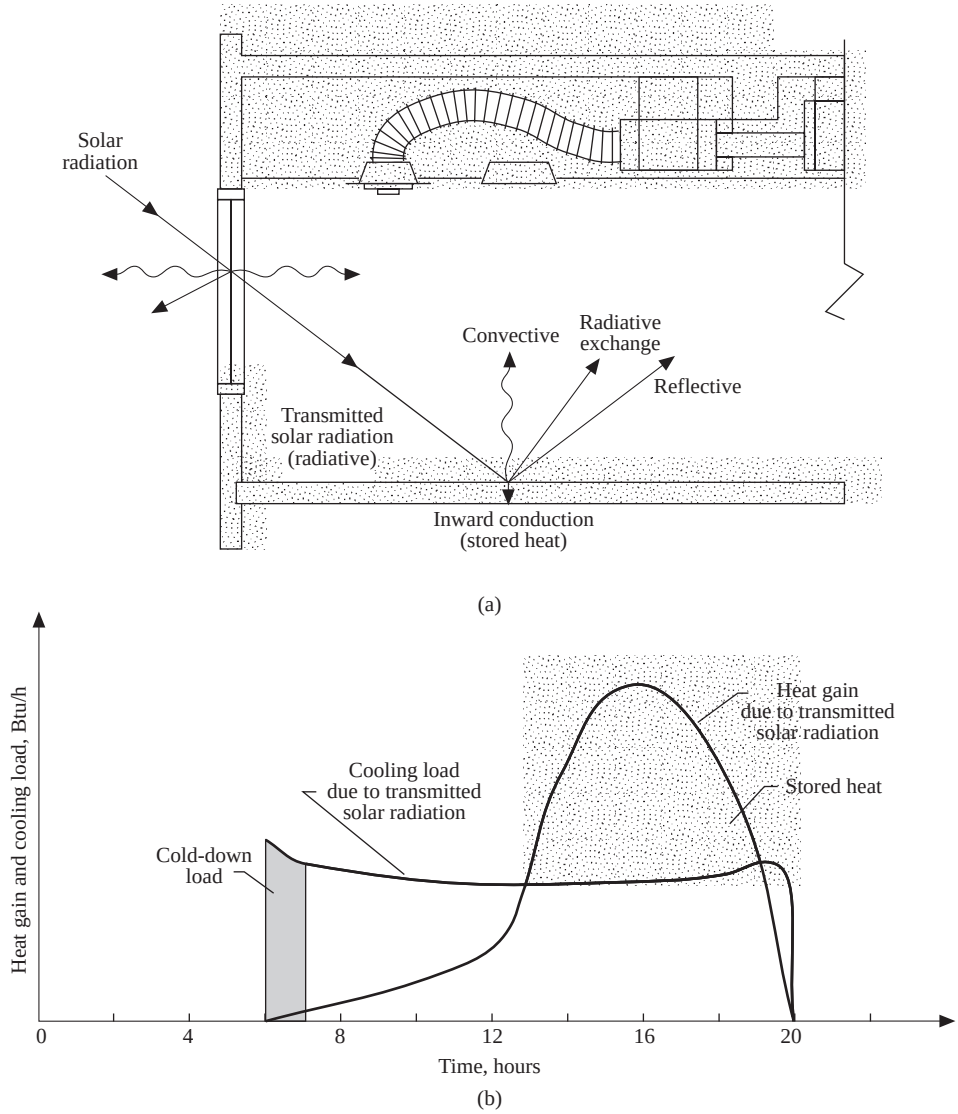


FIGURE 9.6.1 Solar heat gain from west window and its corresponding space cooling load for a night shutdown air system: (a) convective and radiative heat transfer and (b) heat gain and cooling load curves.

The sensible heat gain entering the conditioned space does not equal the sensible cooling load during the same time interval because of the stored heat in the building envelope. Only the convective heat becomes cooling load instantaneously. The sum of the convective heats from the outer surfaces, including the outer surfaces of the internal heat gains in a conditioned space, becomes cooling load. This phenomenon results in a smaller cooling load than heat gain, as shown in Figure 9.6.1(a) and (b). According to *ASHRAE Handbook 1993 Fundamentals*, the percentages of convective and radiative components of the sensible heat gains are as follows:

Sensible heat gains	Convective (%)	Radiative (%)
Solar radiation with internal shading	42	58
Fluorescent lights	50	50
Occupants	67	33
External wall, inner surface	40	60

Load Profile, Peak Load, and Block Load

A *load profile* shows the variation of space, zone, floor, or building load in a certain time period, such as a 24-hr day-and-night cycle. In a load profile, load is always plotted against time. The load profile depends on the outdoor climate as well as the space operating characteristics.

Peak load is the maximum cooling load in a load profile. *Block load* is the sum of the zone loads and floor loads at a specific time. The sum of the zone peak loads in a typical floor does not equal the block load of that floor because the zone peak loads may all not appear at the same time.

Moisture Transfer in Building Envelope

Moisture transfer takes place along two paths:

1. Moisture migrates in the building envelope in both liquid and vapor form. It is mainly liquid if the relative humidity of the ambient air exceeds 50%. Liquid flow is induced by capillary flow and moisture content gradient. Vapor diffusion is induced by vapor pressure gradients. *Moisture content* is defined as the ratio of the mass of moisture contained in a solid to the mass of bone-dry solid. During the migration, the moisture content and the vapor pressure are in equilibrium at a specific temperature and location.
2. Air leakage and its associated water vapor infiltrate or exfiltrate through the cracks, holes, and gaps between joints because of poor construction of the building envelope. The driving potential of this air leakage and associated water vapor is the pressure differential across the building envelope. If the insulating material is of open-cell structure, air leakage and associated water vapor may penetrate the perforated insulating board through cracks and gaps. Condensation, even freezing, will occur inside the perforated insulation board if the temperature of the board is lower than the dew point of the leaked air or the freezing point of the water.

In most comfort air-conditioning systems, usually only the space temperature is controlled within limits. A slight variation of the space relative humidity during the operation of the air system is often acceptable. Therefore, the store effect of moisture is ignored except in conditioned spaces where both temperature and relative humidity need to be controlled or in a hot and humid area where the air system is operated at night shutdown mode. In most cases, latent heat gain is considered equal to latent cooling load instantaneously. For details refer to Wang's *Handbook of Air Conditioning and Refrigeration* (1993), Chapters 6 and 7.

Cooling Load Calculation Methodology

Basic considerations include the following:

- It is assumed that equations of heat transfer for cooling load calculation within a time interval are linear. It is also assumed that the superposition principle holds. When a number of changes occur simultaneously in the conditioned space, they will proceed as if independent of each other. The total change is the sum of the responses caused by the individual changes.
- Space load calculations are often performed by computer-aided design (CAD), with market-available software like DOE-2.1D, TRACE-600, and Carrier E20-II Loads.
- Peak load calculations evaluate the maximum load to size and select the equipment. The energy analysis program compares the total energy use in a certain period with various alternatives in order to determine the optimum one.
- The methodology of various cooling load calculations is mainly due to their differences in the conversion of space radiative heat gains into space cooling loads. Convective heat, latent heat, and sensible heat gains from infiltration are all equal to cooling loads instantaneously.
- Space cooling load is used to calculate the supply volume flow rate and to determine the size of the air system, ducts, terminals, and diffusers. The coil load is used to determine the size of the

cooling coil and the refrigeration system. Space cooling load is a component of the cooling coil load.

The Rigorous Approach

The rigorous approach to the calculation of the space cooling load consists of (1) finding the inside surface temperatures of the building structures that enclose the conditioned space due to heat balance at time t and (2) calculating the sum of the convective heats transferred from these surfaces as well as from the occupants, lights, appliances, and equipment in the conditioned space at time t.

The inside surface temperature of each surface $T_{i,t}$, in °F, can be found from the following simultaneous heat balance equations:

$$q_{i,t} = \left[h_{ci}\left(T_{r,t} - T_{i,t}\right) + \sum_{j=1}^{m} g_{ij}\left(T_{j,t} - T_{i,t}\right) \right] A_i + S_{i,t} + L_{i,t} + P_{i,t} + E_{i,t} \qquad (9.6.2)$$

where h_{ci} = convective heat transfer coefficient, Btu/hr.ft^2.°F
g_{ij} = radiative heat transfer factor between inside surface i and inside surface j, Btu/hr.ft^2.°F
$T_{r,t}$ = space air temperature at time t, °F
$T_{i,t}$, $T_{j,t}$ = average temperature of inside surfaces i and j at time t, °F
A_i = area of inside surface i, ft^2
$S_{i,t}$, $L_{i,t}$, $P_{i,t}$, $E_{i,t}$ = solar radiation transmitted through windows and radiative heat from lights, occupants, and equipment absorbed by inside surface i at time t, Btu/hr

In Equation (9.6.2), $q_{i,t}$, in Btu/hr, is the conductive heat that comes to surface i at time t because of the temperature excitation on the outer opposite surface of i. This conductive heat can be found by solving the partial differential equations or by numerical solutions. The number of inside surfaces i is usually equal to 6, and surface i is different from j so that radiative exchange can proceed. $q_{i,t}$ could also be expressed in Btu/min or even Btu/sec.

The space sensible cooling load q_{rs}, in Btu/hr, is the sum of the convective heat from the inside surfaces, including the convective heat from the inner window glass due to the absorbed solar radiation and the infiltration:

$$q_{rs} = \left[\sum_{i=1}^{6} h_{ci}\left(T_{i,t} - T_{r,t}\right) \right] A_i + 60\dot{V}_{if}\rho_o c_{pa}\left(T_{o,t} - T_{r,t}\right) + S_{c,t} + L_{c,t} + P_{c,t} + E_{c,t} \qquad (9.6.3)$$

where $\dot{V}_{if}$ = volume flow rate of infiltrated air, cfm
ρ_o = air density of outdoor air, lb/ft^3
c_{pa} = specific heat of moist air, Btu/lb.°F
$T_{o,t}$ = temperature of outdoor air at time t, °F
$S_{c,t}$, $L_{c,t}$, $P_{c,t}$, $E_{c,t}$ = heat convected from the windows, lights, occupants, and equipment, Btu/hr

Equations (9.6.2) and (9.6.3), and partial differential equations to determine conductive heat $q_{i,t}$ must be solved simultaneously. Using a rigorous approach to find the space cooling load requires numerous computer calculations. It is laborious and time consuming. The rigorous approach is impractical and is suitable for research work only.

Transfer Function Method (TFM)

The *transfer function* of a system relates its output in Laplace transform Y to its input in Laplace transform G by a ratio K, that is,

$$K = Y/G = \left(v_o + v_1 z^{-1} + v_2 z^{-2} + \ldots\right) \Big/ \left(1 + w_1 z^{-1} + w_2 z^{-2} + \ldots\right)$$
$$z = e^{s\Delta} \tag{9.6.4}$$

where Δ = time interval.

In Equation (9.6.4), K, Y, and G are all expressed in z-transforms of the time series function. Coefficients v_n and w_n are called *transfer function coefficients*, or weighting factors. *Weighting factors* are used to weight the importance of the effect of current and previous heat gains as well as the previous space sensible cooling load on the current space sensible cooling load $q_{rs,t}$. Then, the output $q_{rs,t}$ can be related to the input, the space sensible heat gain $q_{es,t}$, through $q_{rs,t} = Kq_{es,t}$.

Mitalas and Stevenson (1967) and others developed a method for determining the transfer function coefficients of a zone of given geometry and details of the calculated space heat gains and the previously known space sensible cooling load through rigorous computation or through tests and experiments. In DOE 2.1A (1981) software for custom weighting factors (tailor made according to a specific parametric zone) is also provided. Sowell (1988) and Spitler et al. (1993) expanded the application of TFM to zones with various parameters: zone geometry, different types of walls, roof, floor, ceilings, building material, and mass locations. Mass of construction is divided into light construction, 30 lb/ft^2 of floor area; medium construction, 70 lb/ft^2; and heavy construction, 130 lb/ft^2. Data are summarized into groups and listed in tabular form for user's convenience.

Cooling Load Temperature Difference/Solar Cooling Load Factor/Cooling Load Factor (CLTD/SCL/CLF) Method

The *CLTD/SCL/CLF method* is a one-step simplification of the transfer function method. The space cooling load is calculated directly by multiplying the heat gain q_e with CLTD, SCL, or CLF instead of first finding the space heat gains and then converting into space cooling loads through the room transfer function. In the CLTD/SCL/CLF method, the calculation of heat gains is the same as in the transfer function method.

The CLTD/SCL/CLF method was introduced by Rudoy and Duran (1975). McQuiston and Spitler (1992) recommended a new SCL factor. In 1993 they also developed the CLTD and CLF data for different zone geometries and constructions.

Finite Difference Method

Since the development of powerful personal computers, the finite difference or numerical solution method can be used to solve transient simultaneous heat and moisture transfer in space cooling load calculations. Wong and Wang (1990) emphasized the influence of moisture stored in the building structure on the cool-down load during the night shut-down operating mode in locations where the summer outdoor climate is hot and humid. The finite difference method is simple and clear in concept as well as more direct in computation than the transfer function method. Refer to Wang's *Handbook of Air Conditioning and Refrigeration* for details.

Total Equivalent Temperature Differential/Time Averaging (TETD/TA) Method

In this method, the heat gains transmitted through external walls and roofs are calculated from the Fourier series solution of one-dimensional transient heat conduction. The conversion of space heat gains to space cooling loads takes place by (1) averaging the radiative heat gains to the current and successive hours according to the mass of the building structure and experience and (2) adding the instantaneous convective fraction and the allocated radiative fraction in that time period. The TETD/TA method is simpler and more subjective than the TFM.

Conduction Heat Gains

Following are the principles and procedures for the calculation of space heat gains and their conversion to space cooling loads by the TFM. TFM is the method adopted by software DOE-2.1D and is also one

of the computing programs adopted in TRACE 600. Refer to *ASHRAE Handbook 1993 Fundamentals*, Chapters 26 and 27, for detail data and tables.

Surface Heat Transfer Coefficients

ASHRAE Handbook 1993 Fundamentals adopts a constant *outdoor heat transfer coefficient* h_o = 3.0 Btu/hr.ft^2.°F and a constant *inside heat transfer coefficient* h_i = 1.46 Btu/hr.ft^2.°F during cooling load calculations. However, many software programs use the following empirical formula (TRACE 600 *Engineering Manual*):

$$h_{o.t} = 2.00 + 0.27v_{\text{wind}.t} \tag{9.6.5}$$

where $h_{o.t}$ = h_o at time t, Btu/hr.ft^2.°F
$v_{\text{wind}.t}$ = wind velocity at time t, ft/sec

Coefficient h_i may be around 2.1 Btu/hr.ft^2.°F when the space air system is operating, and h_i drops to only about 1.4 Btu/hr.ft^2.°F when the air system is shut down. The *R value*, expressed in hr.ft^2.°F/Btu, is defined as the reciprocal of the overall heat transfer coefficient *U* value, in Btu/hr.ft^2.°F. The *R* value is different from the thermal resistance R^*, since $R^* = 1/UA$, which is expressed in hr.°F/Btu.

Sol-air temperature T_{sol} is a fictitious outdoor air temperature that gives the rate of heat entering the outer surface of walls and roofs due to the combined effect of incident solar radiation, radiative heat exchange with the sky vault and surroundings, and convective heat exchange with the outdoor air. T_{sol}, in °F, is calculated as

$$T_{sol} = T_o + \alpha I_t / h_o - \varepsilon \Delta R / h_o \tag{9.6.6}$$

where T_o = outdoor air temperature, °F
α = absorptance of incident solar radiation on outer surface
ϵ = hemispherical emittance of outer surface, assumed equal to 1
I_t = total intensity of solar radiation including diffuse radiation, Btu/hr.ft^2

Tabulated sol-air temperatures that have been calculated and listed in *ASHRAE Handbook 1993 Fundamentals* are based on the following: if $\epsilon = 1$ and $h_o = 3$ Btu/hr.ft^2.°F, $\Delta R/h_o$ is –7°F for horizontal surfaces and 0°F for vertical surfaces, assuming that the long-wave radiation from the surroundings compensates for the loss to the sky vault.

External Wall and Roof

The sensible heat gain through a wall or roof at time t, $q_{e,t}$, in Btu/hr, can be calculated by using the sol-air temperature at time t-$n\delta$, $T_{sol,t-n\delta}$, in °F, as the outdoor air temperature and a constant indoor space air temperature T_r, in °F, as

$$q_{e,t} = A\left[\sum_{n=0} b_n T_{sol,t-n\delta} - \sum_{n=1} d_n \left(q_{e,t-n\delta}/A\right)\right] - T_r \sum_{n=0} c_n \tag{9.6.7}$$

where A = surface area of wall or roof, ft^2
$q_{e,t-n\delta}$ = heat gain through wall or roof at time t-$n\delta$, Btu/hr
n = number of terms in summation

In Equation (9.6.7), b_n, c_n, and d_n are *conduction transfer function* coefficients.

Ceiling, Floor, and Partition Wall

If the variation of the temperature of the adjacent space T_{aj}, in °F, is small compared with the differential ($T_{aj} - T_r$), or even when T_{aj} is constant, the heat gain to the conditioned space through ceiling, floor, and partition wall $q_{aj,t}$, in Btu/hr, is

$$q_{aj,t} = UA\left(T_{aj} - T_r\right) \tag{9.6.8}$$

where U = overall heat transfer coefficient of the ceiling, floor, and partition (Btu/hr.ft².°F).

Heat Gain through Window Glass

Shading Devices. There are two types of shading devices: indoor shading devices and outdoor shading devices. *Indoor shading devices* increase the reflectance of incident radiation. *Venetian blinds* and draperies are the most widely used indoor shading devices. Most horizontal venetian blinds are made of plastic, aluminum, or rigid woven cloth slats spaced 1 to 2 in. apart. Vertical venetian blinds with wider slats are also used. *Draperies* are made of fabrics of cotton, rayon, or synthetic fibers. They are usually loosely hung, wider than the window, often pleated, and can be drawn open and closed as needed. Draperies also increase thermal resistance in winter.

External shading devices include overhangs, side fins, louvers, and pattern grilles. They reduce the sunlit area of the window glass effectively and therefore decrease the solar heat gain. External shading devices are less flexible and are difficult to maintain.

Shading Coefficient (SC). The shading coefficient is an index indicating the glazing characteristics and the associated indoor shading device to admit solar heat gain. SC of a specific window glass and shading device assembly at a summer design solar intensity and outdoor and indoor temperatures can be calculated as

$$\text{SC} = \frac{\text{solar heat gain of specific type of window glass assembly}}{\text{solar heat gain of reference (DSA) glass}} \tag{9.6.9}$$

Double-strength sheet glass (DSA) has been adopted as the reference glass with a transmittance of $\tau = 0.86$, reflectance $\rho = 0.08$, and absorptance $\alpha = 0.06$.

Solar Heat Gain Factors (SHGFs). SHGFs, in Btu/hr.ft², are the average solar heat gains during cloudless days through DSA glass. SHGF_{max} is the maximum value of SHGF on the 21st day of each month for a specific latitude as listed in *ASHRAE Handbook 1993 Fundamentals*. At high elevations and on very clear days, the actual SHGF may be 15% higher.

Heat Gain through Window Glass. There are two kinds of heat gain through window glass: heat gain due to the solar radiation transmitted and absorbed by the window glass, $q_{es,t}$, and conduction heat gain due to the outdoor and indoor temperature difference, $q_{ec,t}$, both in Btu/hr:

$$q_{es,t} = \left(A_s \times \text{SHGF} \times \text{SC}\right) + \left(A_{sh} \times \text{SHGF}_{sh} \times \text{SC}\right) \tag{9.6.10}$$

where A_s, A_{sh} = sunlit and shaded areas of the glass (ft²).

In Equation (9.6.10), SHGF_{sh} represents the SHGF of the shaded area of the glass having only diffuse radiation. Generally, the SHGF of solar radiation incident on glass facing north without direct radiation can be considered as SHGF_{sh}. Because the mass of window glass is small, its heat storage effect is often neglected; then the conduction heat gain at time t is

$$q_{ec,t} = U_g\left(A_s + A_{sh}\right)\left(T_{o,t} - T_r\right) \tag{9.6.11}$$

where U_g = overall heat transfer coefficient of window glass, Btu/hr.ft^2.°F
$T_{o,t}$ = outdoor air temperature at time t, °F

The inward heat transfer due to the solar radiation absorbed by the glass and the conduction heat transfer due to the outdoor and indoor temperature difference is actually combined. It is simple and more convenient if they are calculated separately.

Internal Heat Gains

Internal heat gains are heat released from the internal sources.

People

The sensible heat gain and latent heat gain per person, $q_{es,t}$ and $q_{el,t}$, both in Btu/hr, are given as

$$q_{es,t} = N_t q_{os}$$
$$q_{el,t} = N_t q_{ol} \tag{9.6.12}$$

where N_t = number of occupants in the conditioned space at time t.

Heat gains q_{os} and q_{ol} depend on the occupant's metabolic level, whether the occupant is an adult or a child, male or female, and the space temperature. For a male adult seated and doing very light work, such as system design by computer in an office of space temperature 75°F, q_{os} and q_{ol} are both about 240 Btu/hr.

Lights

Heat gain in the conditioned space because of the electric lights, $q_{e.li}$, in Btu/hr, is calculated as

$$q_{e.li} = \sum_{n=1}\left[3.413 W_{li} N_{li} F_{ul} F_{al}\left(1 - F_{lp}\right)\right] = 3.413 W_{lA} A_{fl} \tag{9.6.13}$$

where W_{li} = watt input to each lamp, W
W_{lA} = lighting power density, W/ft^2
A_{fl} = floor area of conditioned space, ft^2
N_{li}, n = number of lamps (each type) and number of types of electric lamp
F_{al}, F_{ul} = use factor of electric lights and allowance factor for ballast loss, usually taken as 1.2
F_{lp} = heat gain carried away by the return air to plenum or by exhaust device; it varies from 0.2 to 0.5

Machines and Appliances

The sensible and latent heat gains in the conditioned space from machines and appliances, $q_{e.ap}$ and $q_{l.ap}$, both in Btu/hr, can be calculated as

$$q_{e.ap} = 3.413 W_{ap} F_{ua} F_{load} F_{ra} = q_{is} F_{ua} F_{ra}$$
$$q_{l.ap} = q_{il} F_{ua} \tag{9.6.14}$$

where W_{ap} = rated power input to the motor of the machines and appliances, W
F_{ua} = use factor of machine and appliance
F_{load} = ratio of actual load to the rated power
F_{ra} = radiation reduction factor because of the front shield of the appliance
q_{is}, q_{il} = sensible and latent heat input to the appliance, Btu/hr

Infiltration

Infiltration is the uncontrolled inward flow of unconditioned outdoor air through cracks and openings on the building envelope because of the pressure difference across the envelope. The pressure difference is probably caused by wind pressure, stack effect due to outdoor–indoor temperature difference, and the operation of an air system(s).

Today new commercial buildings have their external windows well sealed. If a positive pressure is maintained in the conditioned space when the air system is operating, infiltration is normally considered as zero.

When the air system is shut down, or for hotels, motels, and high-rise residential buildings, ASHRAE/IES Standard 90.1-1989 specifies an infiltration of 0.038cfm/ft^2 of gross area of the external wall, 0.15 air change per hour (ach) for the perimeter zone.

When exterior windows are not well sealed, the outdoor wind velocity is high at winter design conditions, or there is a door exposed to the outdoors directly, an infiltration rate of 0.15 to 0.4 ach for perimeter zone should be considered.

When the volume flow rate of infiltration is determined, the sensible heat gain due to infiltration $q_{s.if}$ and latent heat gain due to infiltration $q_{l.if}$, in Btu/hr, are

$$q_{s.if} = 60\,\mathring{V}_{if}\,\rho_o c_{pa}\left(T_o - T_r\right)$$
$$q_{l.if} = 60\times 1060\,\mathring{V}_{if}\,\rho_o\left(w_o - w_r\right) \tag{9.6.15}$$

where $\mathring{V}_{if}$ = volume flow rate of infiltration, cfm
ρ_o = air density of outdoor air, lb/ft^3
c_{pa} = specific heat of moist air, 0.243 Btu/lb.°F
w_o,w_r = humidity ratio of outdoor and space air, lb/lb
$h_{fg.58}$ = latent heat of vaporization, 1060 Btu/lb

Infiltration enters the space directly and mixes with space air. It becomes space cooling load instantaneously. Ventilation air is often taken at the AHU or PU and becomes sensible and latent coil load components.

Conversion of Heat Gains into Cooling Load by TFM

The space sensible cooling load $q_{rs,t}$, in Btu/hr, is calculated as

$$q_{rs,t} = q_{s-e,t} + q_{s-c,t} \tag{9.6.16}$$

where $q_{s\text{-}e,t}$ = space sensible cooling load converted from heat gains having radiative and convective components, Btu/hr
$q_{s\text{-}c,t}$ = space sensible cooling load from convective heat gains, Btu/hr

Based on Equation (9.6.4), space sensible cooling load $q_{s\text{-}e,t}$ can be calculated as

$$q_{s-e,t} = \sum_{i=1}\left(v_o q_{e,t} + v_1 q_{e,t-\delta} + v_2 q_{e,t-2\delta} + \ldots\right) - \left(w_1 q_{r,t-\delta} + w_2 q_{r,t-2\delta} + \ldots\right) \tag{9.6.17}$$

where $q_{e,t}$, $q_{e,t\text{-}\delta}$, $q_{e,t\text{-}2\delta}$ = space sensible heat gains having both radiative and convective heats at time t, t-δ, and t-2δ, Btu/hr
$q_{r,t\text{-}\delta}$, $q_{r,t\text{-}2\delta}$ = space sensible cooling load at time t-δ, and t-2δ, Btu/hr

In Equation (9.6.17), v_n and w_n are called *room transfer function coefficients* (RTFs). RTF is affected by parameters like zone geometry; wall, roof, and window construction; internal shades; zone location; types of building envelope; and air supply density. Refer to RTF tables in *ASHRAE Handbook 1993 Fundamentals*, Chapter 26, for details.

Space sensible cooling load from convective heat gains can be calculated as

$$q_{s-c,t} = \sum_{k=1} q_{ec,t} \tag{9.6.18}$$

where $q_{ec,t}$ = each of k space sensible heat gains that have convective heat gains only (Btu/hr). Space latent cooling load at time t, $q_{rl,t}$, in Btu/hr, can be calculated as

$$q_{rl,t} = \sum_{m=1} q_{el,t} \tag{9.6.19}$$

where $q_{el,t}$ = each of m space latent heat gains (Btu/hr).

Space Air Temperature and Heat Extraction Rate

At equilibrium, the space sensible heat extraction rate at time t, $q_{xs,t}$, is approximately equal to the space sensible cooling load, $q_{rs,t}$, when zero offset proportional plus integral or proportional-integral-derivative control mode is used. During the cool-down period, the sensible heat extraction rate of the cooling coil or DX coil at time t, $q_{xs,t}$, or the sensible cooling coil load, in Btu/hr, is greater than the space sensible cooling load at time t and the space temperature T_r, in °F then drops gradually. According to *ASHRAE Handbook 1993 Fundamentals*, the relationship between the space temperature T_r and the sensible heat extraction rate $q_{xs,t}$ can be expressed as

$$\sum_{i=0}^{1} p_i \left(q_{xs,t} - q_{rs,t-i\delta}\right) = \sum_{i=0}^{2} g_i \left(T_r - T_{r,t-i\delta}\right) \tag{9.6.20}$$

where $q_{rs,t\text{-}i\delta}$ = space sensible cooling load calculated on the basis of constant space air temperature T_r, Btu/hr
$T_{r,t\text{-}i\delta}$ = space temperature at time t-iδ

In Equation (9.6.20), p_i, g_i are called *space air transfer function coefficients*. Space air temperature T_r can be considered an average reference temperature within a time interval.

Cooling Coil Load

Cooling coil load q_{cc}, in Btu/hr, can be calculated from Equation (9.3.11). The sensible and latent cooling coil loads can then be calculated as

$$\begin{aligned} q_{cs} &= 60 \mathring{V}_s \rho_s c_{pa} \left(T_m - T_{cc}\right) \\ q_{cl} &= 60 \mathring{V}_s \rho_s \left(w_m - w_{cc}\right) \end{aligned} \tag{9.6.21}$$

where q_{cs}, q_{cl} = sensible and latent cooling coil load, Btu/hr
T_m, T_{cc} = air temperature of the mixture and leaving the cooling coil, °F
w_m, w_{cc} = humidity ratio of the air mixture and air leaving the cooling coil, lb/lb

Heating Load

The *space heating load* or simply *heating load* is always the possible maximum heat energy that must be added to the conditioned space at winter design conditions to maintain the indoor design temperature. It is used to size and select the heating equipment. In heating load calculations, solar heat gain, internal heat gains, and the heat storage effect of the building envelope are usually neglected for reliability and simplicity.

Normally, space heating load q_{rh}, in Btu/hr, can be calculated as

$$\begin{aligned} q_{rl} &= q_{trans} + q_{if.s} + q_{ma} + q_{hu} \\ &= \left[\left(\sum_{n=1} AU\right) + 60\,\mathring{V}_{if}\,\rho_o c_{pa}\,\mathring{m}_m\,c_m\right]\left(T_r - T_o\right) + \mathring{m}_w\,h_{fg.58} \end{aligned} \tag{9.6.22}$$

where A = area of the external walls, roofs, glasses, and floors, ft²
U = overall heat transfer coefficient of the walls, roofs, glasses, and floors, Btu/hr.ft².°F
c_m, $\mathring{m}_m$ = specific heat and mass flow rate of the cold product entering the space per hour, Btu/lb.°F and lb/hr
$\mathring{m}_w$ = mass flow rate of water evaporated for humidification, lb/hr
$h_{fg.58}$ = latent heat of vaporization at 58°F, 1060 Btu/lb

In Equation 9.88, q_{trans} indicates the *transmission loss* through walls, roofs, glasses, and floors, $q_{if.s}$ the *sensible infiltration heat loss*, q_{ma} the heat required to heat the cold product that enters the conditioned space, and q_{hu} the heat required to raise the space air temperature when water droplets from a space humidifier are evaporated in the conditioned space. For details, refer to *ASHRAE Handbook 1993 Fundamentals*.

9.7 Air Handling Units and Packaged Units

Terminals and Air Handling Units

A *terminal unit*, or *terminal*, is a device or equipment installed directly in or above the conditioned space to cool, heat, filter, and mix outdoor air with recirculating air. Fan-coil units, VAV boxes, fan-powered VAV boxes, etc. are all terminals.

An *air handling unit* (AHU) handles and conditions the air, controls it to a required state, and provides motive force to transport it. An AHU is the primary equipment of the air system in a central air-conditioning system. The basic components of an AHU include a supply fan with a fan motor, a water cooling coil, filters, a mixing box except in a makeup AHU unit, dampers, controls, and an outer casing. A return or relief fan, heating coil(s), and humidifier are optional depending on requirements. The supply volume flow rate of AHUs varies from 2000 to about 60,000 cfm.

AHUs are classified into the followings groups according to their structure and location.

Horizontal or Vertical Units

Horizontal AHUs have their fan, coils, and filters installed at the same level as shown in Figure 9.7.1(a). They need more space and are usually for large units. In *vertical units*, as shown in Figure 9.7.1(b), the supply fan is installed at a level higher than coils and filters. They are often comparatively smaller than horizontal units.

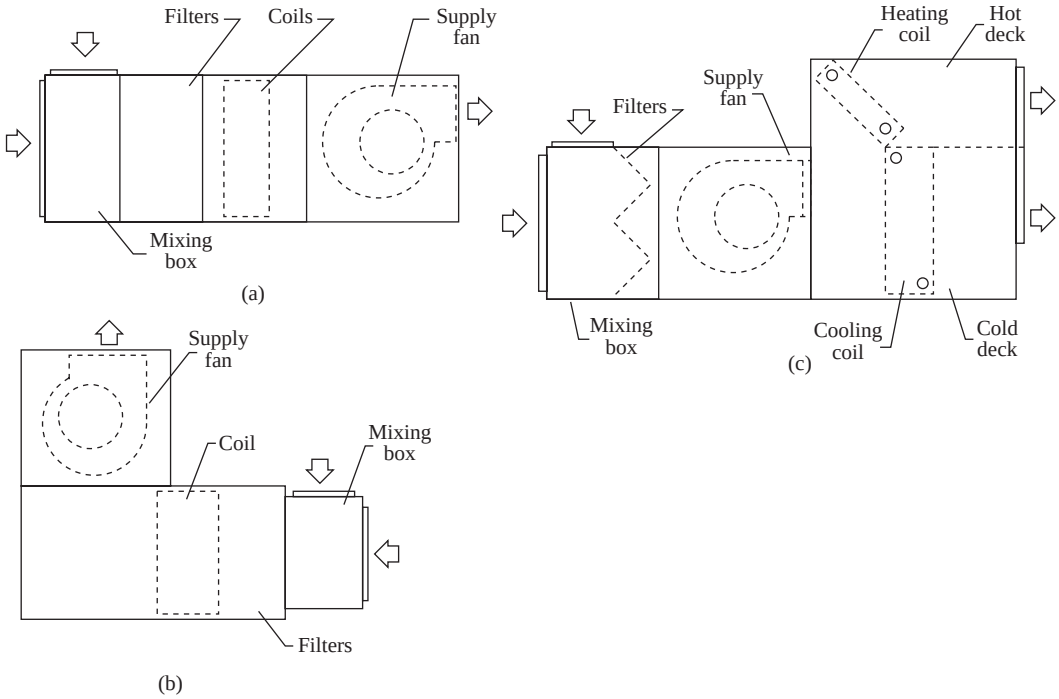


FIGURE 9.7.1 Type of air handling units: (a) horizontal draw-through unit, (b) vertical draw-through unit, and (c) multizone blow-through unit.

Draw-Through or Blow-Through Units

In a *draw-through unit*, as shown in Figure 9.7.1(a), the supply fan is located downstream of the coils. Air is evenly distributed over the coil section, and the fan discharge can easily be connected to a supply duct of nearly the same air velocity. In a *blow-through unit*, as shown in Figure 9.7.1(c), the supply fan

is located upstream of the coils. It usually has hot and cold decks with discharge dampers connected to warm and cold ducts, respectively.

Factory-Fabricated and Field Built-Up Units

Factory-fabricated units are standard in construction and layout, low in cost, of higher quality, and fast in installation. *Field built-up units* or *custom-built units* are more flexible in construction, layout, and dimensions than factory-built standardized units.

Rooftop and Indoor Units

A *rooftop AHU*, sometimes called a penthouse unit, is installed on the roof and will be completely weatherproof. An *indoor AHU* is usually located in a fan room or ceiling and hung like small AHU units.

Make-Up Air and Recirculating Units

A *make-up AHU*, also called a primary-air unit, is used to condition outdoor air entirely. It is a once-through unit. There is no return air and mixing box. *Recirculating units* can have 100% outdoor air intake or mixing of outdoor air and recirculating air.

Packaged Units

A *packaged unit* (PU) is a self-contained air conditioner. It conditions the air and provides it with motive force and is equipped with its own heating and cooling sources. The packaged unit is the primary equipment in a packaged air-conditioning system and is always equipped with a DX coil for cooling, unlike an AHU. R-22, R-134a, and others are used as refrigerants in packaged units. The portion that handles air in a packaged unit is called an *air handler* to distinguish it from an AHU. Like an AHU, an indoor air handler has an indoor fan, a DX coil (indoor coil), filters, dampers, and controls. Packaged units can be classified according to their place of installation: rooftop, indoor, and split packaged units.

Rooftop Packaged Units

A *rooftop packaged unit* is mounted on the roof of the conditioned space as shown in Figure 9.7.2. From the types of heating/cooling sources provided, rooftop units can be subdivided into:

- Gas/electric rooftop packaged unit, in which heating is provided by gas furnace and cooling by electric power-driven compressors.
- Electric/electric rooftop packaged unit, in which electric heating and electric power-driven compressors provide heating and cooling.
- Rooftop packaged heat pump, in which both heating and cooling are provided by the same refrigeration system using a four-way reversing valve (heat pump) in which the refrigeration flow changes when cooling mode is changed to heating mode and vice versa. Auxiliary electric heating is provided if necessary.

Rooftop packaged units are single packaged units. Their cooling capacity may vary from 3 to 220 tons with a corresponding volume flow rate of 1200 to 80,000 cfm. Rooftop packaged units are the most widely used packaged units.

Indoor Packaged Units

An *indoor packaged unit* is also a single packaged and factory-fabricated unit. It is usually installed in a fan room or a machinery room. A small or medium-sized indoor packaged unit could be floor mounted directly inside the conditioned space with or without ductwork. The cooling capacity of an indoor packaged unit may vary from 3 to 100 tons and volume flow rate from 1200 to 40,000 cfm.

Indoor packaged units are also subdivided into:

- Indoor packaged cooling units

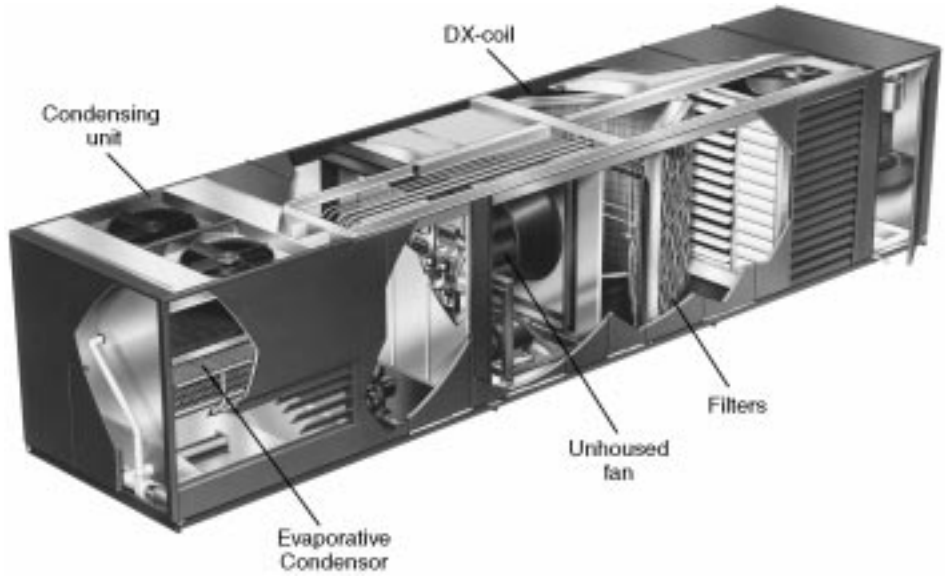


FIGURE 9.7.2 A cut view of a rooftop package unit. (Source: Mammoth, Inc. Reprinted by permission.)

- Indoor packaged cooling/heating units, in which heating may be provided from a hot water heating coil, a steam heating coil, and electric heating
- Indoor packaged heat pumps

Indoor packaged units have either an air-cooled condenser on the rooftop or a shell-and-tube or double-tube water-cooled condenser inside the unit.

Split Packaged Units

A *split packaged unit* consists of two separate pieces of equipment: an indoor air handler and an outdoor condensing unit. The indoor air handler is often installed in the fan room. Small air handlers can be ceiling hung. The condensing unit is usually located outdoors, on a rooftop or podium or on the ground.

A split packaged unit has its compressors and condenser in its outdoor condensing unit, whereas an indoor packaged unit usually has its compressors indoors. The cooling capacity of split packaged units varies from 3 to 75 tons and the volume flow rate from 1200 to 30,000 cfm.

Rating Conditions and Minimum Performance

Air Conditioning and Refrigeration Institute (ARI) Standards and ASHRAE/IES Standard 90.1-1989 specified the following rating indices:

- Energy efficiency ratio (EER) is the ratio of equipment cooling capacity, in Btu/hr, to the electric input, in W, under rating conditions.
- SEER is the seasonal EER, or EER during the normal annual usage period.
- IPLV is the integrated part-load value. It is the summarized single index of part-load efficiency of PUs based on weighted operations at several load conditions.
- HSPF is the heating seasonal performance factor. It is the total heating output of a heat pump during its annual usage period for heating, in Btu, divided by the total electric energy input to the heat pump during the same period, in watt-hours.

According to ARI standards, the minimum performance for air-cooled, electrically operated single packaged units is

	q_{rc} (Btu/hr)	T_o (°F)	EER	T_o (°F)	IPLV
Air-cooled	<65,000	95	9.5		
	65,000 ≤ q_{rc} < 135,000	95	8.9	80	8.3
	135,000 ≤ q_{rc} < 760,000		8.5		7.5

For water- and evaporatively cooled packaged units including heat pumps, refer to ASHRAE/IES Standard 90.1-1989 and also ARI Standards.

Coils

Coils, Fins, and Water Circuits

Coils are indirect contact heat exchangers. Heat transfer or heat and mass transfer takes place between conditioned air flowing over the coil and water, refrigerant, steam, or brine inside the coil for cooling, heating, dehumidifying, or cooling/dehumidifying. Chilled water, brine, and refrigerants that are used to cool and dehumidify the air are called *coolants*. Coils consist of tubes and external fins arranged in rows along the air flow to increase the contact surface area. Tubes are usually made of copper; in steam coils they are sometimes made of steel or even stainless steel. Copper tubes are staggered in 2, 3, 4, 6, 8, or up to 10 rows.

Fins are extended surfaces often called *secondary surfaces* to distinguish them from the *primary surfaces*, which are the outer surfaces of the tubes. Fins are often made from aluminum, with a thickness F_t = 0.005 to 0.008 in., typically 0.006 in. Copper, steel, or sometimes stainless steel fins are also used. Fins are often in the form of continuous plate fins, corrugated plate fins to increase heat transfer, crimped spiral or smooth spiral fins that may be extruded from the aluminum tubes, and spine pipes, which are shaved from the parent aluminum tubes. Corrugated plate fins are most widely used.

Fin spacing S_f is the distance between two fins. *Fin density* is often expressed in fins per inch and usually varies from 8 to 18 fins/in.

In a water cooling coil, *water circuits* or *tube feeds* determine the number of water flow passages. The greater the finned width, the higher the number of water circuits and water flow passages.

Direct Expansion (DX) Coil

In a *direct expansion coil*, the refrigerant, R-22, R-134a, or others, is evaporated and expanded directly inside the tubes to cool and dehumidify the air as shown in Figure 9.7.3(a). Refrigerant is fed to a distributor and is then evenly distributed to various copper tube circuits typically 0.375 in. in diameter. Fin density is usually 12 to 18 fins/in. and a four-row DX coil is often used. On the inner surface of the copper tubes, microfins, typically at 60 fins/in. and a height of 0.008 in., are widely used to enhance the boiling heat transfer.

Air and refrigerant flow is often arranged in a combination of counterflow and cross flow and the discharge header is often located on the air-entering side. Refrigerant distribution and loading in various circuits are critical to the coil's performance. Vaporized vapor refrigerant is superheated 10 to 20°F in order to prevent any liquid refrigerant from flooding back to the reciprocating compressors and damaging them. Finally, the vapor refrigerant is discharged to the suction line through the header.

For comfort air-conditioning systems, the evaporating temperature of refrigerant T_{ev} inside the tubes of a DX coil is usually between 37 and 50°F. At such a temperature, the surface temperature of the coil is often lower than the dew point of the entering air. Condensation occurs at the coil's outside surface, and the coil becomes a wet coil. A condensate *drain pan* is necessary for each vertically banked DX coil, and a trap should be installed to overcome the negative pressure difference between the air in the coil section and the ambient air.

Face velocity of the DX coil v_a, in fpm, is closely related to the blow-off of the water droplets of the condensate, the heat transfer coefficients, the air-side pressure drop, and the size of the air system. For corrugated fins, the upper limit is 600 fpm, with an air-side pressure drop of 0.20 to 0.30 in. WG/row.

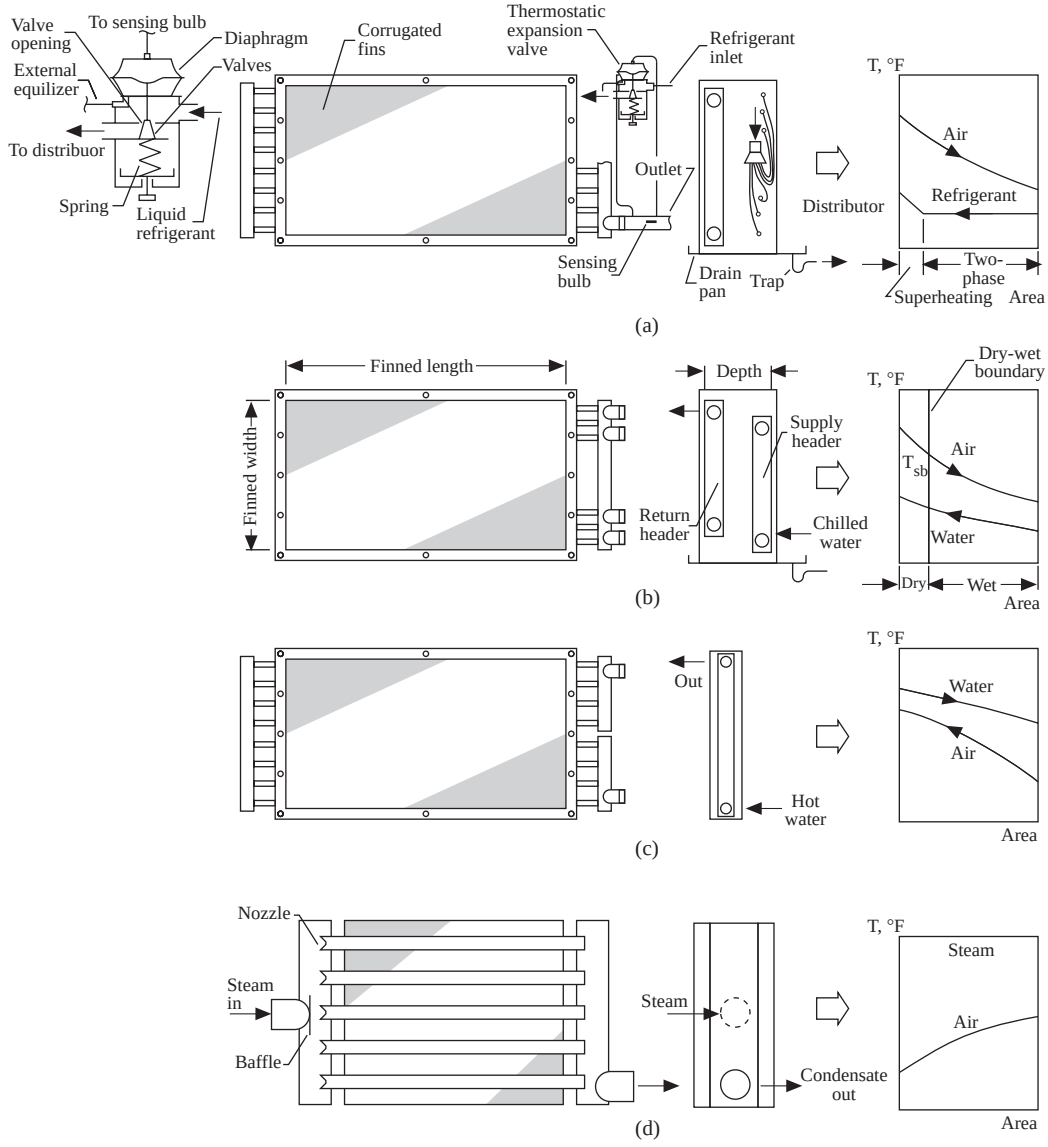


FIGURE 9.7.3 Types of coils: (a) direct expansion coil, (b) water cooling coil, (c) water heating coil, and (d) steam heating coil.

A large DX coil is often divided into two refrigerant sections, each with its own expansion valve, distributor, and discharge header.

For a packaged unit of a specific model, size, face velocity and condition of entering air and outdoor air, the DX coil's cooling capacities in nominal tons, number of rows, and fin density are all fixed values.

Water Cooling Coils — Dry–Wet Coils

In a water cooling coil, chilled water at a temperature of 40 to 50°F, brine, or glycol-water at a temperature of 34 to 40°F during cold air distribution enters the coil. The temperature of chilled water, brine, or glycol-water is usually raised 12 to 24°F before it leaves the water cooling coil.

The water tubes are usually copper tubes of 1/2 to 5/8 in. diameter with a tube wall thickness of 0.01 to 0.02 in. They are spaced at a center-to-center distance of 0.75 to 1.25 in. longitudinally and 1 to 1.5

in. transversely. These tubes may be staggered in 2, 3, 4, 6, 8, or 10 rows. Chilled water coils are often operated at a pressure of 175 to 300 psig.

As in a DX coil, the air flow and water flow are in a combination of counterflow and cross flow. The outer surface of a chilled water cooling coil at the air entering side T_{se} is often greater than the dew point of the entering air T''_{ae}, or $T_{se} > T''_{ae}$. The outer surface temperature of coil at the air leaving side T_{sl} may be smaller than T''_{ae}, or $T_{sl} < T''_{ae}$. Then the water cooling coil becomes a dry–wet coil with part of the dry surface on the air entering side and part of the wet surface on the air leaving side. A *dry–wet boundary* divides the dry and wet surfaces. At the boundary, the tube outer surface temperature $T_{sb} = T''_{ae}$ as shown in Figure 9.7.3(b). A condensate drain pan is necessary for a dry–wet coil.

A water cooling coil is selected from the manufacturer's selection program or from its catalog at (1) a dry and wet bulb of entering air, such as 80°F dry bulb and 67°F wet bulb; (2) an entering water temperature, such as 44 or 45°F; (3) a water temperature rise between 10 and 24°F; and (4) a coil face velocity between 400 and 600 fpm. The number of rows and fins per inch is varied to meet the required sensible and cooling coil load, in Btu/hr.

Water Cooling Coil–Dry Coil

When the temperature of chilled water entering the water cooling coil $T_{we} \geq T''_{ae}$, condensation will not occur on the outer surface of the coil. This coil becomes a sensible cooling–dry coil, and the humidity ratio of the conditioned air w_a remains constant during the sensible cooling process.

The construction of a sensible cooling–dry coil, such as material, tube diameter, number of rows, fin density, and fin thickness, is similar to that of a dry–wet coil except that a dry coil always has a poorer surface heat transfer coefficient than a wet coil, and therefore a greater coil surface area is needed; the maximum face velocity of a dry coil can be raised to $v_a \leq 800$ fpm; and the coil's outer surface is less polluted. The effectiveness of a dry coil $\in_{dry}$ is usually 0.55 to 0.7.

Water Heating Coil

The construction of a water heating coil is similar to that of a water cooling coil except that in water heating coils hot water is supplied instead of chilled water and there are usually fewer rows, only 2, 3, and 4 rows, than in water cooling coils. Hot water pressure in water heating coils is often rated at 175 to 300 psig at a temperature up to 250°F. Figure 9.7.3(c) shows a water heating coil.

Steam Heating Coil

In a steam heating coil, latent heat of condensation is released when steam is condensed into liquid to heat the air flowing over the coil, as shown in Figure 9.7.3(d). Steam enters at one end of the coil, and the condensate comes out from the opposite end. For more even distribution, a baffle plate is often installed after the steam inlet. Steam heating coils are usually made of copper, steel, or sometimes stainless steel.

For a steam coil, the coil core inside the casing should expand or contract freely. The coil core is also pitched toward the outlet to facilitate condensate drainage. Steam heating coils are generally rated at 100 to 200 psig at 400°F.

Coil Accessories and Servicing

Coil accessories include air vents, drain valves, isolation valves, pressure relief valves, flow metering valves, balancing valves, thermometers, pressure gauge taps, condensate drain taps, and even distribution baffles. They are employed depending on the size of the system and operating and serving requirements.

Coil cleanliness is important for proper operation. If a medium-efficiency air filter is installed upstream of the coil, dirt accumulation is often not a problem. If a low-efficiency filter is employed, dirt accumulation may block the air passage and significantly increase the pressure drop across the coil. Coils should normally be inspected and cleaned every 3 months in urban areas when low-efficiency filters are used. Drain pans should be cleaned every month to prevent buildup of bacteria and microorganisms.

Coil Freeze-Up Protection

Improper mixing of outdoor air and recirculating air in the mixing box of an AHU or PU may cause coil freeze-up when the outdoor air temperature is below 32°F. Outdoor air should be guided by a baffle plate and flow in an opposite direction to the recirculating air stream so that they can be thoroughly mixed without stratification.

Run the chilled water pump for the idle coil with a water velocity of 2.5 ft/sec, so that the cooling coil will not freeze when the air temperature drops to 32°F. A better method is to drain the water completely. For a hot water coil, it is better to reset the hot water temperature at part-load operation instead of running the system intermittently. A steam heating coil with inner distributor tubes and outer finned heating tubes provides better protection against freeze-up.

Air Filters

Air Cleaning and Filtration

Air cleaning is the process of removing airborne particles from the air. Air cleaning can be classified into air filtration and industrial air cleaning. Industrial air cleaning involves the removal of dust and gaseous contaminants from manufacturing processes as well as from the space air, exhaust air, and flue gas for air pollution control. In this section, only air filtration is covered.

Air filtration involves the removal of airborne particles presented in the conditioned air. Most of the airborne particles removed by air filtration are smaller than 1 μm, and the concentration of these particles in the airstream seldom exceeds 2 mg/m^3. The purpose of air filtration is to benefit the health and comfort of the occupants as well as meet the cleanliness requirements of the working area in industrial buildings.

An *air filter* is a kind of air cleaner that is installed in AHUs, PUs, and other equipment to filter the conditioned air by inertial impaction or interception and to diffuse and settle fine dust particles on the fibrous medium. The filter medium is the fabricated material that performs air filtration.

Operating performance of air filters is indicated by their:

- *Efficiency* or effectiveness of dust removal
- *Dust holding capacity* m_{dust}, which is the amount of dust held in the air filter, in grains/ft^2
- *Initial pressure drop* when the filter is clean Δp_{fi} and *final pressure drop* Δp_{ff} when the filter's m_{dust} is maximum, both in in. WG
- *Service life*, which is the operating period between Δp_{fi} and Δp_{ff}

Air filters in AHUs and PUs can be classified into low-, medium-, and high-efficiency filters and carbon activated filters.

Test Methods

The performance of air filters is usually tested in a test unit that consists of a fan, a test duct, the tested filter, two samplers, a vacuum pump, and other instruments. Three test methods with their own test dusts and procedures are used for the testing of low-, medium-, and high-efficiency air filters.

The *weight arrestance test* is used for low-efficiency air filters to assess their ability to remove coarse dusts. Standard synthetic dusts that are considerably coarser than atmospheric dust are fed to the test unit. By measuring the weight of dust fed and the weight gain due to the dust collected on the membrane of the sampler after the tested filter, the arrestance can be calculated.

The *atmospheric dust spot efficiency test* is used for medium-efficiency air filters to assess their ability to remove atmospheric dusts. *Atmospheric dusts* are dusts contained in the outdoor air, the outdoor atmosphere. Approximately 99% of atmospheric dusts are dust particles <0.3 μm that make up 10% of the total weight; 0.1% of atmospheric dusts is particles >1 μm that make up 70% of the total weight.

Untreated atmospheric dusts are fed to the test unit. Air samples taken before and after the tested filter are drawn through from identical fiber filter-paper targets. By measuring the light transmission of these discolored white filter papers, the efficiency of the filter can be calculated. Similar atmospheric

dust spot test procedures have been specified by American Filter Institute (AFI), ASHRAE Standard 52.1, and former National Bureau of Standards (NBS).

The *DOP penetration and efficiency test* or simply *DOP test* is used to assess high-efficiency filters removing dusts particles of 0.18 μm. According to U.S. Military Standard MIL-STD-282 (1956), a smoke cloud of uniform dioctyl phthalate (DOP) droplets 0.18 μm in diameter, generated from the condensation of the DOP vapor, is fed to the test unit. By measuring the concentration of these particles in the air stream upstream and downstream of the tested filter using an electronic particle counter or laser spectrometer, the penetration and efficiency of the air filter can be calculated.

Low-Efficiency Air Filters

ASHRAE weight arrestance for low-efficiency filters is between 60 and 95%, and ASHRAE dust spot efficiency for low-efficiency filters is less than 20%. These filters are usually in panels as shown in Figure 9.7.4(a). Their framework is typically 20 × 20 in. or 24 × 24 in. Their thickness varies from 1 to 4 in.

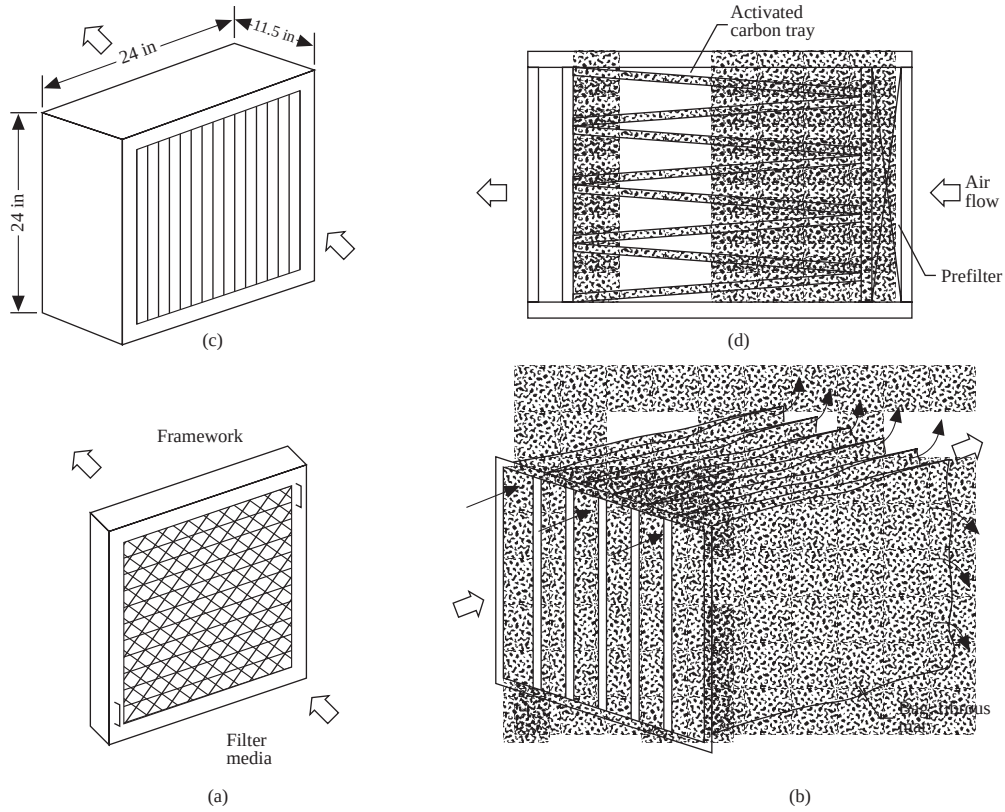


FIGURE 9.7.4 Various types of air filters: (a) low efficiency, (b) medium efficiency, (c) HEPA and ULPA filters, and (d) activated carbon filter.

For low-efficiency filters, the filter media are often made of materials such as

- Corrugated wire mesh and screen strips coated with oil, which act as adhesives to enhance dust removal. Detergents may be used to wash off dusts so that the filter media can be cleaned and reused — they are therefore called *viscous and reusable*.
- Synthetic fibers (nylon, terylene) and polyurethane foam can be washed, cleaned, and reused if required — *dry and reusable*.

- Glass fiber mats with fiber diameter greater than 10 μm. The filter medium is discarded when its final pressure drop is reached — *dry and disposable*. The face velocity of the panel filter is usually between 300 and 600 fpm. The initial pressure drop varies from 0.05 to 0.25 in. WG and the final pressure drop from 0.2 to 0.5 in. WG.

Medium-Efficiency Air Filters

These air filters have an ASHRAE dust spot efficiency usually between 20 and 95%. Filter media of medium-efficiency filters are usually made of glass fiber mat with a fiber diameter of 10 to 1 μm using nylon fibers to join them together. They are usually dry and disposable. In addition:

- As the dust spot efficiency increases, the diameter of glass fibers is reduced, and they are placed closer together.
- Extended surfaces, such as pleated mats or bags, are used to increase the surface area of the medium as shown in Figure 9.7.4(b). Air velocity through the medium is 6 to 90 fpm. Face velocity of the air filter is about 500 fpm to match the face velocity of the coil in AHUs and PUs.
- Initial pressure drop varies from 0.20 to 0.60 in. WG and final pressure drop from 0.50 to 1.20 in. WG.

High-Efficiency Particulate Air (HEPA) Filters and Ultra-Low-Penetration Air (ULPA) Filters

HEPA filters have a DOP test efficiency of 99.97% for dust particles ≥0.3 μm in diameter. *ULPA filters* have a DOP test efficiency of 99.999% for dust particles ≥0.12 μm in diameter.

A typical HEPA filter, shown in Figure 9.7.4(d), has dimensions of 24 × 24 × 11.5 in. Its filter media are made of glass fibers of submicrometer diameter in the form of pleated paper mats. The medium is dry and disposable. The surface area of the HEPA filter may be 50 times its face area, and its rated face velocity varies from 190 to 390 fpm, normally at a pressure drop of 0.50 to 1.35 in. WG for clean filters. The final pressure drop is 0.8 to 2 in. WG. Sealing of the filter pack within its frame and sealing between the frame and the gaskets are critical factors that affect the penetration and efficiency of the HEPA filter.

An ULPA filter is similar to a HEPA filter in construction and filter media. Both its sealing and filter media are more efficient than those of a HEPA filter.

To extend the service life of HEPA filters and ULPA filters, both should be protected by a medium-efficiency filter, or a low-efficiency and a medium-efficiency filter in the sequence low–medium just before the HEPA or ULPA filters. HEPA and ULPA filters are widely used in clean rooms and clean spaces.

Activated Carbon Filters

These filters are widely used to remove objectional odors and irritating gaseous airborne particulates, typically 0.003 to 0.006 μm in size, from the air stream by adsorption. *Adsorption* is physical condensation of gas or vapor on the surface of an activated substance like activated carbon. Activated substances are extremely porous. One pound of activated carbon contains 5,000,000 ft^2 of internal surface.

Activated carbon in the form of granules or pellets is made of coal, coconut shells, or petroleum residues and is placed in trays to form activated carbon beds as shown in Figure 9.7.4(d). A typical carbon tray is 23 × 23 × 5/8 in. thick. Low-efficiency prefilters are used for protection. When air flows through the carbon beds at a face velocity of 375 to 500 fpm, the corresponding pressure drop is 0.2 to 0.3 in. WG.

Humidifiers

A *humidifier* adds moisture to the air. Air is humidified by: (1) heating the liquid to evaporate it; (2) atomizing the liquid water into minute droplets by mechanical means, compressed air, or ultrasonic vibration to create a larger area for evaporation; (3) forcing air to flow through a wetted element in

which water evaporates; and (4) injecting steam into air directly before it is supplied to the conditioned space.

For comfort air-conditioning systems, a steam humidifier with a separator as shown in Figure 9.7.5(a) is widely used. Steam is supplied to a jacketed distribution manifold. It enters a separating chamber with its condensate. Steam then flows through a control valve, throttles to a pressure slightly above atmospheric, and enters a dry chamber. Due to the high temperature in the surrounding separating chamber, the steam is superheated. Dry steam is then discharged into the ambient air stream through the orifices on the inner steam discharge tubes.

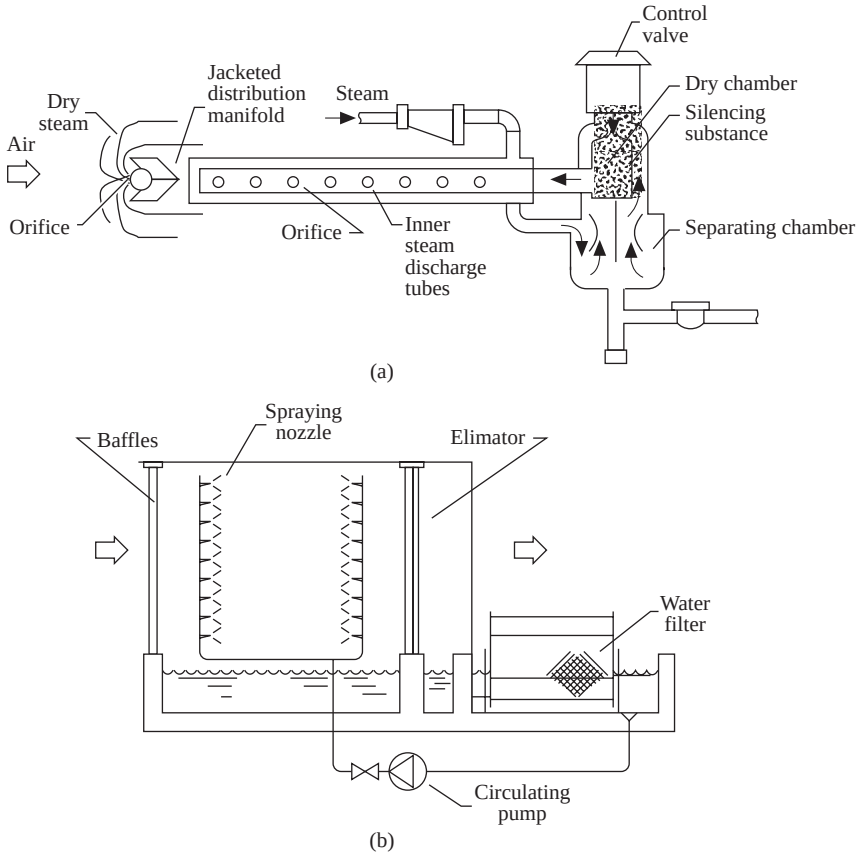


FIGURE 9.7.5 Steam grid humidifier (a) and air washer (b).

For an air system of cold air supply with humidity control during winter mode operation, an air washer is economical for large-capacity humidification in many industrial applications.

An air washer is a humidifier, a cooler, a dehumidifier, and an air cleaner. An air washer usually has an outer casing, two banks of spraying nozzles, one bank of guide baffles at the entrance, one bank of eliminators at the exit, a water tank, a circulating pump, a water filter, and other accessories as shown in Figure 9.7.5(b). Outer casing, baffles, and eliminators are often made of plastics or sometimes stainless steel. Spraying nozzles are usually made of brass or nylon, with an orifice diameter of 1/16 to 3/16 in., a smaller orifice for humidification, and a larger orifice for cooling and dehumidification. An eccentric inlet connected to the discharge chamber of the spraying nozzle gives centrifugal force to the water stream and atomizes the spraying water. Water is supplied to the spraying nozzle at a pressure of 15 to 30 psig. The distance between two spraying banks is 3 to 4.5 ft, and the total length of the air water from 4 to 7 ft. The air velocity inside an air washer is usually 500 to 800 fpm.

Selection of AHUs and PUs

- The size of an AHU is usually selected so that the face velocity of its coil is 600 fpm or less in order to prevent entrained condensate droplets. The cooling and heating capacities of an AHU can be varied by using coils of different numbers of rows and fin densities. The size of a PU is determined by its cooling capacity. Normally, the volume flow rate per ton of cooling capacity in PUs is 350 to 400 cfm. In most packaged units whose supply fans have belt drives, the fan speed can be selected so that the volume flow rate is varied and external pressure is met.
- ASHRAE/IES Standard 90.1-1989 specifies that the selected equipment capacity may exceed the design load only when it is the smallest size needed to meet the load. Selected equipment in a size larger always means a waste of energy and investment.
- To improve the indoor air quality, save energy, and prevent smudging and discoloring building interiors, a medium-efficiency filter of dust spot efficiency ≥50% and an air economizer are preferable for large AHUs and PUs.

9.8 Refrigeration Components and Evaporative Coolers

Refrigeration Compressors

A refrigeration compressor is the heart of a vapor compression system. It raises the pressure of refrigerant so that it can be condensed into liquid, throttled, and evaporated into vapor to produce the refrigeration effect. It also provides the motive force to circulate the refrigerant through condenser, expansion valve, and evaporator.

According to the compression process, refrigeration compressors can be divided into *positive displacement* and *nonpositive displacement* compressors. A positive displacement compressor increases the pressure of the refrigerant by reducing the internal volume of the compression chamber. Reciprocating, scroll, rotary, and screw compressors are all positive displacement compressors. The centrifugal compressor is the only type of nonpositive displacement refrigeration compressor widely used in refrigeration systems today.

Based on the sealing of the refrigerant, refrigeration compressors can be classified as

- *Hermetic compressors*, in which the motor and the compressor are sealed or welded in the same housing to minimize leakage of refrigerant and to cool the motor windings by using suction vapor
- *Semihermetic compressors*, in which motor and compressor are enclosed in the same housing but are accessible from the cylinder head for repair and maintenance
- *Open compressors*, in which compressor and motor are enclosed in two separate housings

Refrigeration compressors are often driven by motor directly or by gear train.

Performance Indices

Volumetric efficiency η_v of a refrigeration compressor is defined as

$$\eta_v = \mathring{V}_{a.v} \Big/ \mathring{V}_p \tag{9.8.1}$$

where $\mathring{V}_{a.v}$ = actual induced volume of the suction vapor at suction pressure, cfm
$\mathring{V}_p$ = calculated displacement of the compressor, cfm

Isentropic efficiency η_{isen}, *compression efficiency* η_{cp}, compressor efficiency η_{com}, and *mechanical efficiency* η_{mec} are defined as

$$\begin{aligned} \eta_{isen} &= (h_2 - h_1) \Big/ (h_2' - h_1) = \eta_{cp}\eta_{mec} = \eta_{com} \\ \eta_{cp} &= W_{sen} / W_v \\ \eta_{mec} &= W_v / W_{com} \end{aligned} \tag{9.8.2}$$

where h_1, h_2, h_2' = enthalpy of the suction vapor, ideal discharged hot gas, and actual discharged hot gas, respectively, Btu/lb
W_{isen}, W_v, W_{com} = isentropic work = $(h_2 - h_1)$, work delivered to the vapor refrigerant, and work delivered to the compressor shaft, Btu/lb

The actual power input to the compressor P_{com}, in hp, can be calculated as

$$\begin{aligned} P_{com} &= \mathring{m}_r (h_2 - h_1) \Big/ (42.41\eta_{isen}\eta_{mo}) \\ \mathring{m}_r &= \mathring{V}_p \eta_v \rho_{suc} \\ \eta_{mo} &= P_{com} / P_{mo} \end{aligned} \tag{9.8.3}$$

where $\dot{m}_r$ = mass flow rate of refrigerant, lb/min
ρ_{suc} = density of suction vapor, lb/ft^3
P_{mo} = power input to the compressor motor, hp

Power consumption, kW/ton refrigeration, is an energy index used in the HVAC&R industry in addition to EER and COP.

Currently used refrigeration compressors are reciprocating, scroll, screw, rotary, and centrifugal compressors.

Reciprocating Compressors

In a reciprocating compressor, as shown in Figure 9.8.1(a), a crankshaft connected to the motor shaft drives 2, 3, 4, or 6 single-acting pistons moving reciprocally in the cylinders via a connecting rod.

The refrigeration capacity of a reciprocating compressor is a fraction of a ton to about 200 tons. Refrigerants R-22 and R-134a are widely used in comfort and processing systems and sometimes R-717 in industrial applications. The maximum compression ratio R_{com} for a single-stage reciprocating compressor is about 7. Volumetric efficiency η_v drops from 0.92 to 0.65 when R_{com} is raised from 1 to 6. Capacity control of reciprocating compressor including: on-off and cylinder unloader in which discharge gas is in short cut and return to the suction chamber.

Although reciprocating compressors are still widely used today in small and medium-sized refrigeration systems, they have little room for significant improvement and will be gradually replaced by scroll and screw compressors.

Scroll Compressors

A scroll compressor consists of two identical spiral scrolls assembled opposite to each other, as shown in Figure 9.8.1(b). One of the scrolls is fixed, and the other moves in an orbit around the motor shaft whose amplitude equals the radius of the orbit. The two scrolls are in contact at several points and therefore form a series of pockets.

Vapor refrigerant enters the space between two scrolls through lateral openings. The lateral openings are then sealed and the formation of the two trapped vapor pockets indicates the end of the suction process. The vapor is compressed and the discharge process begins when the trapped gaseous pockets open to the discharge port. Compressed hot gas is then discharged through this opening to the discharge line. In a scroll compressor, the scrolls touch each other with sufficient force to form a seal but not enough to cause wear.

The upper limit of the refrigeration capacity of currently manufactured scroll compressors is 60 tons. A scroll compressor has $\eta_v > 95\%$ at $R_{com} = 4$ and $\eta_{isen} = 80\%$. A scroll compressor also has only about half as many parts as a reciprocating compressor at the same refrigeration capacity. Few components result in higher reliability and efficiency. Power input to the scroll compressor is about 5 to 10% less than to the reciprocating compressor. A scroll compressor also operates more smoothly and is quieter.

Rotary Compressors

Small rotary compressors for room air conditioners and refrigerators have a capacity up to 4 tons. There are two types of rotary compressors: rolling piston and rotating vane. A typical rolling piston rotary compressor is shown in Figure 9.8.1(c). A rolling piston mounted on an eccentric shaft is kept in contact with a fixed vane that slides in a slot. Vapor refrigerant enters the compression chamber and is compressed by the eccentric motion of the roller. When the rolling piston contacts the top housing, hot gas is squeezed out from the discharge valve.

Screw Compressors

These are also called *helical rotary compressors*. Screw compressors can be classified into single-screw compressors, in which there is a single helical rotor and two star wheels, and twin-screw compressors. Twin-screw compressors are widely used.

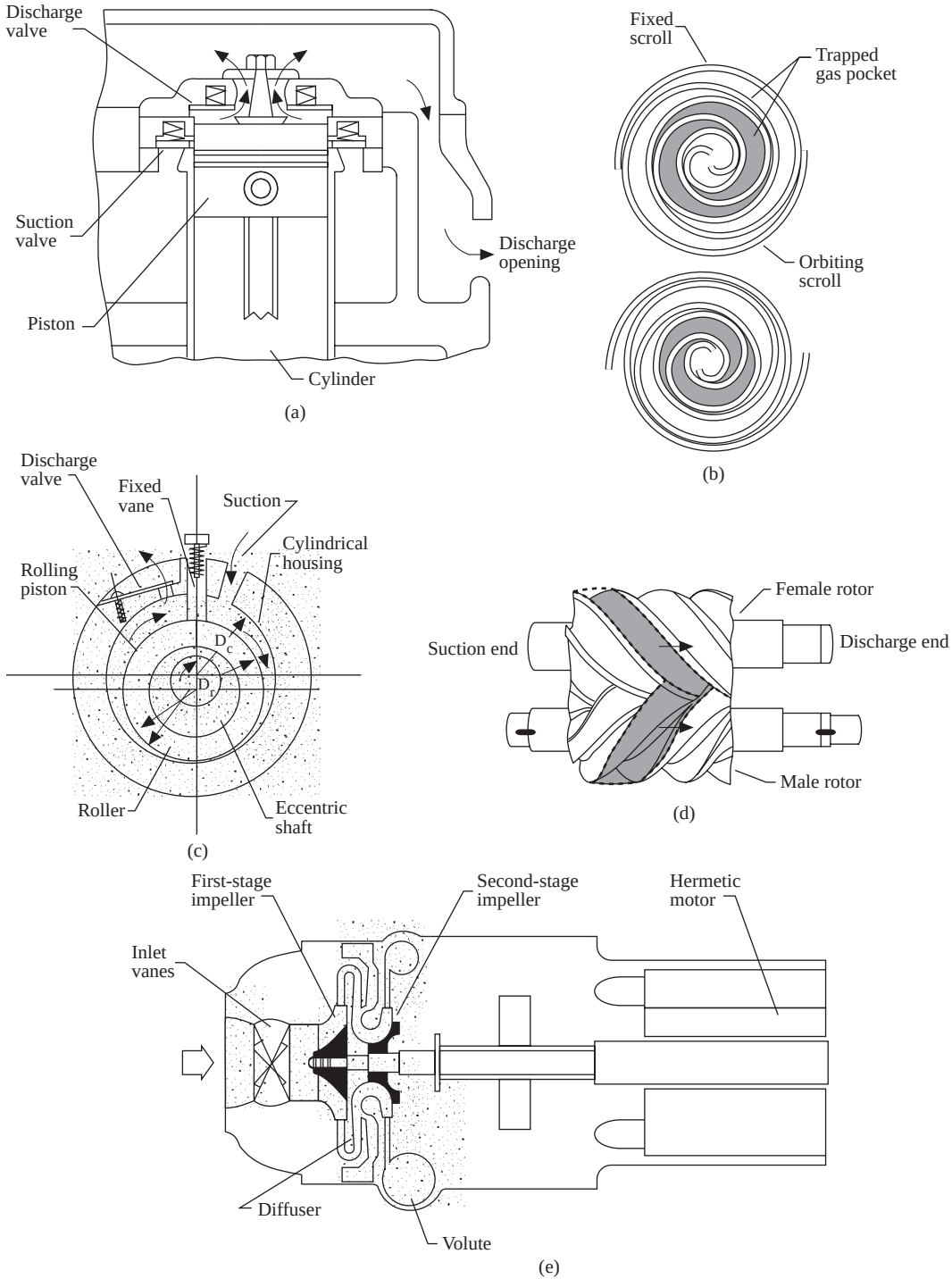


FIGURE 9.8.1 Various types of refrigeration compressors: (a) reciprocating, (b) scroll, (c) rotary, (d) twin-screw, and (e) centrifugal.

A typical twin-screw compressor, as shown in Figure 9.8.1(d) consists of a four-lobe male rotor and a six-lobe female rotor, a housing with suction and discharge ports, and a sliding valve to adjust the

capacity during part load. Normally, the male rotor is the driver. Twin-screw compressors are often direct driven and of hermetic type.

Vapor refrigerant is extracted into the interlobe space when the lobes are separated at the suction port. During the successive rotations of the rotor, the volume of the trapped vapor is compressed. When the interlobe space is in contact with the discharge port, the compressed hot gas discharges through the outlet. Oil injection effectively cools the rotors and results in a lower discharge temperature. Oil also provides a sealing effect and lubrication. A small clearance of 0.0005 in. as well as the oil sealing minimizes leakage of the refrigerant.

The refrigeration capacity of twin-screw compressors is 50 to 1500 tons. The compression ratio of a twin-screw compressor can be up to 20:1. R-22 and R-134a are the most widely used refrigerants in comfort systems. In a typical twin-screw compressor, η_v decreases from 0.92 to 0.87 and η_{isen} drops from 0.82 to 0.67 when R_{com} increases from 2 to 10. Continuous and stepless capacity control is provided by moving a sliding valve toward the discharge port, which opens a shortcut recirculating passage to the suction port.

Twin-screw compressors are more efficient than reciprocating compressors. The low noise and vibration of the twin-screw compressor together with its positive displacement compression results in more applications today.

Centrifugal Compressors

A *centrifugal compressor* is a turbomachine and is similar to a centrifugal fan. A hermetic centrifugal compressor has an outer casing with one, two, or even three impellers internally connected in series and is driven by a motor directly or by a gear train. At the entrance to the first-stage impeller are inlet guide vanes positioned at a specific opening to adjust refrigerant flow and therefore the capacity of the centrifugal compressor.

Figure 9.8.1(e) shows a two-stage hermetic centrifugal compressor. The total pressure rise in a centrifugal compressor, often called head lift, in psi, is due to the conversion of the velocity pressure into static pressure. Although the compression ratio R_{com} of a single-stage centrifugal compressor using R-123 and R-22 seldom exceeds 4, two or three impellers connected in series satisfy most of the requirements in comfort systems.

Because of the high head lift to raise the evaporating pressure to condensing pressure, the discharge velocity at the exit of the second-stage impeller approaches the acoustic velocity of saturated vapor v_{ac} of R-123, 420 ft/sec at atmospheric pressure and a temperature of 80°F. Centrifugal compressors need high peripheral velocity and rotating speeds (up to 50,000 rpm) to produce such a discharge velocity. It is not economical to manufacture small centrifugal compressors. The available refrigeration capacity for centrifugal compressors ranges from 100 to 10,000 tons. Centrifugal compressors have higher volume flow per unit refrigeration capacity output than positive displacement compressors. Centrifugal compressors are efficient and reliable. Their volumetric efficiency almost equals 1. At design conditions, their η_{isen} may reach 0.83, and it drops to 0.6 during part-load operation. They are the most widely used refrigeration compressors in large air-conditioning systems.

Refrigeration Condensers

A *refrigeration condenser* or simply a *condenser* is a heat exchanger in which hot gaseous refrigerant is condensed into liquid and the latent heat of condensation is rejected to the atmospheric air, surface water, or well water. In a condenser, hot gas is first desuperheated, then condensed into liquid, and finally subcooled.

The capacity of a condenser is rated by its *total heat rejection* Q_{rej}, in Btu/hr, which is defined as the total heat removed from the condenser during desuperheating, condensation, and subcooling. For a refrigeration system using a hermetic compressor, Q_{rej} can be calculated as

$$Q_{rej} = U_{con} A_{con} \Delta T_m = 60 \dot{m}_r \left(h_2 - h_3'\right) = q_{rl} + \left(2545 P_{com}\right) / \eta_{mo} \tag{9.8.4}$$

where U_{con} = overall heat transfer coefficient across the tube wall in the condenser, Btu/hr.ft^2.°F
A_{con} = condensing area in the condenser, ft^2
ΔT_m = logarithmic temperature difference, °F
$\dot{m}_r$ = mass flow rate of refrigerant, lb/min
h_2, h_3' = enthalpy of suction vapor refrigerant and hot gas, Btu/lb
q_{rl} = refrigeration load at the evaporator, Btu/hr

A factor that relates Q_{rej} and q_{rl} is the *heat rejection factor* F_{rej}, which is defined as the ratio of total heat rejection to the refrigeration load, or

$$F_{rej} = Q_{rej} / q_{rl} = 1 + \left(2545 P_{com}\right) / \left(q_{rl} \eta_{mo}\right) \tag{9.8.5}$$

Fouling factor R_f, in hr.ft^2.°F/Btu, is defined as the additional resistance caused by a dirty film of scale, rust, or other deposits on the surface of the tube. ARI Standard 550-88 specifies the following for evaporators and condensers:

Field fouling allowance	0.00025 hr.ft^2.°F/Btu
New evaporators and condensers	0

According to the cooling process used during condensation, refrigeration condensers can be classified as air-cooled, water-cooled, and evaporative-cooled condensers.

Air-Cooled Condensers

In an *air-cooled condenser*, air is used to absorb the latent heat of condensation released during desuperheating, condensation, and subcooling.

An air-cooled condenser consists of a condenser coil, a subcooling coil, condenser fans, dampers, and controls as shown in Figure 9.8.2(a). There are refrigeration circuits in the condensing coil. Condensing coils are usually made of copper tubes and aluminum fins. The diameter of the tubes is 1/4 to 3/4 in., typically 3/8 in., and the fin density is 8 to 20 fins/in. On the inner surface of the copper tubes, microfins, typically 60 fins/in. with a height of 0.008 in., are used. A condensing coil usually has only two to three rows due to the low pressure drop of the propeller-type condenser fans. A subcooling coil is located at a lower level and is connected to the condensing coil.

Hot gas from the compressor enters the condensing coil from the top. When the condensate increases, part of the condensing area can be used as a subcooling area. A receiver is necessary only when the liquid refrigerant cannot all be stored in the condensing and subcooling coils during the shut-down period in winter.

Cooling air is drawn through the coils by a condenser fan(s) for even distribution. Condenser fans are often propeller fans for their low pressure and large volume flow rate. A damper(s) may be installed to adjust the volume flow of cooling air.

In air-cooled condensers, the volume flow of cooling air per unit of total heat rejection $\dot{V}_{ca} / Q_{u.rej}$ is 600 to 1200 cfm/ton of refrigeration capacity at the evaporator, and the optimum value is about 900 cfm/ton. The corresponding cooling air temperature difference — cooling air leaving temperature minus outdoor temperature ($T_{ca.l} - T_o$) — is around 13°F.

The condenser temperature difference (CTD) for an air-cooled condenser is defined as the difference between the saturated condensing temperature corresponding to the pressure at the inlet and the air intake temperature, or ($T_{con.i} - T_o$). Air-cooled condensers are rated at a specific CTD, depending on the evaporating temperature of the refrigeration system T_{ev} in which the air-cooled condenser is installed. For a refrigeration system having a lower T_{ev}, it is more economical to equip a larger condenser with a smaller CTD. For a comfort air-conditioning system having a T_{ev} of 45°F, CTD = 20 to 30°F.

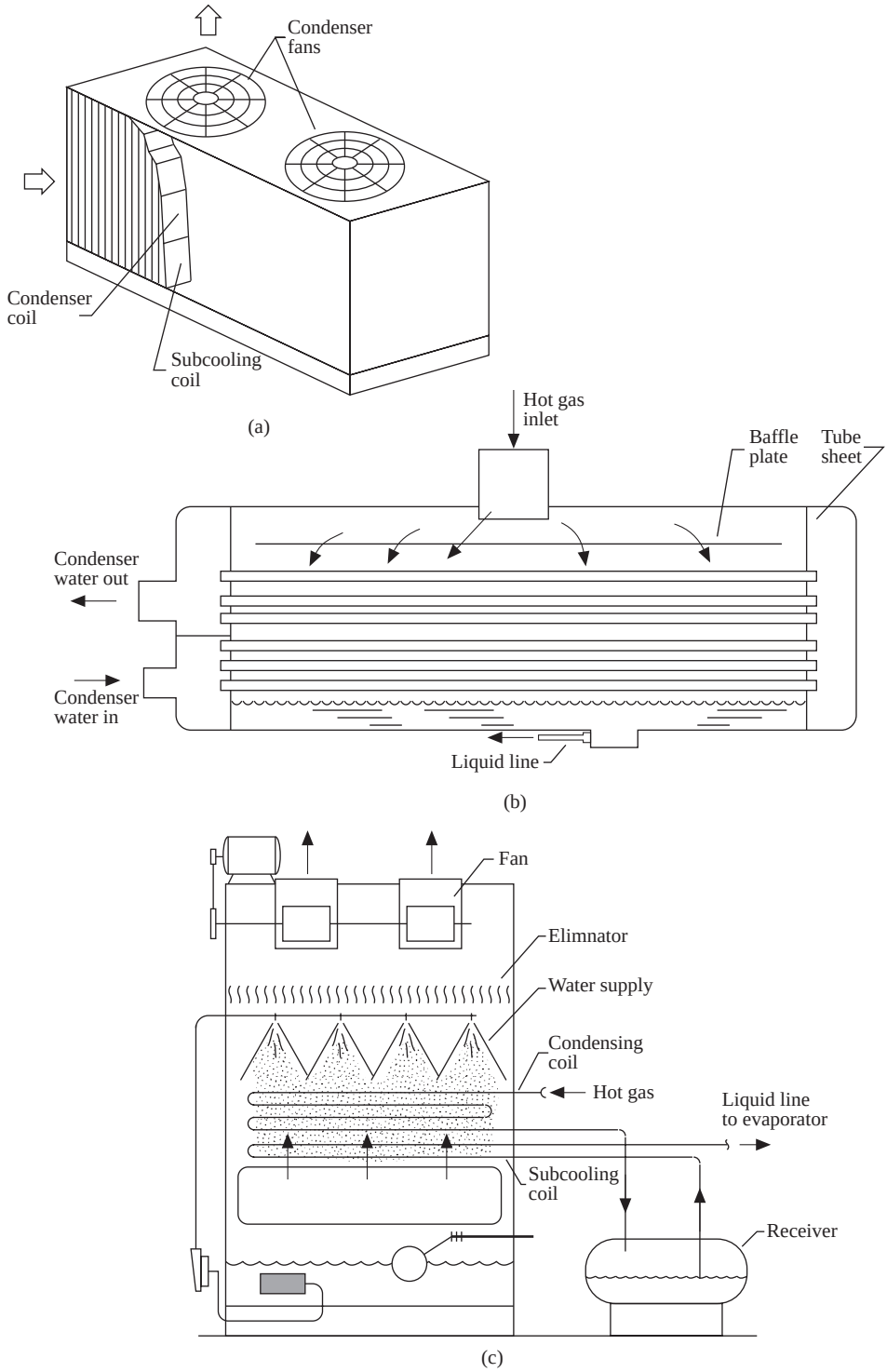


FIGURE 9.8.2 Various types of refrigeration condensers: (a) air-cooled, (b) shell-and-tube water-cooled, and (c) evaporative cooled.

A higher condensing temperature T_{con}, a higher condensing pressure p_{con}, and a higher compressor power input may be due to an undersized air-cooled condenser, lack of cooling air or low $\mathring{V}_{ca}/Q_{u.rej}$ value, a high entering cooling air temperature at the roof, a dirty condensing coil, warm air circulation because of insufficient clearance between the condenser and the wall, or a combination of these. The clearance should not be less than the width of the condensing coil.

If p_{con} drops below a certain value because of a lower outdoor temperature, the expansion valve in a reciprocating vapor compression system may not operate properly. At a low ambient temperature T_o, the following controls are often used:

- Duty cycling, turning the condenser fans on and off until all of them are shut down, to reduce cooling air volume flow
- Modulating the air dampers to reduce the volume flow
- Reducing the fan speed

Some manufacturers' catalogs start low ambient control at T_o = 65°F and some specify a minimum operating temperature at T_o = 0°F.

Water-Cooled Condensers

In a *water-cooled condenser*, latent heat of condensation released from the refrigerant during condensation is extracted by water. This cooling water, often called condenser water, is taken directly from river, lake, sea, underground well water or a cooling tower.

Two types of water-cooled condensers are widely used for air-conditioning and refrigeration: double-tube condensers and horizontal shell-and-tube condensers.

A *double-tube condenser* consists of two tubes, one inside the other. Condenser water is pumped through the inner tube and refrigerant flows within the space between the inner and outer tubes in a counterflow arrangement. Because of its limited condensing area, the double-tube condenser is used only in small refrigeration systems.

A horizontal *shell-and-tube water-cooled condenser* using halocarbon refrigerant usually has an outer shell in which copper tubes typically 5/8 to 3/4 in. in diameter are fixed in position by tube sheets as shown in Figure 9.8.2(b). Integral external fins of 19 to 35 fins/in. and a height of 0.006 in. and spiral internal grooves are used for copper tubes to increase both the external and the inner surface area and their heat transfer coefficients.

Hot gas from the compressor enters the top inlet and is distributed along the baffle to fill the shell. Hot gas is then desuperheated, condensed, subcooled into liquid, and discharged into the liquid line at the bottom outlet. Usually one sixth of the volume is filled with subcooled liquid refrigerant. Subcooling depends on the entering temperature of condenser water T_{ce}, in °F, and usually varies between 2 and 8°F.

Condenser water enters the condenser from the bottom for effective subcooling. After extracting heat from the gaseous refrigerant, condenser water is discharged at a higher level. Two-pass or three-pass water flow arrangements are usually used in shell-and-tube water-cooled condensers. The two-pass arrangement means that water flows from one end to the opposite end and returns to the original end. Two-pass is the standard setup. In a shell-and-tube water-cooled condenser, the condensing temperature T_{con} depends mainly on the entering temperature of condenser water T_{ce}, the condenser area, the fouling factor, and the configuration of the copper tube.

Evaporative Condenser

An *evaporative condenser* uses the evaporation of water spray on the outer surface of the condensing tubes to remove the latent heat of condensation of refrigerant during condensation.

An evaporative condenser consists of a condensing coil, a subcooling coil, a water spray, an induced draft or sometimes forced draft fan, a circulating water pump, a water eliminator, a water basin, an outer casing, and controls as shown in Figure 9.8.2(c). The condensing coil is usually made of bare copper, steel, or sometimes stainless steel tubing.

Water is sprayed over the outside surface of the tubing. The evaporation of a fraction of condenser water from the saturated air film removes the sensible and latent heat rejected by the refrigerant. The wetted outer surface heat transfer coefficient h_{wet} is about four or five times greater than the dry surface heat transfer coefficient h_o, in Btu/hr.ft^2.°F. The rest of the spray falls and is collected by the basin. Air enters from the inlet just above the basin. It flows through the condensing coil at a face velocity of 400 to 700 fpm, the water spray bank, and the eliminator. After air absorbs the evaporated water vapor, it is extracted by the fan and discharged at the top outlet. The water circulation rate is about 1.6 to 2 gpm/ton, which is far less than that of the cooling tower.

An evaporative condenser is actually a combination of a water-cooled condenser and a cooling tower. It is usually located on the rooftop and should be as near the compressor as possible. Clean tube surface and good maintenance are critical factors for evaporative condensers. An evaporative condenser also needs low ambient control similar as in an air-cooled condenser.

Comparison of Air-Cooled, Water-Cooled, and Evaporative Condensers

An air-cooled condenser has the highest condensing temperature T_{con} and therefore the highest compressor power input. For an outdoor dry bulb temperature of 90°F and a wet bulb temperature of 78°F, a typical air-cooled condenser has T_{con} = 110°F. An evaporative condenser has the lowest T_{con} and is most energy efficient. At the same outdoor dry and wet bulb temperatures, its T_{con} may be equal to 95°F, even lower than that of a water-cooled condenser incorporating with a cooling tower, whose T_{con} may be equal to 100°F. An evaporative condenser also consumes less water and pump power. The drawback of evaporative condensers is that the rejected heat from the interior zone is difficult to recover and use as winter heating for perimeter zones and more maintenance is required.

Evaporators and Refrigerant Flow Control Devices

An *evaporator* is a heat exchanger in which the liquid refrigerant is vaporized and extracts heat from the surrounding air, chilled water, brine, or other substance to produce a refrigeration effect.

Evaporators used in air-conditioning can be classified according to the combination of the medium to be cooled and the type of refrigerant feed, as the following.

Direct expansion DX coils are air coolers, and the refrigerant is fed according to its degree of superheat after vaporization. DX coils were covered earlier.

Direct expansion ice makers or *liquid overfeed ice makers* are such that liquid refrigerant is forced through the copper tubes or the hollow inner part of a plate heat exchanger and vaporized. The refrigeration effect freezes the water in the glycol-water that flows over the outside surface of the tubes or the plate heat exchanger. In direct expansion ice makers, liquid refrigerant completely vaporizes inside the copper tubes, and the superheated vapor is extracted by the compressor. In liquid overfeed ice makers, liquid refrigerant floods and wets the inner surface of the copper tubes or the hollow plate heat exchanger. Only part of the liquid refrigerant is vaporized. The rest is returned to a receiver and pumped to the copper tubes or plate heat exchanger again at a circulation rate two to several times greater than the evaporation rate.

Flooded shell-and-tube liquid coolers, or simply *flooded liquid coolers*, are such that refrigerant floods and wets all the boiling surfaces and results in high heat transfer coefficients. A flooded shell-and-tube liquid cooler is similar in construction to a shell-and-tube water-cooled condenser, except that its liquid refrigeration inlet is at the bottom and the vapor outlet is at the top. Water velocity inside the copper tubes is usually between 4 and 12 ft/sec and the water-side pressure normally drops below 10 psi. Flooded liquid coolers can provide larger evaporating surface area and need minimal space. They are widely used in large central air-conditioning systems.

Currently used refrigerant flow control devices include thermostatic expansion valves, float valves, multiple orifices, and capillary tubes.

A *thermostatic expansion valve* throttles the refrigerant pressure from condensing to evaporating pressure and at the same time regulates the rate of refrigerant feed according to the degree of superheat

of the vapor at the evaporator's exit. A thermostatic expansion valve is usually installed just prior to the refrigerant distributor in DX coils and direct-expansion ice makers.

A thermostatic expansion valve consists of a valve body, a valve pin, a spring, a diaphragm, and a sensing bulb near the outlet of the DX coil, as shown in Figure 9.7.3(a). The sensing bulb is connected to the upper part of the diaphragm by a connecting tube.

When the liquid refrigerant passes through the opening of the thermostatic expansion valve, its pressure is reduced to the evaporating pressure. Liquid and a small fraction of vaporized refrigerant then flow through the distributor and enter various refrigerant circuits. If the refrigeration load of the DX coil increases, more liquid refrigerant evaporizes. This increases the degree of superheat of the leaving vapor at the outlet and the temperature of the sensing bulb. A higher bulb temperature exerts a higher saturated pressure on the top of the diaphragm. The valve pin then moves downward and widens the opening. More liquid refrigerant is allowed to enter the DX coil to match the increase of refrigeration load. If the refrigeration load drops, the degree of superheat at the outlet and the temperature of the sensing bulb both drop, and the valve opening is narrower. The refrigeration feed decreases accordingly. The degree of superheat is usually 10 to 20°F. Its value can also be adjusted manually by varying the spring tension.

A *float valve* is a valve in which a float is used to regulate the valve opening to maintain a specific liquid refrigerant level. A lower liquid level causes a lower valve pin and therefore a wider opening and vice versa.

In a centrifugal refrigeration system, two or more orifice plates, *multiple orifices*, are sometimes installed in the liquid line between the condenser and the flash cooler and between the flash cooler and the flooded liquid cooler to throttle their pressure as well as to regulate the refrigerant feed.

A *capillary tube*, sometimes called a *restrictor tube*, is a fixed length of small-diameter tubing installed between the condenser and the evaporator to throttle the refrigerant pressure from p_{con} to p_{ev} and to meter the refrigerant flow to the evaporator. Capillary tubes are usually made of copper. The inside diameter D_{cap} is 0.05 to 0.06 in. and the length L_{cap} from an inch to several feet. There is a trend to use short capillary tubes of L_{cap}/D_{cap} between 3 and 20. Capillary tubes are especially suitable for a heat pump system in which the refrigerant flow may be reversed.

Evaporative Coolers

An evaporative cooling system is an air-conditioning system in which air is cooled evaporatively. It consists of evaporative coolers, fan(s), filters, dampers, controls, and others. A mixing box is optional. An evaporative cooler could be a stand-alone cooler or installed in an air system as a component. There are three types of evaporative coolers: (1) direct evaporative coolers, (2) indirect evaporative coolers, and (3) indirect–direct evaporative coolers.

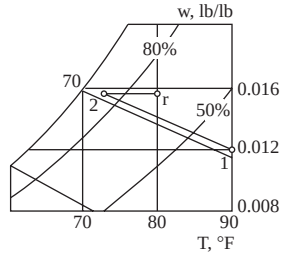
Direct Evaporative Cooler

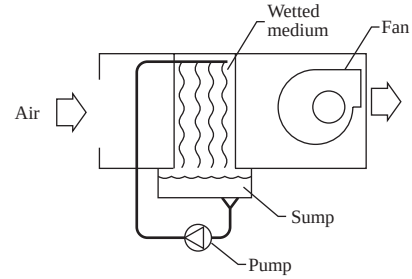
In a *direct evaporative cooler*, the air stream to be cooled directly contacts the water spray or wetted medium as shown in Figure 9.8.3(a). Evaporative pads made of wooden fibers with necessary treatment at a thickness of 2 in., rigid and corrugated plastics, impregnated cellulose, or fiber glass all dripping with water are wetted mediums.

The direct evaporation process 12 takes place along the thermodynamic wet bulb line on the psychrometric chart. Saturation effectiveness ε_{sat} is an index that assesses the performance of a direct evaporative cooler:

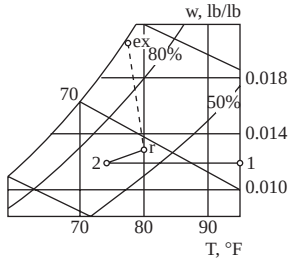
$$\varepsilon_{sat} = \left(T_{ae} - T_{al}\right) \Big/ \left(T_{ae} - T_{ae}^{*}\right) \tag{9.8.6}$$

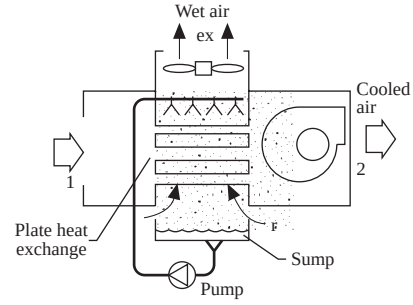
where T, T^* = temperature and thermodynamic wet bulb temperature of air stream, °F. Subscript ae indicates the entering air and al the leaving air. ε_{sat} usually varies between 0.75 and 0.95 at a water–air ratio of 0.1 to 0.4.



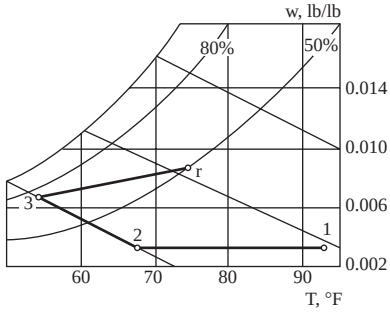


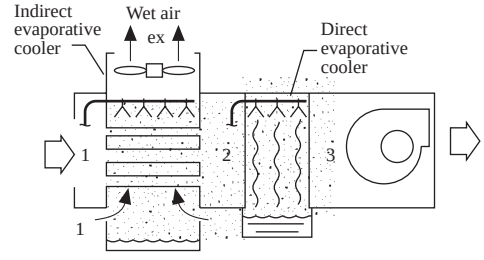
(a)





(b)





(c)

FIGURE 9.8.3 Types of evaporative coolers: (a) direct, (b) indirect, and (c) indirect–direct.

Indirect Evaporative Coolers

In an *indirect evaporative cooler*, the cooled-air stream to be cooled is separated from a wetted surface by a flat plate or tube wall as shown in Figure 9.8.3(b). A wet-air stream flows over the wetted surface so that liquid water is evaporated and extracts heat from the cooled-air stream through the flat plate or tube wall. The cooled-air stream is in contact with the wetted surface indirectly.

The core part of an indirect evaporative cooler is a plate heat exchanger. It is made of thin polyvinyl chloride plates 0.01 in. thick and spaced from 0.08 to 0.12 in. apart to form horizontal passages for cooled air and vertical passages for wet air and water. As in a direct evaporative cooler, there are also fan(s), water sprays, circulating pump, air intake, dampers, controls, etc.

An indirect evaporative cooling process is represented by a horizontal line on a psychrometric chart, which shows that humidity ratio remains constant. If the space air is extracted and used as the wet air intake, the wet air will be exhausted at point x at nearly saturated state.

The performance of an indirect evaporative cooler can be assessed by its performance factor e_{in}, which is calculated as:

$$e_{in} = \left(T_{ca.e} - T_{ca.l}\right) / \left(T_{ca.e} - T_{s.a}\right) \tag{9.8.7}$$

where $T_{ca.e}$, $T_{ca.l}$ = temperature of cooled air entering and leaving the indirect evaporative cooler, °F, and $T_{s.a}$ = temperature of the saturated air film on the wet air side and is about 3°F higher than the wet bulb temperature of the entering air, °F.

An indirect evaporative cooler could be so energy efficient as to provide evaporative cooling with an EER up to 50 instead of 9 to 12 for a reciprocating compression refrigeration system.

Direct–Indirect Evaporative Cooler. A direct–indirect evaporative cooler is a two-stage evaporating cooler, as shown in Figure 9.8.3(c), in which the first-stage indirect evaporative cooler is connected in series with a second-stage direct evaporative cooler for the purpose of increasing the evaporating effect.

Operating Characteristics. The saturation effectiveness ε_{sat} and performance factor e_{in} are both closely related to the air velocity flowing through the air passages. For a direct evaporative cooler, face velocity is usually less than 600 fpm to reduce drift carryover. For an indirect evaporative cooler, face velocity v_s is usually between 400 to 1000 fpm. A higher v_s results at a greater air-side pressure drop.

Scofield et al. (1984) reported the performance of an indirect–direct evaporative cooler in Denver, Colorado. Outdoor air enters the indirect cooler at a dry bulb of 93°F and a wet bulb of 67.5° and was evaporatively cooled to 67.5°F dry bulb and 49.8°F wet bulb with an e_{in} = 0.76 as shown in Figure 9.8.3(c). In the direct cooler, conditioned air was further cooled to a dry bulb of 53.5°F and the wet bulb remained at 49.8°F at a saturation effectiveness ε_{sat} = 0.8.

In locations where outdoor wet bulb $T_o' \leq 60$°F, a direct evaporative can often provide an indoor environment of 78°F and a relative humidity of 60%. In locations $T_o' \leq 68$°F, an indirect–direct evaporative cooler can maintain a comfortable indoor environment. In locations $T_o' \geq 72$°F, an evaporative cooler with a supplementary vapor compression refrigeration may be cost effective. Because the installation cost of an indirect–direct cooler is higher than that of refrigeration, cost analysis is required to select the right choice. Evaporative coolers are not suitable for dehumidification except in locations where $T_o' \leq 60$°F.

9.9 Water Systems

Types of Water Systems

In central and space conditioning air-conditioning systems, water that links the central plant and the air handling units or terminals, that extracts condensing heat, or that provides evaporative cooling may be classified as

- *Chilled water system*, in which chilled water is first cooled in the centrifugal, screw, and reciprocating chillers in a central plant. Chilled water is then used as a cooling medium to cool the air in the cooling coils in AHUs and terminals.
- *Evaporative-cooled water system*, used to cool air directly or indirectly in evaporative coolers.
- *Hot water system*, in which hot water is heated in the boiler and then used to heat the air through heating coils in AHUs, terminals, or space finned-tube heaters.
- *Dual-temperature water system*, in which chilled water and hot water are supplied to and returned from the coils in AHUs and terminals through separate or common main and branch pipes. Using common main and branch pipes requires a lengthy changeover from chilled water to hot water or vice versa for a period of several hours.
- *Condenser water system*, which is a kind of cooling water system used to extract the latent heat of condensation from the condensing refrigerant in a water-cooled condenser and heat of absorption from the absorber.

Water systems can also be classified according to their operating characteristics.

Closed System

In a closed system, water forms a closed loop for water conservation and energy saving when it flows through the coils, chillers, boilers, heaters, or other heat exchangers and water is not exposed to the atmosphere.

Open System

In an open system, water is exposed to the atmosphere.

Once-Through System

In a once-through system, water flows through a heat exchanger(s) only once without recirculation.

Basics

Volume Flow and Temperature Difference

The rate of heat transfer between water and air or water and refrigerant when water flows through a heat exchanger q_w, in Btu/hr, can be calculated as

$$q_w = 500\,\mathring{V}_{gal}\left(T_{wl} - T_{we}\right) = 500\,\mathring{V}_{gal}\,\Delta T_w \tag{9.9.1}$$

where $\mathring{V}_{gal}$ = volume flow rate of water, gpm
T_{wl}, T_{we} = temperature of water leaving and entering the heat exchanger, °F
ΔT_w = temperature rise or drop of water when it flows through a heat exchanger, °F

The temperature of chilled water leaving the water chiller T_{el} should not be lower than 38°F in order to prevent freezing in the evaporator. Otherwise, brine or glycol-water should be used. The T_{el} of chilled water entering the coil T_{we} and the temperature difference of chilled water leaving and entering the coil ΔT_w directly affect the temperature of air leaving the cooling coil T_{cc}. The lower T_{we}, the higher will be the compressor power input. The smaller ΔT_w, the greater will be the water volume flow rate, the pipe

size, and the pump power. For chilled water in conventional comfort systems, T_{we} is usually 40 to 45°F and ΔT_w 12 to 24°F. Only in cold air distribution, T_{we} may drop to 34°F. For a cooling capacity of 1 ton refrigeration, a ΔT_w of 12°F requires a $\mathring{V}_{gal}$ = 2 gpm.

For hot water heating systems in buildings, hot water often leaves the boiler and enters the heating coil or heaters at a temperature T_{we} of 190 to 200°F. It returns at 150 to 180°F. For dual-temperature systems, hot water is usually supplied at 100 to 150°F and returns at a ΔT_w of 20 to 40°F.

Pressure Drop

Usually the pressure drop of water in pipes due to friction for HVAC&R systems, H_f, is in the range 0.75 ft/100 ft length of pipe to 4 ft/100 ft. A pressure loss of 2.5 ft/100 ft is most often used. ASHRAE/IES Standard 90.1-1989 specifies that water piping systems should be designed at a pressure loss of no more than 4.0 ft/100 ft. Figure 9.9.1(a), (b), and (c) shows the friction charts for steel, copper, and plastic pipes for closed water systems.

Water Piping

The piping materials of various water systems for HVAC&R are as follows:

Chilled water	Black and galvanized steel
Hot water	Black steel, hard copper
Condenser water	Black steel, galvanized ductile iron, polyvinyl chloride (PVC)

The pipe thickness varies from Schedule 10, a light wall pipe, to Schedule 160, a very heavy wall pipe. Schedule 40 is the standard thickness for a pipe of up to 10 in. diameter. For copper tubing, type K is the heaviest, and type L is generally used as the standard for pressure copper tubes.

Steel pipes of small diameter are often joined by threaded cast-iron fittings. Steel pipes of diameter 2 in. and over, welded joints, and bolted flanges are often used.

In a water system, the maximum allowable working pressure for steel and copper pipes at 250°F varies from 125 to 400 psig, depending on the pipe wall thickness. Not only pipes, but also their joints and fittings should be considered.

During temperature changes, pipes expand and contract. Both operating and shut-down periods should be taken into consideration. Bends like U-, Z-, and L-bends, loops, and sometimes packed expansion joints, bellows, or flexible metal hose mechanical joints are used.

ASHRAE/IES Standard 90.1-1989 specifies that for chilled water between 40 to 55°F, the minimum thickness of the external pipe insulation varies from 0.5 to 1 in. depending on the pipe diameter. For chilled water temperature below 40°F, the minimum thickness varies from 1 to 1.5 in. For hot water temperatures not exceeding 200°F, the minimum thickness varies from 0.5 to 1.5 in. depending on the pipe diameter and the hot water temperature.

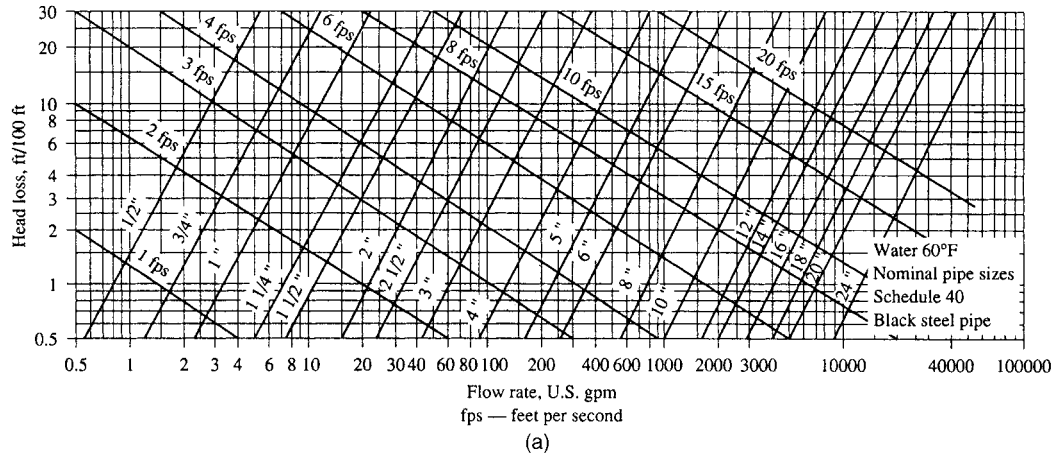
Corrosion, Impurities, and Water Treatments

Corrosion is a destructive process caused by a chemical or electrochemical reaction on metal or alloy. In water systems, dissolved *impurities* cause corrosion and scale and the growth of microbiologicals like algae, bacteria, and fungi. *Scale* is the deposit formed on a metal surface by precipitation of the insoluble constituents. In addition to the dissolved solids, unpurified water may contain suspended solids.

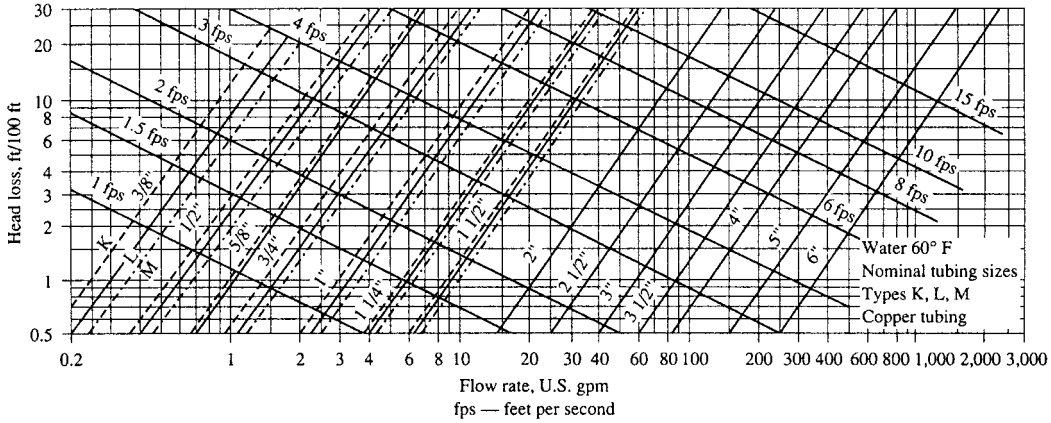
Currently used chemicals include crystal modifiers to change the crystal formation of scale and sequestering chemicals. Growth of bacteria, algae, and fungi is usually treated by biocides to prevent the formation of an insulating layer resulting in lower heat transfer as well as restricted water flow. Chlorine and its compounds are effective and widely used. Blow-down is an effective process in water treatment and should be considered as important as chemical treatments.

Piping Arrangements

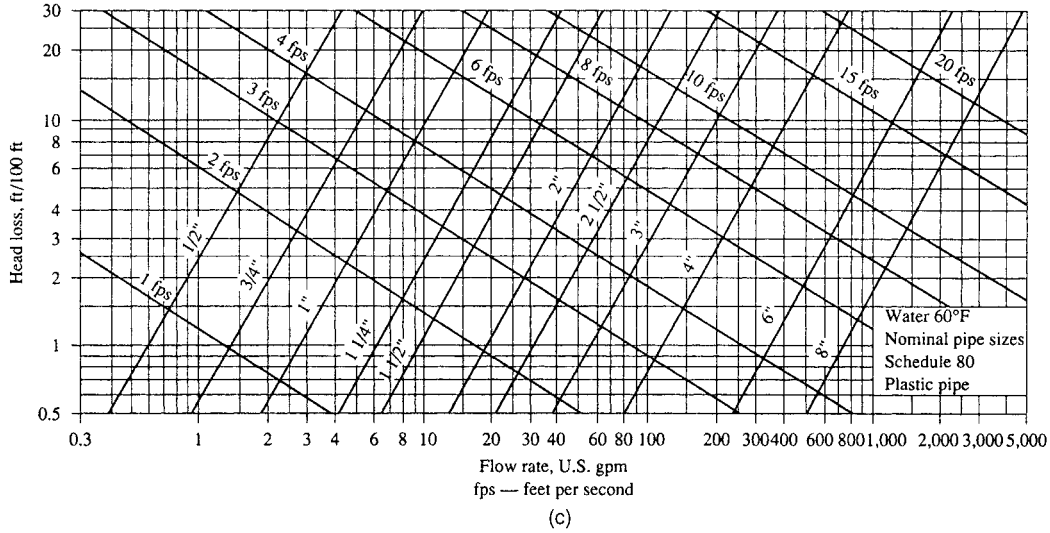
Main and Branch Pipes. In a piping circuit as shown in Figure 9.9.2(a), chilled water from a chiller or hot water from a boiler is often supplied to a *main pipe* and then distributed to *branch pipes* that



(a)



(b)



(c)

FIGURE 9.9.1 Friction chart for water in pipes: (a) steel pipe (schedule 40), (b) copper tubing, and (c) plastic pipe (schedule 80). (Souce: *ASHRAE Handbook 1993 Fundamentals*. Reprinted with permission.)

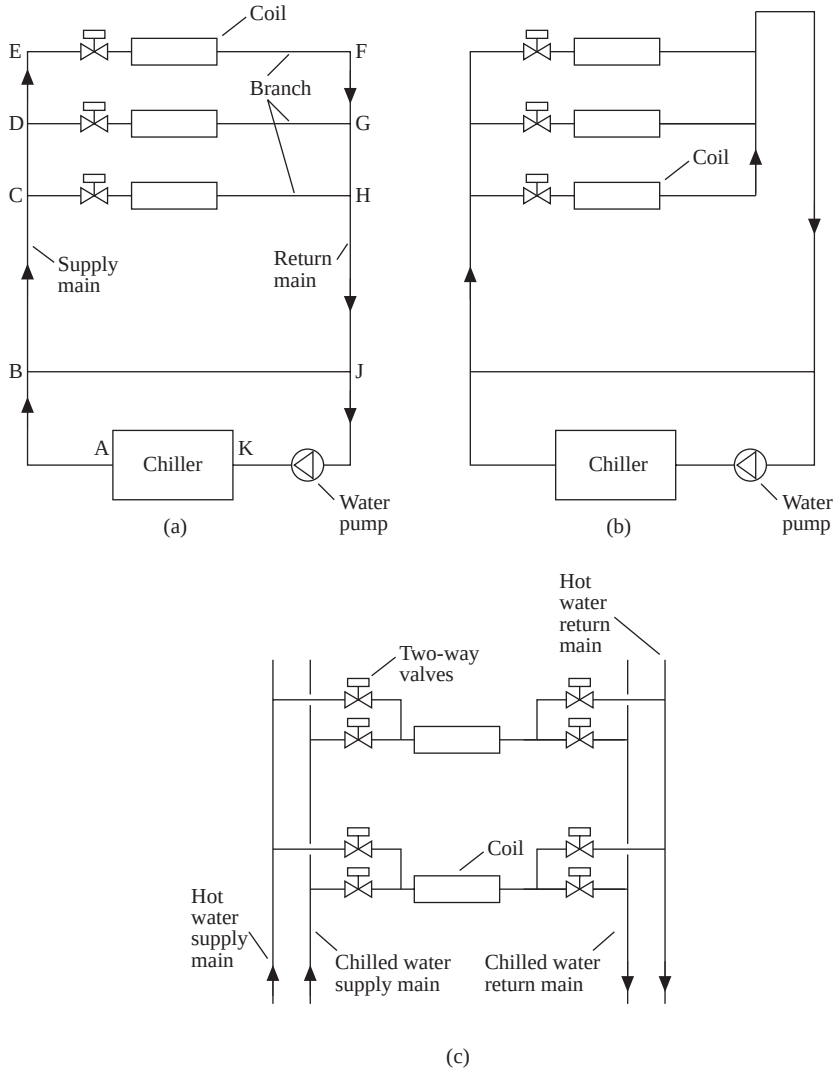


FIGURE 9.9.2 Piping arrangements: (a) two-pipe direct return system, (b) two-pipe reverse system, and (c) four-pipe system.

connect to coils and heat exchangers. Chilled or hot water from the coils and heat exchangers is accumulated by the return main pipe through return branch pipes and then returned to the chiller or boiler.

Constant Flow and Variable Flow. In a constant-flow water system, the volume flow rate at any cross-sectional plane of the supply and return mains remains constant during the entire operating period. In a variable-flow water system, the volume flow rate varies when the system load changes during the operating period.

Direct Return and Reverse Return. In a *direct return* system, the water supplies to and returns from various coils through various piping circuits. ABCHJKA, ... ABCDEFGHJKA are not equal in length, as shown in Figure 9.9.2(a). Water flow must be adjusted and balanced by using balance valves to provide required design flow rates at design conditions. In a *reverse-return* system, as shown in Figure 9.9.2(b), the piping lengths for various piping circuits including the branch and coil are almost equal. Water flow rates to various coils are easier to balance.

Two-Pipe or Four-Pipe. In a dual-temperature water system, the piping from the chiller or boiler to the coils can be either a *two-pipe* system with a supply main and return main as shown in Figure 9.9.2(a) or (b) or a *four-pipe* system with a chilled water supply main, a hot water supply main, a chilled water return main, and a hot water return main as shown in Figure 9.9.2(c). The two-pipe system needs a changeover from chilled to hot water and vice versa. A four-pipe system is more expensive to install.

Plant-Building Loop

System Description

The chillers/boilers in a central plant are often located in the basement, machinery room, or on the rooftop of the building. Generally, a fairly constant-volume water flow is required in the evaporator to protect it from freezing at part load. ASHRAE/IES Standard 90.1-1989 specifies that water systems using control valves to modulate the coil load must be designed for variable flow for energy savings.

Current chilled water systems or dual-temperature water systems often adopt a *plant-building loop* that consists of two piping loops as shown in Figure 9.9.3(a) for reliable and energy-efficient operation:

Plant Loop. A plant loop ABJKA is comprised of chiller(s)/boiler(s); plant chilled/hot water pump(s) usually of split-case, double suction, or end suction type; a diaphragm expansion tank to allow water expansion and contraction during water temperature changes; an air separator to purge dissolved air from water; corresponding piping and fittings; and control systems. The plant loop is operated at constant flow.

Building Loop. A building loop BCDEFGHJB contains coils; building circulating water pumps, often variable-speed pumps, one being a standby pump; corresponding pipes and fittings; and control systems. A differential pressure transmitter is often installed at the farthest end from the building pump between the supply and return mains. The building loop is operated at variable flow.

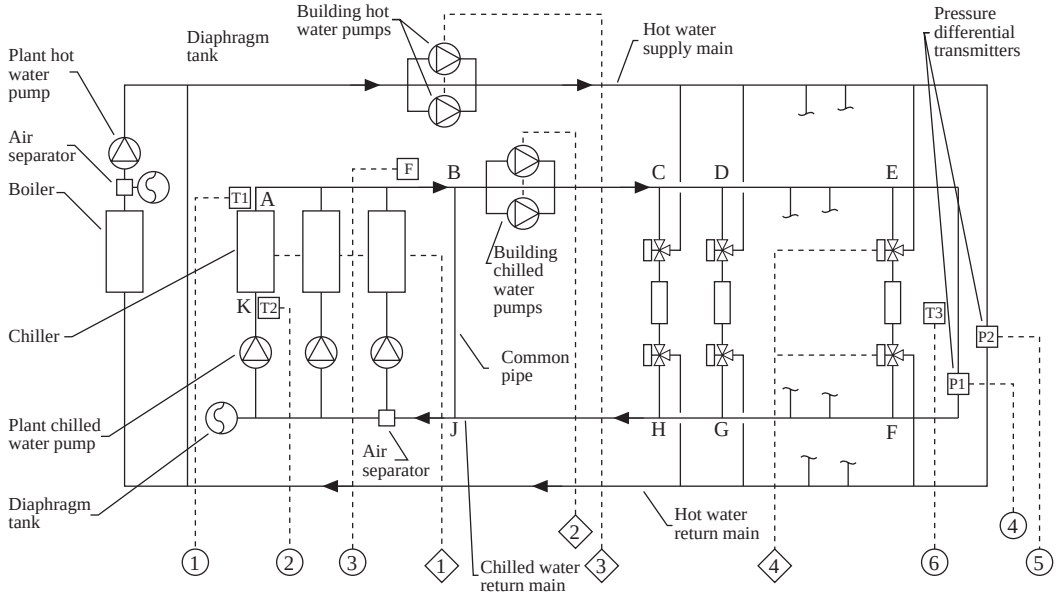
A short common pipe connects these two loops and combines them into a plant-building loop combination. It is also called a *primary-secondary loop*, with the plant loop the primary loop and the building loop the secondary loop. The location of the pump in a water system should be arranged so that the pressure at any point in the system is greater than atmospheric pressure to prevent air leaking into the system.

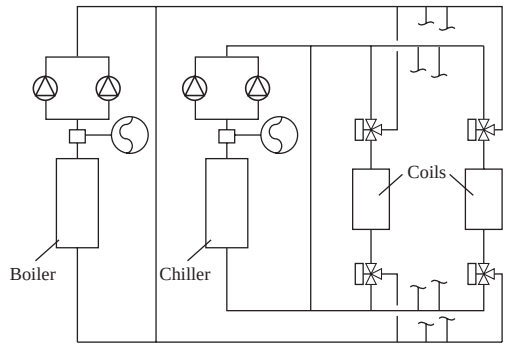
Operating Characteristics

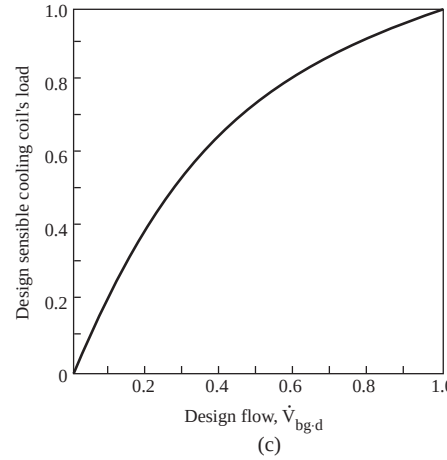
In a dual-temperature water system, when the coil load q_{cc}, in Btu/hr, drops, two-way valves close to a smaller opening. The pressure drop across the coils then increases. As soon as the increase of the pressure differential is sensed by the transmitter, a DDC controller reduces the flow rate of the variable-speed pump to match the reduction of the coil load during part load. The pressure differential transmitter functions similarly to a duct static pressure sensor in a VAV system.

Since the plant loop is at constant flow, excess chilled water bypasses the building loop and flows back to the chiller(s) through the common pipe. When the reduction of coil loads in the building loop is equal to or greater than the refrigeration capacity of a single chiller, the DDC controller will shut down a chiller and its associated chilled water pump. The operating characteristics of the hot water in a dual-temperature water system are similar to those in a chilled water system.

Due to the coil's heat transfer characteristics, the reduction of a certain fraction of the design sensible cooling coil load $q_{cs.d}$ does not equal the reduction of the same fraction of design water volume flow $\mathring{V}_{w.d}$ (Figure 9.9.3[c]). Roughly, a 0.7 $q_{cs.d}$ needs only a 0.45 $\mathring{V}_{w.d}$, and the temperature difference $\Delta T_w = (T_{wl} - T_{we})$ will be about 150% of that at the design condition. When an equal-percentage contour two-way control valve is used, the valve stem displacement of the cooling coil's two-way valve is approximately linearly proportional to the sensible cooling coil's load change.







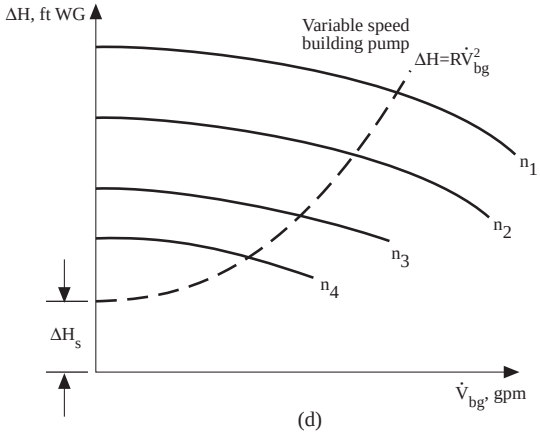


FIGURE 9.9.3 A dual-temperature water system: (a) plant-building loop, (b) single-loop, (c) design sensible load and design volume flow relationship, and (d) variable-speed building pump curves.

Comparison of a Single-Loop and a Plant-Building Loop

Compared with a single-loop water system as shown in Figure 9.9.3(b) that has only a single-stage water pump(s), the plant-building loop:

- Provides variable flow at the building loop and thus saves pump power at reduced flow during part load.
- Separates plant and building loops and makes design, operation, maintenance, and control simpler and more stable as stated in Carlson (1968). The pressure drop across the two-way control valve is considerably reduced as the control valve is only a component of the building loop.

Plant-Distribution-Building Loop

In universities, medical centers, and airports, buildings are often scattered from the central plant. A campus-type chilled water system using a plant-distribution-building loop is often reliable in operation, requires less maintenance, has minimal environmental impact, and sometimes provides energy cost savings.

In a plant-distribution-building loop, plant loop water is combined with many building loops through a distribution loop as shown in Figure 9.9.4(a). The water flow in the plant loop is constant, whereas in both distribution and building loops it is variable. Chilled water often leaves the chiller at a temperature of 40 to 42°F. It is extracted by the distribution pumps and forced into the distribution supply main. At the entrance of each building, chilled water is again extracted by the building pumps and supplied to the cooling coils in AHUs and fan-coil units. From the coils, chilled water is returned through the building return main and the distribution return main and enters the chillers at a temperature around 60°F at design volume flow. System performances of the plant and building loops in a plant-distribution-building loop are similar to those in a plant-building loop.

A campus-type central plant may transport several thousand gallons of chilled water to many buildings, and the farthest building may be several thousand feet away from the central plant. A smaller pressure gradient and end pressure differential between distribution supply and return mains, $\Delta H_{s.d}$, save energy. Meanwhile, a larger diameter of pipes is needed. For a distribution loop, a pressure drop of 0.5 to 1 ft head loss per 100 ft of piping length may be used. A low pressure differential may be offset at the control valves without affecting the normal operation of the coils. Usually, direct return is used for distribution loops.

Using a two-way distribution from the central plant with two separate supply and return loops often reduces the main pipe length and diameter.

A differential pressure transmitter may be located at the farthest end of the distribution mains. As the pressure loss in the building loop is taken care of by the variable-speed building pump(s), a set point of 5 ft WG pressure differential across the supply and return mains seems appropriate.

Chilled water and hot water distribution pipes are often mounted in underground accessible tunnels or trenches. They are well insulated except when chilled water return temperature $T_{ret} > 55$°F. Under such circumstances, the return main may not be insulated, depending on local conditions and cost analysis.

Many retrofits of campus-type chilled water systems required an existing refrigeration plant(s) and a new developed plant(s) connected to the same plant-distribution-building loop. A stand-alone DDC panel is often used to optimize the turning on and off of the chillers in the existing and developed plants according to the system loads and available sources (chillers).

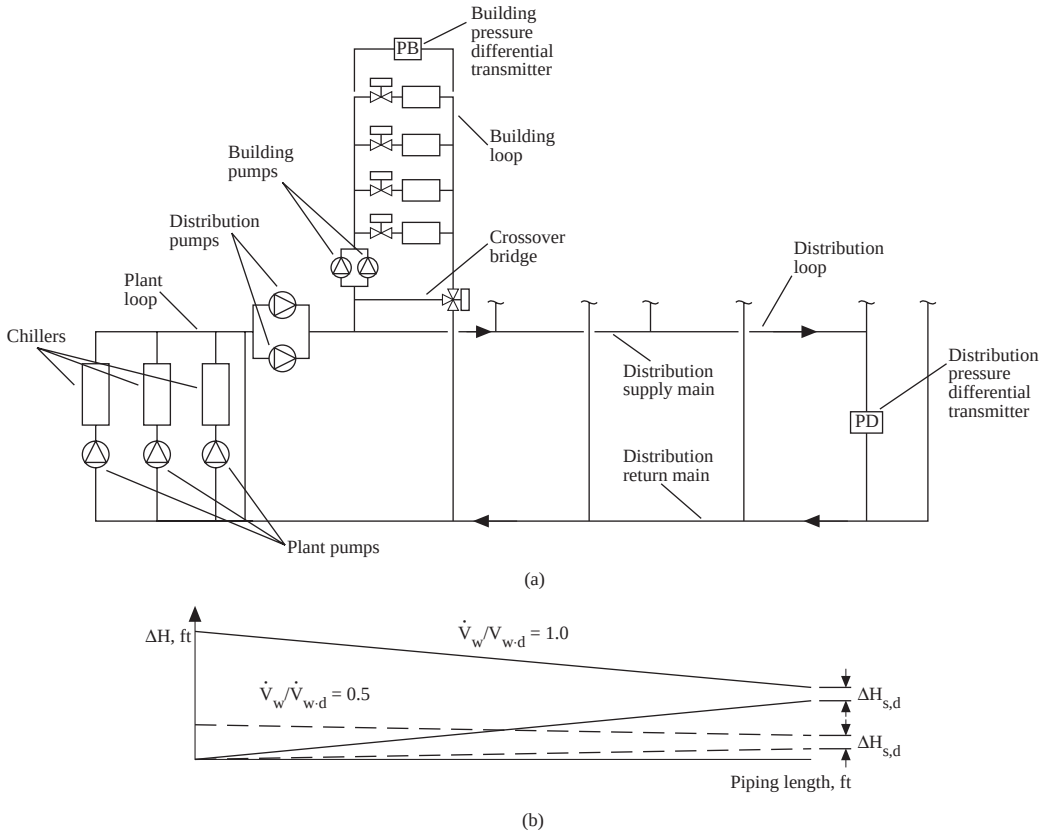


FIGURE 9.9.4 A chilled water system using a plant-distribution-building loop: (a) schematic diagram and (b) pressure gradient for distribution loop.

9.10 Heating Systems

Types of Heating Systems

According to EIA Commercial Buildings Characteristics 1992, for the 57.8 billion ft^2 of heated commercial buildings in the United States in 1992, the following types of heating systems were used:

Warm air heating systems using warm air furnace	27%
Hot water heating systems using boilers	33%
Heat pumps	13%
District heating	8%
Individual space heaters and others heaters	19%

Modera (1989) reported that nearly 50% of U.S. residential houses are using warm air heating systems with direct-fired warm air furnaces.

Warm Air Furnaces

A *warm air furnace* is a device in which gaseous or liquid fuel is directly fired or electric resistance heaters are used to heat the warm supply air. Natural gas, liquefied petroleum gas (LPG), oil, electric energy, or occasionally wood may be used as the fuel or energy input. Among these, natural gas is most widely used. In a warm air furnace, the warm air flow could be *upflow*, in which the warm air is discharged at the top, as shown in Figure 9.10.1(a) and (b); downflow, with the warm air discharged at the bottom; or *horizontal flow*, with the warm air discharged horizontally.

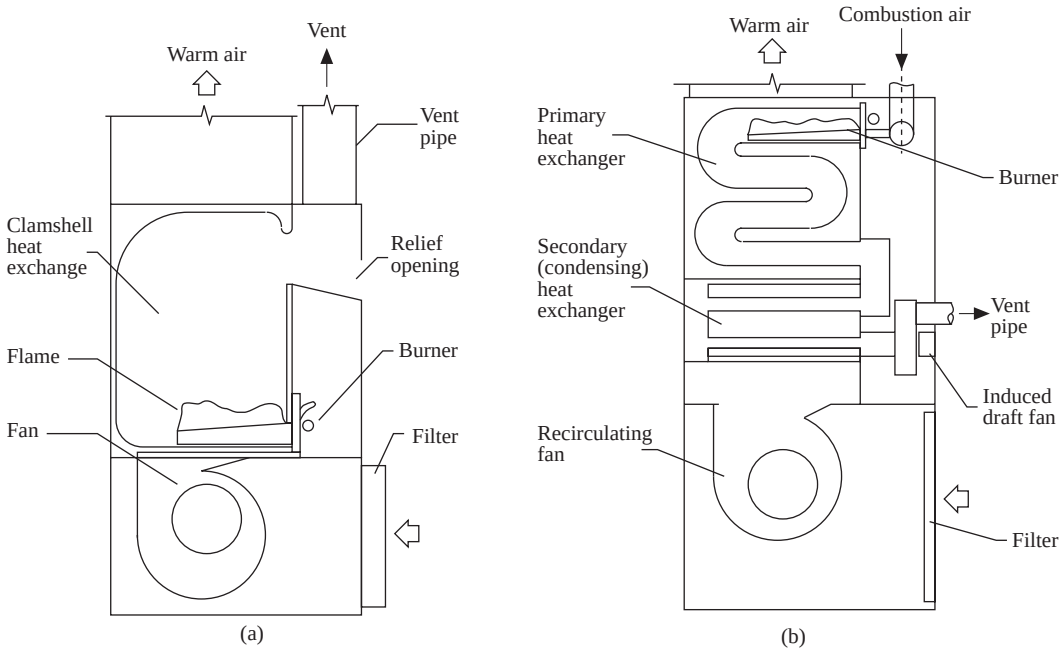


FIGURE 9.10.1 Upflow warm air gas furnace: (a) a natural-vent gas furnace and (b) a power-vent high-efficiency gas furnace.

Natural Vent Combustion Systems. There are two types of combustion systems in a natural gas-fired warm air furnace: natural vent or power vent combustion systems. In a *natural vent* or *atmospheric vent* combustion system, the buoyancy of the combustion products carries the flue gas flowing through the heat exchanger and draft hood, discharging from the chimney or vent pipe. The gas burner is an

atmospheric burner. In an atmospheric burner, air is exracted for combustion by the suction effect of the high-velocity discharged gas and the buoyance effect of the combustion air. An atmospheric burner can be either an in-shot or an up-shot burner or multiple ports. Atmospheric burners are simple, require only a minimal draft of air, and need sufficient gas pressure for normal functioning.

Two types of ignition have been used in burners: standing pilot and spark ignition. In *standing pilot ignition*, the small pilot flame is monitored by a sensor and the gas supply is shut off if the flame is extinguished. *Spark ignition* fires intermittently only when ignition is required. It saves gas fuel if the furnace is not operating.

In a natural vent combustion system, the heat exchanger is often made from cold-rolled steel or aluminized steel in the shape of a clamshell or S. A fan or blower is always used to force the recirculating air flowing over the heat exchanger and distribute the heated air to the conditioned space. A low-efficiency disposable air filter is often located upstream of the fan to remove dust from the recirculating air. A draft hood is also installed to connect the flue gas exit at the top of the heat exchanger to a vent pipe or chimney. A relief air opening is employed to guarantee that the pressure at the flue gas exit is atmospheric and operates safely even if the chimney is blocked. The outer casing of the furnace is generally made of heavy-gauge steel with access panels.

Power Vent Combustion Systems. In a *power vent* combustion system, either a forced draft fan is used to supply the combustion air or an induced draft fan is used to induce the flue gas to the vent pipe or chimney. A power vent is often used for a large gas furnace or a high-efficiency gas furnace with condensing heat exchangers.

Gas burners in a power vent system are called *power burners*. The gas supply to the power burner is controlled by a pressure regulator and a gas valve to control the firing rate. Intermittent spark ignition and *hot surface ignition* that ignites the main burners directly are often used.

Usually, there are two heat exchangers in a power vent combustion system: a primary heat exchanger and a secondary or condensing heat exchanger. The primary heat exchanger constitutes the heating surface of the combustion chamber. When the water vapor in the flue gas is condensed by indirect contact with the recirculating air, part of the latent heat of condensation released is absorbed by the air. Thus the furnace efficiency is increased in the *secondary* or *condensing heat exchanger.* Both primary and secondary heat exchangers are made from corrosion-resistant steel. A fan is also used to force the recirculating air to flow over the heat exchangers and to distribute the heated air to the conditioned space.

Most natural gas furnaces can use LPG. LPG needs a pressure of 10 in. WG at the manifold, compared with 3 to 4 in. for natural gas. It also needs more primary air for gas burners. Oil furnaces are usually of forced draft and installed with pressure-atomizing burners. The oil pressure and the orifice size of the injection nozzle control the firing rate.

Furnace Performance Indices. The performance of a gas-fired furnace is usually assessed by the following indices:

- Thermal efficiency E_t, in percent, is the ratio of the energy output of heated air or water to the fuel energy input during specific test periods using the same units:

$$E_t = 100(\text{fuel energy output})/(\text{fuel energy input}) \tag{9.10.1}$$

- Annual fuel utilization efficiency (AFUE), in percent, is the ratio of the annual output energy from heated air or water to the annual input energy using the same units:

$$\text{AFUE} = (100 \text{ annual output energy})/(\text{annual input energy}) \tag{9.10.2}$$

- Steady-state efficiency (SSE) is the efficiency of a given furnace according to an ANSI test procedure, in percent:

$$\text{SSE} = 100(\text{fuel input} - \text{fuel loss})/(\text{fuel input}) \tag{9.10.3}$$

Jakob et al. (1986) and Locklin et al. (1987), in a report on ASHRAE Special Project SP43, gave the following performance indices based on a nighttime setback period of 8 hr with a setback temperature of 10°F:

Description	AFUE (%)	SSE (%)
Natural vent		
Pilot ignition	64.5	77
Intermittent ignition	69	77
Intermittent ignition plus vent damper	78	77
Power vent		
Noncondensing	81.5	82.5
Condensing	92.5	93

ASHRAE/IES Standard 90.1-1989 specifies a minimum AFUE of 78% for both gas-fired and oil-fired furnaces of heating capacity <225,000 Btu/hr. For a gas-fired warm air furnace ≥225,000 Btu/hr at maximum rating capacity (steady state) the minimum E_t is 80%, and at minimum rating capacity, E_t is 78%.

Operation and Control. Gas is usually brought from the main to the pressure regulator, where its pressure is reduced to 3.5 in. WG. Gas then flows through a valve and mixes with the necessary amount of outside primary air. The mixture mixes again with the ambient secondary air and is burned. The combustion products flow through the heat exchanger(s) due either to natural draft or power vent by a fan. The flue gas is then vented to the outside atmosphere through a vent pipe or chimney.

Recirculating air is pulled from the conditioned space through the return duct and is mixed with the outdoor air. The mixture is forced through the heat exchanger(s) by a fan and is then heated and distributed to the conditioned space. The fan is often started 1 min after the burner is fired in order to prevent a cold air supply.

At part-load operation, the reduction of the heating capacity of the warm air furnace is usually controlled by the gas valve and the ignition device. For small furnaces, the gas valve is often operated at on–off control. For large furnaces, a two-stage gas valve operates the furnace at 100, 50, and 0% heating capacity as required.

Low NO_x Emissions. NO_x means nitrogen oxides NO and NO_2. They are combustion products in the flue gas and become air pollutants with other emissions like SO_2 and CO when they are discharged to the atmosphere. NO_x cause ozone depletion as well as smog.

Southern California regulations required that NO_x emissions should be 30 ppm or less for gas-fired warm air furnaces and boilers. Many gas burner manufactures use induced flue gas recirculation to cool the burner's flame, a cyclonic-type burner to create a swirling high-velocity flame, and other technologies to reduce NO_x and other emissions while retaining high furnace and boiler efficiencies.

Hot Water Boilers

Types of Hot Water Boilers. A *hot water boiler* is an enclosed pressure vessel used as a heat source for space heating in which water is heated to a required temperature and pressure without evaporation. Hot water boilers are fabricated according to American Society of Mechanical Engineers (ASME) codes for boilers and pressure vessels. Boilers are generally rated on the basis of their gross output delivered at the boiler's outlet. Hot water boilers are available in standard sizes from 50 to 50,000 MBtu/hr (1 MBtu/hr = 1000 Btu/hr).

According to EIA Characteristics of Commercial Buildings (1991), the percentages of floor area served by different kinds of fuel used in hot water and steam boilers in 1989 in the United States are gas-fired, 69%; oil-fired, 19%; electric, 7%; others, 5%.

Hot water boilers can be classified as *low-pressure boilers*, whose working pressure does not exceed 160 psig and working temperature is 250°F or less, and *medium-* and *high-pressure boilers*, whose working pressure is above 160 psig and working temperature above 250°F. Most of the hot water boilers are low-pressure boilers except those in campus-type or district water heating systems.

Based on their construction and material, hot water boilers can be classified as fire tube boilers, water tube boilers, cast iron sectional boilers, and electric boilers. Water tube boilers are used mainly to generate steam. Cast iron sectional boilers consist of many vertical inverted U-shaped cast iron hollow sections. They are lower in efficiency and used mainly for residential and small commercial buildings. Electric boilers are limited in applications because of their higher energy cost in many locations in the United States.

Scotch Marine Boiler. In a *fire tube* hot water boiler, the combustion chamber and flue gas passages are in tubes. These tubes are enclosed by an outer shell filled with water. A recently developed fire tube model is the modified Scotch Marine packaged boiler, which is a compact, efficient, and popular hot water boiler used today.

A *Scotch Marine* boiler, as shown in Figure 9.10.2, consists of a gas, oil, or gas/oil burner, a mechanical forced-draft fan, a combustion chamber, fire tubes, flue gas vent, outer shell, and corresponding control system. A packaged boiler is a one-piece factory-assembled boiler.

This figure is not available.

FIGURE 9.10.2 A Scotch Marine packaged boiler.

Gas or oil and air are measured, mixed, and injected into the combustion chamber in which they are initially ignited by an ignition device. The mixture burns and the combustion process is sustained when there is a high enough temperature in the combustion chamber. Because the number of fire tubes decreases continuously in the second, third, and fourth passes due to the volume contraction at lower flue gas temperatures, the gas velocity and heat transfer coefficients are maintained at reasonably high values.

Return water enters the side of the boiler. It sinks to the bottom, rises again after heating, and then discharges at the top outlet.

The dew point of the flue gas is often 130°F. If the temperature of the return water in a *condensing boiler* is below 125°F, it can be used as a condensing cooling medium to condense the water vapor contained in the flue gas. Latent heat of condensation of water vapor can then be recovered. Corrosion in the condensing heat exchanger and the flue gas passage should be avoided.

The *chimney*, or *stack*, is the vertical pipe or structure used to discharge flue gas, which usually has a temperature rise of 300 to 400°F above the ambient temperature. The chimney or stack should be extended to a certain height above adjacent buildings according to local codes.

Operation and Safety Controls. During part-load operation, reduction of heating capacity is achieved by sensing the temperature of return water and controlling the firing rate of the gas burners in on–off, high–low–off, or modulating modes.

For gas burners, two pressure sensors are often provided to maintain the gas pressure within a narrow range. For modulating controls, the ratio of maximum to minimum input is called the *turn-down* ratio. The minimum input is usually 5 to 25% of the maximum input, that is, a turn-down ratio of 20:1 to 4:1. The boiler should be shut off if the input is less than the minimum.

Pressure and temperature relief valves should be installed in each boiler. An additional high limit control is often equipped to shut down the boiler in case the sensed water pressure or temperature exceeds the predetermined limits. A flame detector is often used to monitor the flame and an airflow sensor to verify the combustion airflow. As soon as the flame is extinguished or the combustion airflow is not sensed, the fuel valve closes and the boiler is shut down. ASHRAE/IES Standard 90.1-1989 specifies that gas-fired boilers should have a minimum AFUE of 80%.

Low-Pressure Warm Air Heating Systems

A *low-pressure warm air heating system* is often equipped with an upflow gas-fired furnace having a furnace heat capacity Q_f to air flow $\dot{V}_a$ ratio, $Q_f / \dot{V}_a$, of 50 to 70 Btu/hr.cfm and a temperature rise immediately after the furnace of 50 to 70°F. The external pressure loss of the supply and return duct system is usually not higher than 0.5 in. WG. The supply temperature differential ($T_s - T_r$) is often 20 to 35°F. The heating system is often integrated with a cooling system, forming a heating/cooling system.

Low-pressure warm air heating systems usually have a heating capacity not exceeding 100,000 Btu/hr. They are often used in residences and sometimes in small commercial buildings.

System Characteristics. A low-pressure warm air heating system is equipped with either supply and return ducts or a supply duct and a return plenum. Recirculating air is then returned from living, dining, bed, and study rooms to the return plenum through door undercuts in case the doors are closed.

The location of the furnace has a significant effect on the efficiency of the heating system. According to Locklin et al. (1987), if the gas furnace of a low-pressure warm air heating system is installed in a closet and its supply duct is mounted inside the conditioned space, its system efficiency may be 20% higher than that of installations whose furnace and supply duct are in the attic or basement.

When a low-pressure warm air heating system is operating, the supply duct leakage in the attic or basement raises its pressure to a positive value and promotes exfiltration. Return duct leakage extracts the ambient air, lowers the attic or basement pressure to a negative value, and induces infiltration. Gammage et al. (1986) reported that both types of leakage increase the whole house infiltration to 0.78 ach when the low-pressure warm air heating system is operating. The infiltration rate is only 0.44 ach when the low-pressure warm air heating system is shut off.

If the supply temperature differential $\Delta T_s = (T_s - T_r)$ exceeds 30°F, or if there is a high ceiling, thermal stratification may form in the conditioned space. Greater supply volume flow rates and suitable locations of the supply and return outlets may reduce thermal stratification and vertical temperature difference.

Part-Load Operation and Control. For a low-pressure warm air heating system, a space thermostat is often used to control the gas valve of the furnace operated in on–off or high–low–off mode. The proportion of on and off times in an operating cycle can be varied to meet the change of space heating load. The time period of an on–off operating cycle is usually 5 to 15 min.

A warm-air heating system that has an external pressure higher than 0.5 in. WG is often integrated with a cooling system and becomes a part of an air-conditioning system.

Low-Temperature Hot Water Heating System Using Fin-Tube Heaters

In a *low-temperature hot water heating system*, the operating temperature is 250°F or less with a maximum working pressure not exceeding 150 psig, usually less than 30 psig. Low-temperature hot water heating systems are widely used for space heating in residences and commercial buildings.

Fin-Tube Heaters. A *fin-tube heater* is a device installed directly inside the conditioned space to add heat to the space through radiant and convective heat transfer. A fin-tube heater consists of a finned-tube element and an outer casing as shown in Figure 9.10.3(a). The tubes are often made of copper and steel. Copper tubes are generally 0.75, 1, and 1.25 in. in diameter and steel tubes 1.25 and 2 in. in diameter. The fins are usually made of aluminum for copper tubes and of steel for steel tubes. Fin density may vary from 24 to 60 fins per foot. A fin heater may have a maximum length of 12 ft. The outer casing of a finned-tube heater always has a bottom inlet and top outlet for better convection.

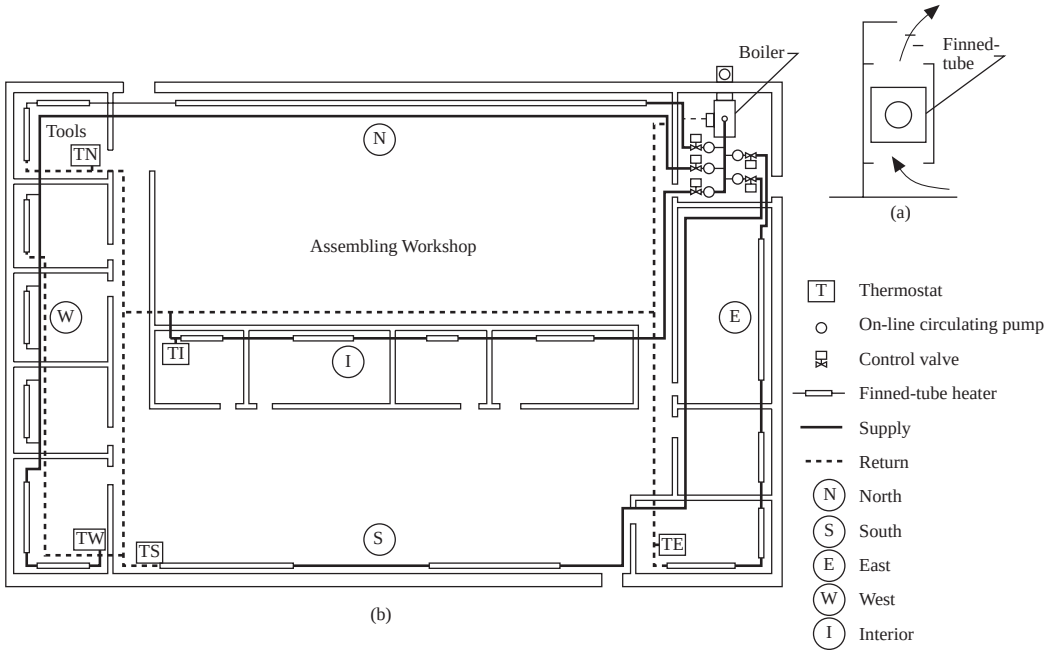


FIGURE 9.10.3 A two-pipe individual-loop low-temperature hot water heating system: (a) finned-tube heater and (b) piping layout.

The most widely used finned-tube heater is the *baseboard heater*, which is often mounted on cold walls at a level 7 to 10 in. from the floor. It is usually 3 in. deep and has one fin-tube row. A wall finned-tube heater has a greater height. A convector has a cabinet-type enclosure and is often installed under the windowsill.

Two-Pipe Individual-Loop Systems. Current low-temperature hot water heating systems using finned-tube heaters are often equipped with zone controls. Zone controls can avoid overheating rooms facing south and underheating rooms facing north because of the effects of solar radiation.

Figure 9.10.3(b) shows the piping layout of a *two-pipe individual-loop system* that is widely used in low-temperature hot water heating systems. Two-pipe means that there are a supply main and a return main pipe instead of one common main for both supply and return. Individual-loop means that there is an individual loop for each control zone. Several finned-tube heaters in a large room can be connected in series, while finned-tube heaters in several small rooms can be connected in reverse return arrangement.

The sizing of low-temperature hot water pipes is usually based on a pressure drop of 1 to 3 ft per 100 ft of pipe length. For a small low-temperature hot water heating system, an open-type expansion tank is often used. A diaphragm tank is often used for a large system. On-line circulating pumps with low head are often used.

Part Load and Control. Usually a hot water sensor located at the exit of the hot water boiler is used to control the firing rate of the boiler at part-load operation. Its set point is usually reset according to the outdoor temperature. Zone control is provided by sensing the return hot water temperature from each individual loop or zone and then varying the water volume flow rate supplied to that zone by modulating the speed of each corresponding on-line circulating pump or its control valve. For hot water heating systems using multiple boilers, on and off for each specific boiler depend not only on the heating demand, but also on minimizing the energy cost.

Infrared Heating

Infrared heating is a process that uses radiant heat transfer from a gas-fired or electrically heated high-temperature device to provide space heating on a localized area for the health and comfort of the occupants or to maintain a suitable indoor environment for a manufacturing process.

An *infrared heater* is a device used to provide infrared heating. Heat radiates from an infrared heater in the form of electromagnetic waves and scatters in all directions. Most infrared heaters have reflectors to focus the radiant beam onto a localized area. Therefore, they are often called *beam radiant heaters.* Infrared heaters are widely used in high-ceiling supermarkets, factories, warehouses, gymnasiums, skating rinks, and outdoor loading docks.

Gas Infrared Heaters. Infrared heaters can be divided into two categories: gas and electric infrared heaters. *Gas infrared heaters* are again divided into porous matrix gas infrared heaters and indirect gas infrared heaters. In a *porous matrix gas infrared heater,* a gas and air mixture is supplied and distributed evenly through a porous ceramic, a stainless steel panel, or a metallic screen, which is exposed to the ambient air and backed by a reflector. Combustion takes place at the exposed surface with a maximum temperature of about 1600°F. An indirect infrared heater consists of a burner, a radiating tube, and a reflector. Combustion takes place inside the radiating tube at a temperature not exceeding 1200°F.

Gas infrared heaters are usually vented and have a small conversion efficiency. Only 10 to 20% of the input energy of an open combustion gas infrared heater is radiated in the form of infrared radiant energy. Usually 4 cfm of combustion air is required for 1000 Btu/hr gas input. A thermostat often controls a gas valve in on–off mode. For standing pilot ignition, a sensor and a controller are used to cut off the gas supply if the flame is extinguished.

Electric Infrared Heaters. An *electric infrared heater* is usually made of nickel–chromium wire or tungsten filaments mounted inside an electrically insulated metal tube or quartz tube with or without inert gas. The heater also contains a reflector that directs the radiant beam to the localized area requiring heating. Nickel–chromium wires often operate at a temperature of 1200 to 1800°F. A thermostat is also used to switch on or cut off the electric current. An electric infrared heater has a far higher conversion efficiency and is cleaner and more easily managed.

Design Considerations. An acceptable radiative temperature increase ($T_{rad} - T_r$) of 20 to 25°F is often adopted for normal clothed occupants using infrared heating. The corresponding required watt density

for infrared heaters is 30 to 37 W/ft^2. At a mounting height of 11 ft, two heaters having a spacing of 6.5 ft can provide a watt density of 33 W/ft^2 and cover an area of 12 × 13 ft. The mounting height of the infrared heaters should not be lower than 10 ft. Otherwise the occupants may feel discomfort from the overhead radiant beam. Refer to Grimm and Rosaler (1990), *Handbook of HVAC Design*, for details.

Gas and electric infrared heaters should not be used in places where there is danger of ignitable gas or materials that may decompose into toxic gases.

9.11 Refrigeration Systems

Classifications of Refrigeration Systems

Most of the refrigeration systems used for air-conditioning are vapor compression systems. Because of the increase in the energy cost of natural gas in the 1980s, the application of absorption refrigeration systems has dropped sharply. According to Commercial Buildings Characteristics 1992, absorption refrigeration system have a weight of less than 3% of the total amount of refrigeration used in commercial buildings in the United States. Air expansion refrigeration systems are used mainly in aircraft and cryogenics.

Refrigeration systems used for air-conditioning can be classified mainly in the following categories:

- Direct expansion (DX) systems and heat pumps
- Centrifugal chillers
- Screw chillers
- Absorption systems

Each can be either a single-stage or a multistage system.

Direct Expansion Refrigeration Systems

A *direct expansion refrigeration (DX) system*, or simply *DX system*, is part of the packaged air-conditioning system. The DX coil in the packaged unit is used to cool and dehumidify the air directly as shown in Figure 9.11.1(a). According to EIA Commercial Buildings Characteristics 1992, about 74% of the floor space of commercial buildings in the United States was cooled by DX refrigeration systems.

Refrigerants R-22 and R-134a are widely used. Azeotropics and near azeotropics are the refrigerants often used for low-evaporating-temperature systems like those in supermarkets. Because of the limitation of the size of the air system, the refrigeration capacity of DX systems is usually 3 to 100 tons.

Components and Accessories. In addition to the DX coil, a DX refrigeration system has the following components and accessories:

- *Compressor(s)* — Both reciprocating and scroll compressors are widely used in DX systems. Scroll compressors are gradually replacing reciprocating compressors because they have fewer parts and comparatively higher efficiency. For large DX systems, multiple compressors are adopted.
- *Condensers* — Most DX systems in rooftop packaged units are air cooled. Water-cooled condensers are adopted mainly for DX systems in indoor packaged units due to their compact volume. Evaporative-cooled condensers are also available.
- *Refrigeration feed* — Thermostatic expansion valves are widely used as the throttling and refrigerant flow control devices in medium and large DX systems, whereas capillary tubes are used in small and medium-sized systems.
- *Oil lubrication* — R-22 is partly miscible with mineral oil. Since R-134a is not miscible with mineral oil, synthetic polyolester oil should be used. For medium and large reciprocating compressors, an oil pump of vane, gear, or centrifugal type is used to force the lubricating oil to the bearings and moving surfaces via grooves. For small reciprocating compressors, splash lubrication using the rotation of the crankshaft and the connecting rod to splash oil onto the bearing surface and the cylinder walls is used.

A scroll compressor is often equipped with a centrifugal oil pump to force the oil to lubricate the orbiting scroll journal bearing and motor bearing. For the scroll contact surfaces, lubrication is provided by the small amount of oil entrained in the suction vapor.

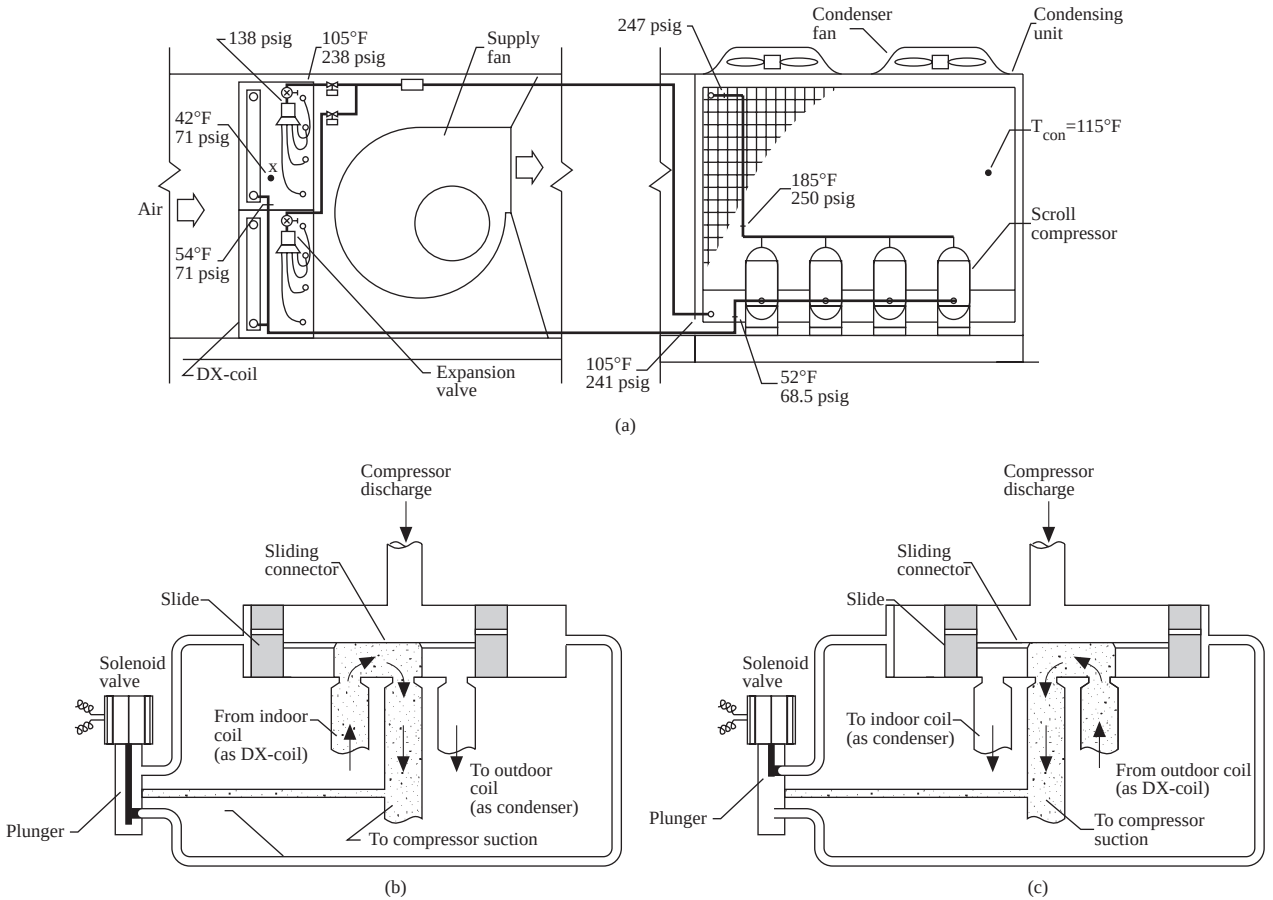


FIGURE 9.11.1 A DX refrigeration system: (a) schematic diagram: (b) four-way reversing valve, cooling mode; and (c) four-way reversing valve, heating mode.

- *Refrigerant piping* — Refrigerant piping transports refrigerant through the compressor, condenser, expansion valve, and DX coil to provide the required refrigeration effect. As shown in Figure 9.11.1(a), from the exit of the DX coil to the inlet of the compressor(s) is the *suction line*. From the outlet of the compressor to the inlet of the air-cooled condenser is the *discharge line*. From the exit of the condenser to the inlet of the expansion valve is the *liquid line*.

Halocarbon refrigerant pipes are mainly made of copper tubes of L type. In a packaged unit, refrigerant pipes are usually sized and connected in the factory. However, the refrigerant pipes in field-built and split DX systems for R-22 are sized on the basis of a pressure drop of 2.91 psi corresponding to a change of saturated temperature ΔT_{suc} of 2°F at 40°F for the suction line and a pressure drop of 3.05 psi corresponding to 1°F at 105°F for the discharge and liquid line. The pressure drop includes pressure losses of pipe and fittings. Refrigerant pipes should also be sized to bring back the entrained oil from the DX coil and condenser through the discharge and suction lines.

Accessories include a filter dryer to remove moisture from the refrigerant, strainer to remove foreign matter, and sight glass to observe the condition of refrigerant flow (whether bubbles are seen because of the presence of flash gas in the liquid line).

Capacity Controls. In DX systems, control of the mass flow rate of refrigerant through the compressor(s) is often used as the primary refrigeration capacity control. Row or intertwined face control at the DX coil is also used in conjunction with the capacity control of the compressor(s).

Three methods of capacity controls are widely used for reciprocating and scroll compressors in DX systems:

- *On–off control* — Starting or stopping the compressor is a kind of step control of the refrigerant flow to the compressor. It is simple and inexpensive, but there is a 100% variation in capacity for DX systems installed with only a single compressor. On–off control is widely used for small systems or DX systems with multiple compressors.
- *Cylinder unloader* — For a reciprocating compressor having multiple cylinders, a cylinder unloader mechanism bypasses the compressed gas from one, two, or three cylinders to the suction chamber to reduce the refrigeration capacity and compressing energy.
- *Speed modulation* — A two-speed motor is often used to drive scroll or reciprocating compressors so that the capacity can be reduced 50% at lower speed.

Safety Controls. In *low- and high-pressure control*, the compressor is stopped when suction pressure drops below a preset value, the cut-in pressure, or the discharge pressure of the hot gas approaches a dangerous level, the cut-out pressure.

In *low-temperature control*, a sensor is mounted on the outer pipe surface of the DX coil. When the surface temperature drops below 32°F, the controller stops the compressor to prevent frosting.

If the pressure of the oil discharged from the pump does not reach a predetermined level within a certain period, a mechanism in *oil pressure failure control* opens the circuit contact and stops the compressor.

In *motor overload control*, a sensor is used to measure the temperature of the electric winding or the electric current to protect the motor from overheating and overloading.

Pump-down control is an effective means of preventing the migration of the refrigerant from the DX coil (evaporator) to the crankcase of the reciprocating compressor. This prevents mixing of refrigerant and oil to form slugs, which may damage the compressor.

When a rise of suction pressure is sensed by a sensor, a DDC controller opens a solenoid valve and the liquid refrigerant enters the DX coil. As the buildup of vapor pressure exceeds the cut-in pressure, the compressor starts. When the DX system needs to shut down, the solenoid valve is closed first; the compressor still pumps the gaseous refrigerant to the condenser. As the suction pressure drops below the cut-in pressure, the compressor stops.

Full- and Part-Load Operations. Consider a DX system in a rooftop packaged unit using four scroll compressors, each with a refrigeration capacity of 10 tons at full load. Performance curves of the condensing unit and the DX coil of this DX system are shown in Figure 9.11.2(a). A DDC controller actuates on–off for these scroll compressors based on the signal from a discharge air temperature T_{dis} sensor.

On a hot summer day, when the rooftop packaged unit starts with a DX coil load or refrigeration capacity, $q_{rc} \approx 40$ tons, all four scroll compressors are operating. The operating point is at A′ with a suction temperature T_{suc} of about 42°F, and the discharge air temperature T_{dis} is maintained around 53°F. As the space cooling load q_{rc} as well as T_{dis} decreases, the required DX coil load q_{rl} drops to 35 tons. Less evaporation in the DX coil causes a decrease of T_{suc} to about 40°F, and the operating point may move downward to point A with a DX coil refrigeration capacity of 39 tons. Since $q_{rc} > q_{rl}$, T_{dis} drops continually until it reaches 50°F, point A in Figure 9.11.2(b), and the DDC controller shuts down one of the scroll compressors. The operating point immediately shifts to B′ on the three-compressor curve.

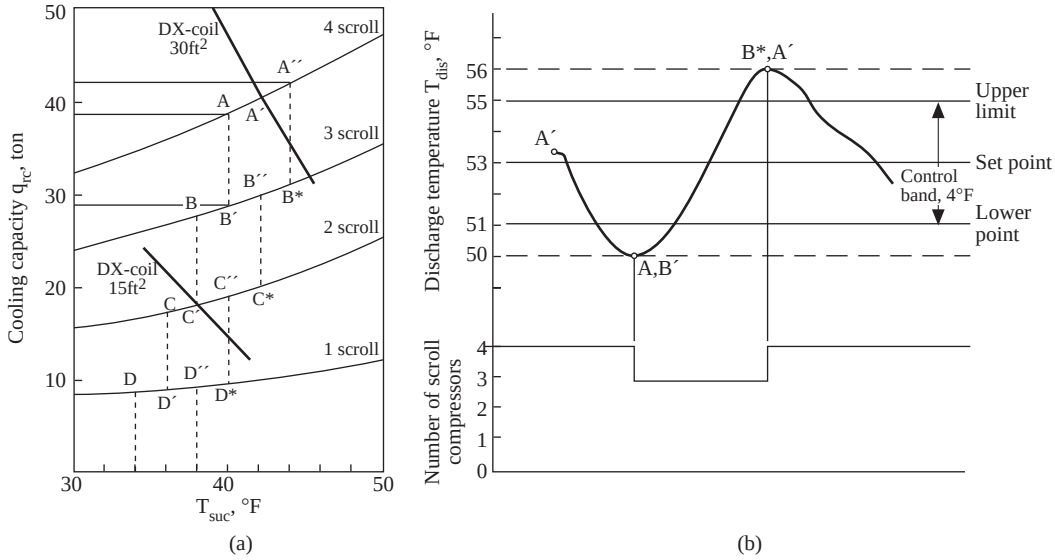


FIGURE 9.11.2 Capacity control of a DX refrigeration system: (a) performance curves and operating points and (b) locus of control point.

Because the refrigeration capacity at point B′ q_{rc} is 29 tons, which is less than the required q_{rl} = 35 tons, both T_{dis} and T_{suc} rise. When the operating point moves up to B* and T_{dis} reaches 56°F, the DDC controller starts all four scroll compressors at operating point A″ with a refrigeration capacity of 42 tons. Since $q_{rc} > q_{rl}$, the operating point again moves downward along the four-compressor curve and forms an operating cycle A″AB′ and B*. The timing of the operating period on four- or three-compressor performance curves balances any required q_{rl} between 29 and 42 tons. Less evaporation at part load in the DX coil results in a greater superheating region and therefore less refrigeration capacity to balance the reduction of refrigeration capacity of the compressor(s) as well as the condensing unit. The condition will be similar when $q_{rl} < 30$ tons, only three- or two-compressor, or two- or one-compressor, or even on–off of one compressor forms an operating cycle.

Main Problems in DX Systems

- *Liquid slugging* is formed by a mixture of liquid refrigerant and oil. It is formed because of the flooding back of liquid refrigerant from the DX coil to the crankcase of the reciprocating compressor due to insufficient superheating. It also may be caused by migration of liquid refrig-

erant from the warmer indoor DX coil to the colder outdoor compressor during the shut-down period in a split packaged unit. Liquid slugging dilutes the lubricating oil and causes serious loss of oil in the crankcase of the reciprocating compressor. Liquid slugging is incompressible. When it enters the compression chamber of a reciprocating compressor, it may damage the valve, piston, and other components. Pump-down control and installation of a crankcase heater are effective means of preventing liquid refrigerant migration and flooding back.

- *Compressor short cycling* — For on–off control, too short a cycle, such as less than 3 min, may pump oil away from the compressor or damage system components. It is due mainly to a too close low-pressure control differential or to reduced air flow.
- *Defrosting* — If the surface of a DX coil is 32°F or lower, frost accumulates on it. Frost blocks the air passage and reduces the rate of heat transfer. It should be removed periodically. The process of removing frost is called defrosting.

  Air at a temperature above 36°F, hot gas inside the refrigerant tubes and an installed electric heating element can be used for defrosting. The defrosting cycle is often controlled by sensing the temperature or pressure difference of air entering the DX coil during a fixed time interval.
- *Refrigerant charge* — Insufficient refrigerant charge causes lower refrigeration capacity, lower suction temperature, and short on–off cycles. Overcharging refrigerant may cause a higher condensing pressure because part of the condensing surface becomes flooded with liquid refrigerant.

Heat Pumps

A *heat pump* in the form of a packaged unit is also a *heat pump system*. A heat pump can either extract heat from a heat source and reject heat to air and water at a higher temperature for heating, or provide refrigeration at a lower temperature and reject condensing heat at a higher temperature for cooling. During summer, the heat extraction, or refrigeration effect, is the useful effect for cooling in a heat pump. In winter, the rejected heat and the heat from a supplementary heater provide heating in a heat pump.

There are three types of heat pumps: air-source, water-source, and ground-coupled heat pumps. Ground-coupled heat pumps have limited applications. Water-source heat pump systems are covered in detail in a later section.

Air-Source Heat Pump. An *air-source heat pump*, or *air-to-air heat pump*, is a DX system with an additional four-way reversing valve to change the refrigerant flow from cooling mode in summer to heating mode in winter and vice versa. The variation in connections between four means of refrigerant flow — compressor suction, compressor discharge, DX coil exit, and condenser inlet — causes the function of the indoor and outdoor coils to reverse. In an air-source heat pump, the coil used to cool or to heat the recirculating/outdoor air is called the *indoor coil*. The coil used to reject heat to or absorb heat from the outside atmosphere is called the *outdoor coil*. A short capillary or restrict tube is often used instead of a thermostatic expansion valve. Both reciprocating and scroll compressors are used in air-source heat pumps. R-22 is the refrigerant widely used. Currently available air-source heat pumps usually have a cooling capacity of $1^1/_2$ to 40 tons.

Cooling and Heating Mode Operation. In *cooling mode operation*, as shown in Figure 9.11.1(b), the solenoid valve is deenergized and drops downward. The high-pressure hot gas pushes the sliding connector to the left end. The compressor discharge connects to the outdoor coil, and the indoor coil connects to the compressor inlet.

In *heating mode operation*, as shown in Figure 9.11.1(c), the solenoid plunger moves upward and pushes the slide connector to the right-hand side. The compressor discharge connects to the indoor coil, and the outdoor coil exit connects to the compressor suction.

System Performance. The performance of an air-source heat pump depends on the outdoor air temperature T_o, in °F as well as the required space heating load q_{rh}. During cooling mode operation, both the refrigeration capacity q_{rc}, in Btu/hr, and EER for the heat pump EER_{hp}, in Btu/hr.W, increase as T_o drops.

During heating mode operation, the heating capacity q_{hp}, in Btu/hr, and COP_{hp} decrease, and q_{rh} increases as the T_o drops. When $q_{rh} > q_{hp}$, supplementary heating is required. If COP_{hp} drops below 1, electric heating may be more economical than a heat pump.

If on–off is used for compressor capacity control for an air-source heat pump in a split packaged unit, refrigerant tends to migrate from the warmer outdoor coil to the cooler indoor coil in summer and from the warmer indoor coil to the cooler outdoor coil in winter during the off period. When the compressor starts again, 2 to 5 min of reduced capacity is experienced before the heat pump can be operated at full capacity. Such a loss is called a *cycling loss*.

In winter, most air-source heat pumps switch from the heating mode to cooling mode operation and force the hot gas to the outdoor coil to melt frost. After the frost is melted, the heat pump is switched back to heating mode operation. During defrosting, supplementary electric heating is often necessary to prevent a cold air supply from the air-source heat pump.

Minimum Performance. ASHRAE/IES Standard 90.1-1989 specifies a minimum performance for air-cooled DX systems in packaged units as covered in Section 9.7. For air-cooled, electrically operated rooftop heat pumps (air-source heat pumps), minimum performance characteristics are:

Cooling EER	8.9
Heating, COP (at T_oF = 47°F)	3.0
Heating, COP (at T_oF = 17°F)	2.0

Centrifugal Chillers

A chiller is a refrigeration machine using a liquid cooler as an evaporator to produce chilled water as the cooling medium in a central air-conditioning system. A *centrifugal chiller*, as shown in Figure 9.11.3(a), is a refrigeration machine using a centrifugal compressor to produce chilled water. It is often a factory-assembled unit with an integrated DDC control system and sometimes may separate into pieces for transportation. A centrifugal chiller is also a *centrifugal vapor compression refrigeration system.*

Refrigerants. According to Hummel et al. (1991), in 1988 there were about 73,000 centrifugal chillers in the United States. Of these, 80% use R-11, 10% use R-12, and the remaining 10% use R-22 and others. As mentioned in Section 9.4, production of CFCs, including R-11 and R-12, ceased at the end of 1995 with limited exception for service. R-123 (HCFC) will replace R-11. The chiller's efficiency may drop 2 to 4%, and a capacity reduction of 5% is possible. R-123 has low toxicity. Its allowable exposure limit was raised to 50 ppm in 1997 from 30 ppm in 1993 by its manufacturers. A monitor and alarm device to detect R-123 in air should be installed in plant rooms and places where there may be refrigerant leaks.

R-134a (HFC) will replace R-12. According to Lowe and Ares (1995), as a result of the changeout from R-12 to R-134a for a 5000-hp centrifugal chiller in Sears Tower, Chicago, its speed increased from 4878 to 5300 rpm, its cooling capacity is 12 to 24% less, and its efficiency is 12 to 16% worse.

System Components. A centrifugal chiller consists of a centrifugal compressor, an evaporator or liquid cooler, a condenser, a flash cooler, throttling devices, piping connections, and controls. A purge unit is optional.

- Centrifugal compressor — According to the number of internally connected impellers, the centrifugal compressor could have a single, two, or more than two stages. A two-stage impeller with a flash cooler is most widely used because of its higher system performance and comparatively simple construction. Centrifugal compressors having a refrigeration capacity less than 1200 tons are often hermetic. Very large centrifugal compressors are of open type. A gear train is often required to raise the speed of the impeller except for very large impellers using direct drive.
- Evaporator — Usually a liquid cooler of flooded shell-and-tube type evaporator is adopted because of its compact size and high rate of heat transfer.
- Condenser — Water-cooled, horizontal shell-and-tube condensers are widely used.

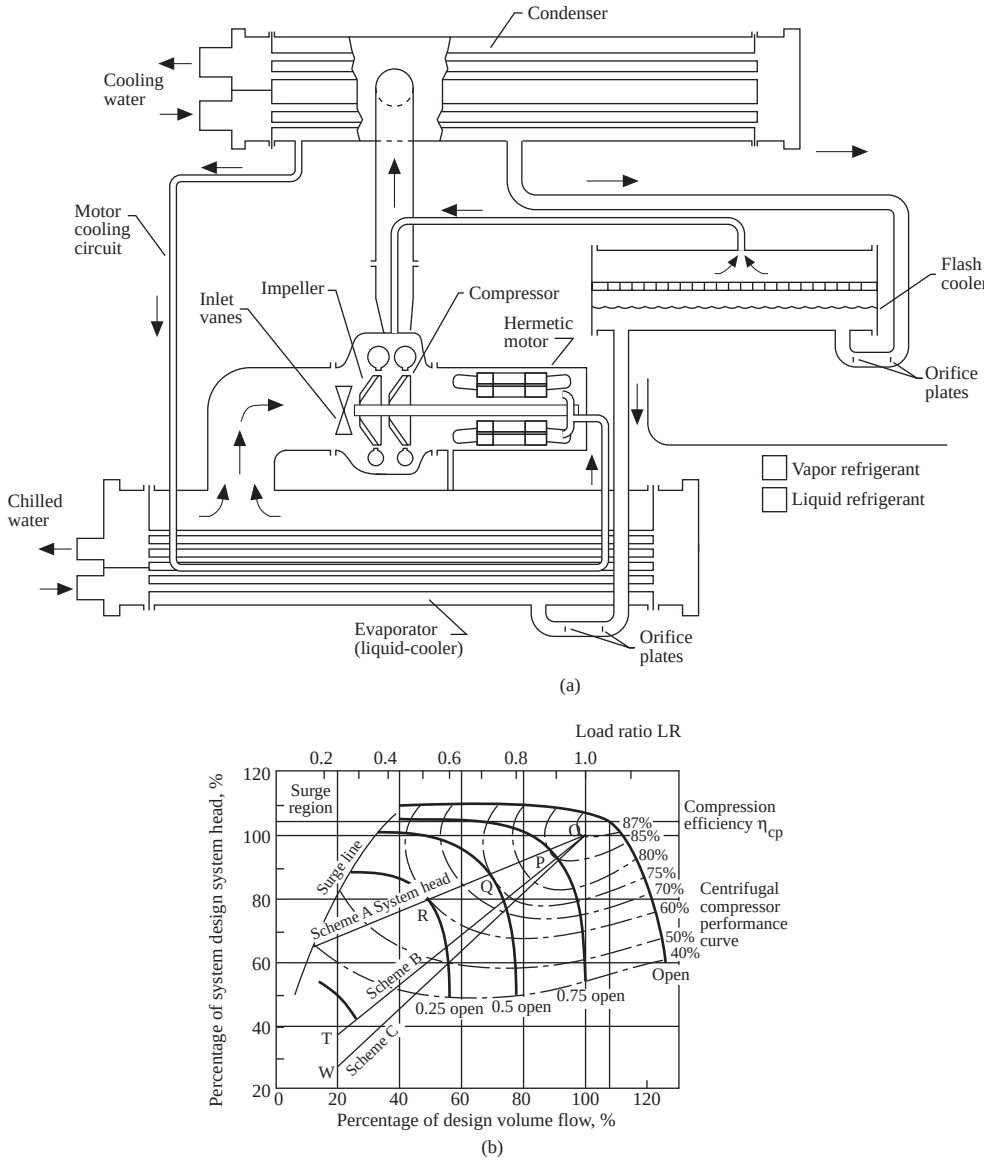


FIGURE 9.11.3 A two-stage water-cooled centrifugal chiller: (a) schematic diagram and (b) centrifugal compressor performance map at constant speed.

- Flash cooler — For a two-stage centrifugal compressor, a single-stage flash cooler is used. For a three-stage compressor, a two-stage flash cooler is used.
- Orifice plates and float valves — Both multiple-orifice plates such as that shown in Figure 9.11.3(a) and float valves are used as throttling devices in centrifugal chillers.
- Purge unit — R-123 has an evaporating pressure p_{ev} = 5.8 psia at 40°F, which is lower than atmospheric pressure. Air and other noncondensable gases may leak into the evaporator through cracks and gaps and usually accumulate in the upper part of the condenser. These noncondensable gases raise the condensing pressure, reduce the refrigerant flow, and lower the rate of heat transfer. A purge unit uses a cooling coil to separate the refrigerant and water from the noncondensable gases and purge the gases by using a vacuum pump.

Performance Ratings. The refrigeration cycle of a typical water-cooled, two-stage centrifugal chiller with a flash cooler was covered in Section 9.4. Centrifugal chillers have the same refrigeration capacity as centrifugal compressors, 100 to 10,000 tons. According to ARI Standard 550-88, the refrigeration capacity of a centrifugal chiller is rated as follows:

Chilled water temperature leaving evaporator T_{el}:	100% load 44°F
	0% load 44°F
Chilled water flow rate:	2.4 gpm/ton
Condenser water temperature entering condenser T_{ce}:	100% load 85°F
	0% load 60°F
Condenser water flow rate:	3.0 gpm/ton
Fouling factor in evaporator and condenser:	0.00025 hr.ft^2.°F/Btu

The integrated part-load value (IPLV) of a centrifugal chiller or other chillers at standard rating conditions can be calculated as:

$$\text{IPLV} = 0.1(A+B)/2 + 0.5(B+C)/2 + 0.3(C+D)/2 + 0.1D \tag{9.11.1}$$

where A, B, C, and D = kW/ton or COP at 100, 75, 50, and 25% load, respectively. If the operating conditions are different from the standard rating conditions, when T_{el} is 40 to 50°F, for each °F increase or decrease of T_{el}, there is roughly a 1.5% difference in refrigeration capacity and energy use; when T_{ce} is between 80 to 90°F, for each °F of increase or decrease of T_{ce}, there is roughly a 1% increase or decrease in refrigeration capacity and 0.6% in energy use.

ASHRAE/IES Standard 90.1-1989 and ARI Standard 550-88 specify the minimum performance for water-cooled water chillers from January 1, 1992:

	COP	**IPLV**
≥300 tons	5.2	5.3
≥150 tons < 300 tons	4.2	4.5
<150 tons	3.8	3.9

COP = 5.0 is equivalent to about 0.70 kW/ton. New, installed centrifugal chillers often have an energy consumption of 0.50 kW/ton.

Air-cooled centrifugal chillers have COPs from 2.5 to 2.8. Their energy performance is far poorer than that of water-cooled chillers. Their application is limited to locations where city water is not allowed to be used as makeup water for cooling towers.

Capacity Control. The refrigeration capacity of a centrifugal chiller is controlled by modulating the refrigerant flow at the centrifugal compressor. There are mainly two types of capacity controls: varying the opening and angle of the inlet vanes, and using an adjustable-frequency AC inverter to vary the rotating speed of the centrifugal compressor.

When the opening of the inlet vanes has been reduced, the refrigerant flow is throttled and imparted with a rotation. The result is a new performance curve at lower head and flow. If the rotating speed of a centrifugal compressor is reduced, it also has a new performance curve at lower volume flow and head. Inlet vanes are inexpensive, whereas the AC inverter speed modulation is more energy efficient at part-load operation.

Centrifugal Compressor Performance Map. Figure 9.11.3(b) shows a single-stage, water-cooled *centrifugal compressor performance map* for constant speed using inlet vanes for capacity modulation. A performance map consists of the compressor's performance curves at various operating conditions. The *performance curve* of a centrifugal compressor shows the relationship of volume flow of refrigerant $\dot{V}_r$

and its head lift Δp or compression efficiency η_{cp} at that volume flow. It is assumed that η_{cp} for a two-stage compressor is equal to the average of the two single-stage impellers having a head of their sum.

On the map, the required *system head curve* indicates the required system head lift at that volume flow of refrigerant. The intersection of the performance curve and the required system head curve is called the *operating point* O, P, Q, R, … as shown in Figure 9.11.3(b). One of the important operating characteristics of a centrifugal chiller (a centrifugal vapor compression refrigeration system as well) is that the required system head lift is mainly determined according to the difference in condensing and evaporating pressure $\Delta p_{c\text{-}e} = (p_{con} - p_{ev})$. The pressure losses due to the refrigerant piping, fittings, and components are minor.

In Figure 9.11.3(b), the abscissa is the percentage of design volume flow of refrigerant, % $\mathring{V}_r$, or load ratio; the ordinate is the percentage of design system head $\Delta H_{s.d}$, or percentage of design temperature lift $(T_{con} - T_{ev})$. Here load ratio LR is defined as the ratio of the refrigeration load to the design refrigeration load $q_{rl}/q_{rl.d}$. There are three schemes of required system head curves:

- Scheme A — T_{ce} = constant and T_{el} = constant
- Scheme B — T_{el} = constant and a drop of 2.5°F of T_{ce} for each 0.1 reduction of load ratio
- Scheme C — A reset of T_{el} of 1°F increase and a drop of 2.5°F of T_{ce} for each 0.1 reduction of load ratio

At design $\mathring{V}_r$ and system head $H_{s.d}$, $\eta_{cp} = 0.87$. As $\mathring{V}_r$, load ratio, and required system head Δp decrease, η_{cp} drops accordingly.

Surge is a unstable operation of a centrifugal compressor or fan resulting in vibration and noise. In a centrifugal chiller, surge occurs when the centrifugal compressor is unable to develop a discharge pressure that satisfies the requirement at the condenser. A centrifugal compressor should never be operated in the surge region.

Part-Load Operation. During part-load operation, if T_{el} and T_{ce} remain constant, the evaporating temperature T_{ev} tends to increase from the design load value because there is a greater evaporating surface and a smaller temperature difference $(T_{el} - T_{ev})$. Similarly, T_{con} tends to decrease.

The ratio of actual compressor power input at part load to the power input at design load may be slightly higher or lower than the load ratios, depending on whether the outdoor wet bulb is constant or varying at part load or whether there is a T_{el} reset; it also depends on the degree of drop of η_{cp} at part load.

Specific Controls. In addition to generic controls, specific controls for a centrifugal chiller include:

- Chilled water leaving temperature T_{el} and reset
- Condenser water temperature T_{ce} control
- On and off of multiple chillers based on actual measured coil load
- Air purge control
- Safety controls like oil pressure, low-temperature freezing protection, high condensing pressure control, motor overheating, and time delaying

Centrifugal Chillers Incorporating Heat Recovery. A *HVAC&R heat recovery system* converts waste heat or waste cooling from any HVAC&R process into useful heat and cooling. A heat recovery system is often subordinate to a parent system, such as a heat recovery system to a centrifugal chiller.

A centrifugal chiller incorporating a heat recovery system often uses a double-bundle condenser in which water tubes are classified as tower bundles and heating bundles. Heat rejected in the condenser may be either discharged to the atmosphere through the tower bundle and cooling tower or used for heating through the heating bundle. A temperature sensor is installed to sense the temperature of return hot water from the heating coils in the perimeter zone. A DDC controller is used to modulate a bypass three-way valve which determines the amount of condenser water supplied to the heating bundle. The tower and heating bundles may be enclosed in the same shell, but baffle sheets are required to guide the water flows.

A centrifugal chiller incorporating a heat recovery system provides cooling for the interior zone and heating for the perimeter zone simultaneously in winter with a higher COP_{hr}. However, it needs a higher condenser water-leaving temperature T_{cl} of 105 to 110°F, compared with 95°F or even lower in a cooling-only centrifugal chiller. An increase of 10 to 12°F of the difference ($T_{con} - T_{ev}$) requires an additional 10 to 15% power input to the compressor. For a refrigeration plant equipped with multiple chillers, it is more energy efficient and lower in first cost to have only part of them equipped with double-bundle condensers.

Screw Chillers

A *screw chiller* or a *helical rotary chiller* is a refrigeration machine using a screw compressor to produce chilled water. A factory-fabricated and assembled screw chiller itself is also a screw vapor compression refrigeration system.

Twin-screw chillers are more widely used than single-screw chillers. A twin-screw chiller consists of mainly a twin-screw compressor, a flooded shell-and-tube liquid cooler as evaporator, a water-cooled condenser, throttling devices, an oil separator, an oil cooler, piping, and controls as shown in Figure 9.11.4(a). The construction of twin-screw compressors has already been covered. For evaporator, condenser, and throttling devices, they are similar to those in centrifugal chillers. Most twin-screw chillers have a refrigeration capacity of 100 to 1000 tons.

Following are the systems characteristics of screw chillers.

Variable Volume Ratio. The ratio of vapor refrigerant trapped within the interlobe space during the intake process V_{in} to the volume of trapped hot gas discharged V_{dis} is called the *built-in volume ratio* of the twin-screw compressor $V_i = V_{in}/V_{dis}$, or simply *volume ratio*, all in ft^3.

There are two types of twin-screw chiller: fixed and variable volume ratio. For a twin-screw chiller of *fixed volume ratio*, the isentropic efficiency η_{isen} becomes maximum when the system required compression ratio $R_{s.com} \approx V_i$. Here $R_{s.com} = p_{con}/p_{ev}$. If $p_{dis} > p_{con}$, overcompression occurs, as shown in Figure 9.11.4(b). The discharged hot gas reexpands to match the condensing pressure. If $p_{dis} < p_{con}$, undercompression occurs (Figure 9.11.4[c]). A volume of gas at condensing pressure reenters the trapped volume at the beginning of the discharge process. Both over- and undercompression cause a reduction of η_{isen}.

For a twin-screw chiller of *variable volume ratio*, there are two slides: a sliding valve is used for capacity control and a second slide. By moving the second slide back and forth, the radial discharge port can be relocated. This allows variation of suction and discharge pressure levels and still maintains maximum efficiency.

Economizer. The hermetic motor shell is connected to an intermediate point of the compression process and maintains an intermediate pressure p_i between p_{con} and p_{ev}. Liquid refrigerant at condensing pressure p_{con} is throttled to p_i, and a portion of the liquid is flashed into vapor. This causes a drop in the temperature of the remaining liquid refrigerant down to the saturated temperature corresponding to p_i. Although the compression in a twin-screw compressor is in continuous progression, the mixing of flashed gas with the compressed gas at the intermediate point actually divides the compression process into two stages. The resulting economizing effect is similar to that of a two-stage compound refrigeration system with a flash cooler: an increase of the refrigeration effect and a saving of the compression power from ($p_{con} - p_{ev}$) to ($p_{con} - p_i$).

Oil Separation, Oil Cooling, and Oil Injection. Oil entrained in the discharged hot gas enters an oil separator. In the separator, oil impinges on an internal perforated surface and is collected because of its inertia. Oil drops to an oil sump through perforation. It is then cooled by condenser water in a heat exchanger. A heater is often used to vaporize the liquid refrigerant in the oil sump to prevent dilution of the oil. Since the oil sump is on the high-pressure side of the refrigeration system, oil is forced to the rotor bearings and injected to the rotors for lubrication.

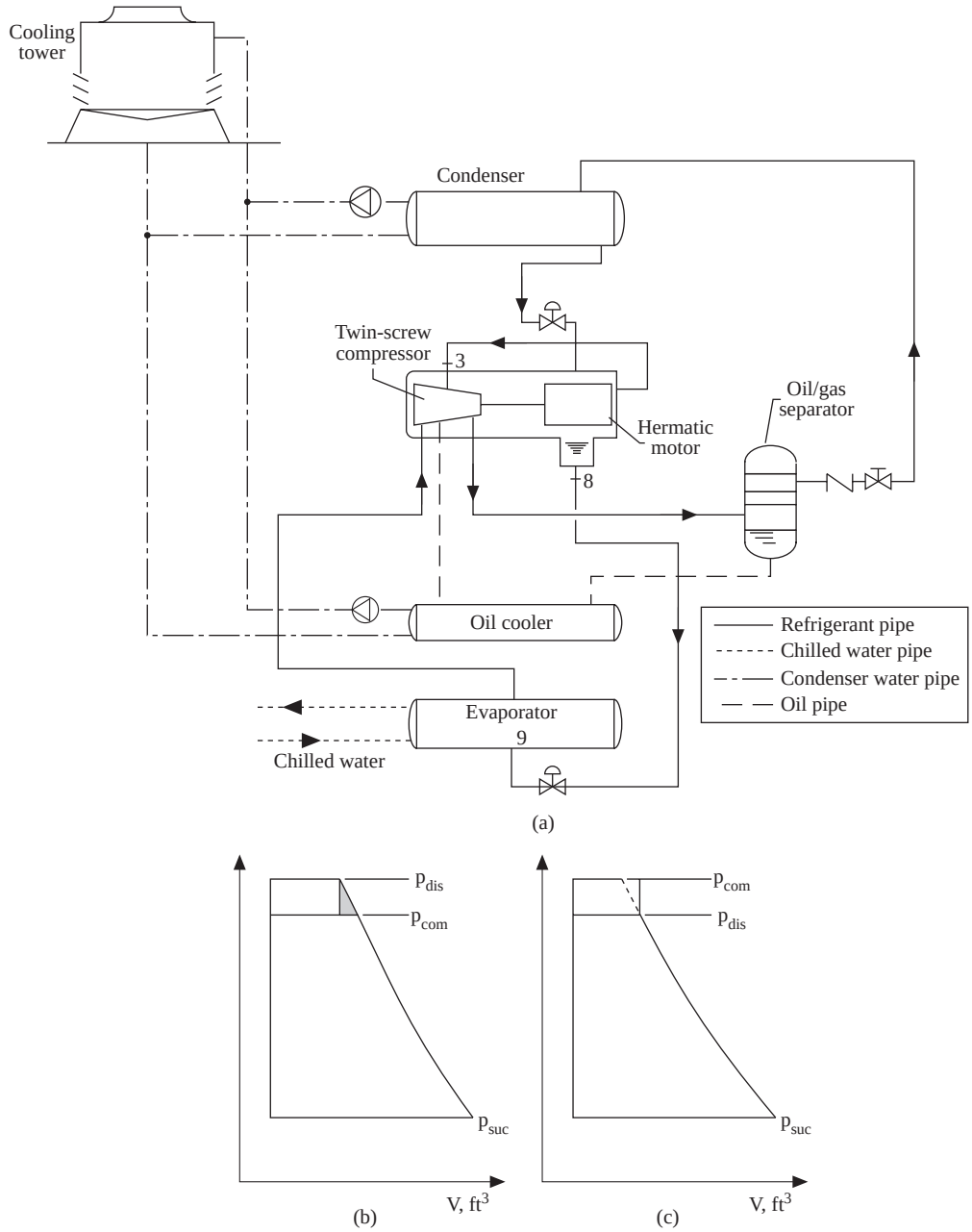


FIGURE 9.11.4 A typical twin-screw chiller: (a) schematic diagram, (b) over-compression, and (c) under-compression.

Oil slugging is not a problem for twin-screw compressors. When suction vapor contains a small amount of liquid refrigerant that carries over from the oil separator, often called *wet suction*, it often has the benefit of scavenging the oil from the evaporator.

Twin-screw compressors are positive displacement compressors. They are critical in oil lubrication, sealing, and cooling. They are also more energy efficient than reciprocating compressors. Twin-screw chillers are gaining more applications, especially for ice-storage systems with cold air distribution.

9.12 Thermal Storage Systems

Thermal Storage Systems and Off-Peak Air-Conditioning Systems

Many electric utilities in the United States have their *on-peak hours* between noon and 8 p.m. during summer weekdays, which include the peak-load hours of air-conditioning. Because the capital cost of a new power plant is so high, from $1200 to $4000 per kW, electric utilities tend to increase their power output by using customers' thermal energy storage (TES) systems, or simply thermal storage systems, which are much less expensive.

A *thermal storage system* as shown in Figure 9.12.1(a) may have the same refrigeration equipment, like chillers, additional storage tank(s), additional piping, pumps, and controls. The electric-driven compressors are operated during off-peak, partial-peak, and on-peak hours. *Off-peak hours* are often nighttime hours. *Partial-peak hours* are hours between on-peak and off-peak hours in a weekday's 24-hr day-and-night cycle. Chilled water or ice is stored in tanks to provide cooling for buildings during on-peak hours when higher electric demand and electric rates are effective. Although thermal storage systems operate during nighttime when outdoor dry and wet bulbs are lower, they are not necessarily energy saving due to lower evaporating temperature, additional pump power, and energy losses. Thermal storage systems significantly reduce the electric energy cost.

Utilities in the United States often use higher electric demand and rates as well as incentive bonus to encourage the shift of electric load from on-peak to off-peak hours by using thermal storage systems and others. Such a shift not only saves expensive capital cost, but also increases the use of base-load high-efficiency coal and nuclear plants instead of inefficient diesel and gas turbine plants.

The air-conditioning systems that operate during off-peak and partial-peak hours for thermal storage, or those that use mainly natural gas to provide cooling to avoid higher electric demand and rates during on-peak hours, are called *off-peak air-conditioning systems*. These systems include ice-storage and chilled-water storage systems, desiccant cooling systems, absorption systems, and gas engine chillers.

Absorption chillers and desiccant cooling systems are covered in other sections. Gas engine-driven reciprocating chillers are often a cogeneration plant with heat recovery from engine jacket and exhaust gas, and will not be covered here.

Full Storage and Partial Storage

Ice and chilled-water storage systems are the two most common thermal energy storage systems today. Knebel (1995) estimated that more than 4000 cool storage systems are operated in various commercial buildings.

The unit of stored thermal energy for cooling is ton-hour, or ton.hr. One *ton.hr* is the refrigeration capacity of one refrigeration ton during a 1-hr period, or 12,000 Btu.

In order to achieve minimum life-cycle cost, thermal storage systems could be either full storage or partial storage. For a *full-storage*, or *load shift*, thermal storage system, all refrigeration compressors cease to operate during on-peak hours. Building refrigeration loads are entirely offset by the chilled water or brine from the thermal storage system within an on-peak period. In a *partial storage*, or *load-leveling*, thermal storage system as shown in Figure 9.12.1(b) all or some refrigeration compressor(s) are operated during on-peak hours.

Direct cooling is the process in which refrigeration compressors produce refrigeration to cool the building directly. During a specific time interval, if the cost of direct cooling is lower than the stored energy, the operation of a thermal storage system is said to be in *chiller priority.* On the contrary, if the cost of direct cooling is higher than the cost of stored energy, the operation is said to be at *storage priority.*

The optimum size of a thermal storage system is mainly determined according to the utility's electric rate structure, especially a time-of-day structure whose electric rates are different between on-peak, partial-peak, and off-peak hours. Not only the design day's instantaneous building peak cooling load is important, but also an hour-by-hour cooling load profile of the design day is required for thermal storage design. A simple payback or a life-cycle cost analysis is usually necessary.

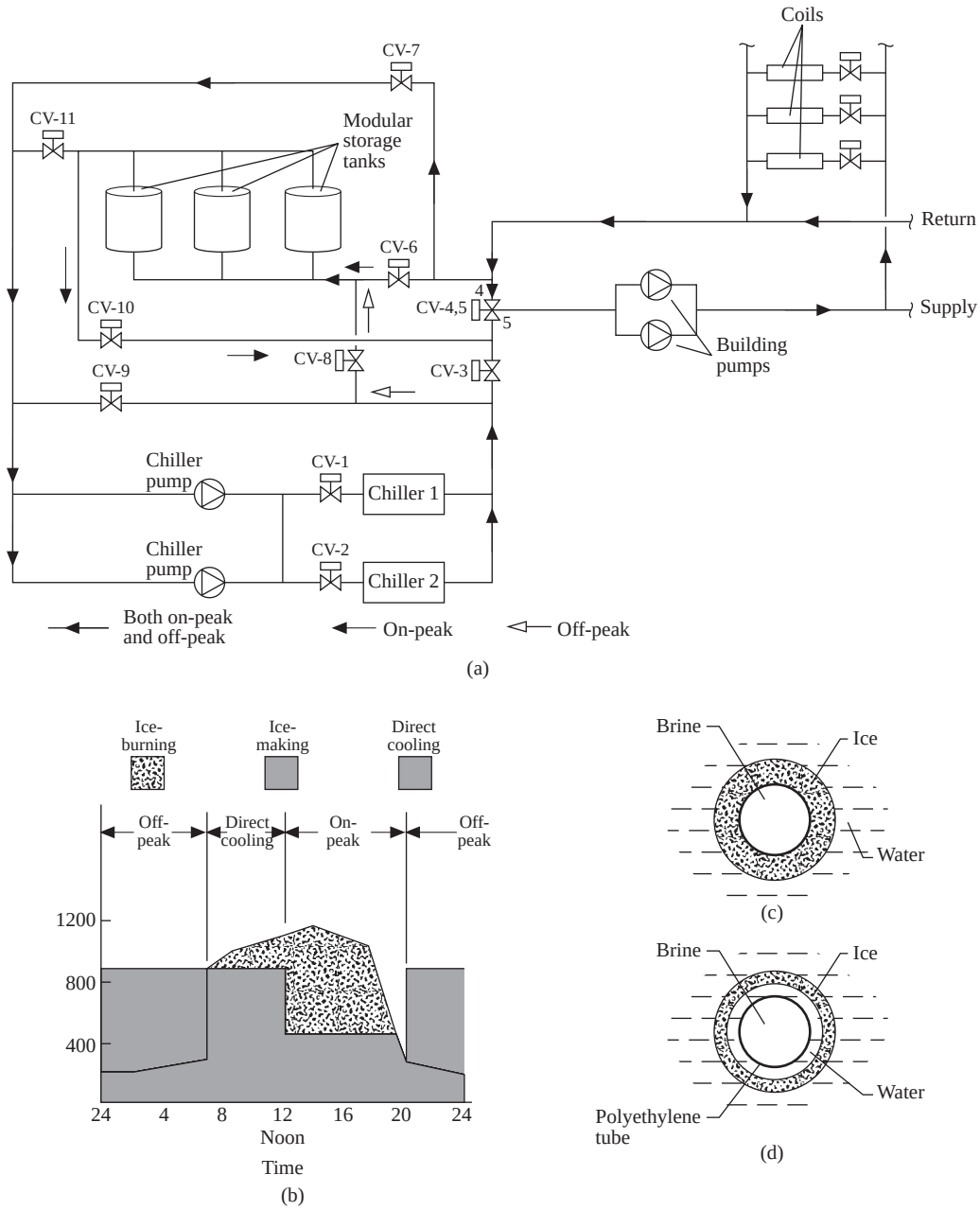


FIGURE 9.12.1 A brine-coil ice-storage system: (a) schematic diagram, (b) partial-storage time schedule, (c) ice making, and (d) ice burning.

Ice-Storage Systems

System Characteristics

In an *ice-thermal-storage* system, or simply an *ice-storage* system, ice is stored in a tank or other containers to provide cooling for buildings in on-peak hours or on- and partial-peak hours. One pound of ice can store [(1 × 144) + (55 – 35)] = 164 Btu instead of (60 – 44) = 16 Btu for chilled water storage. For the same cooling capacity, the storage volume of ice is only about 12% of chilled water. In addition,

an air-conditioning system using ice storage often adopts cold air distribution to supply conditioned air at a temperature typically at 44°F. Cold air distribution reduces the size of air-side equipment, ducts, and investment as well as fan energy use. It also improves the indoor air quality of the conditioned space. Since the late 1980s, ice storage has gained more applications than chilled water storage.

Brine is a coolant without freezing and flashing during normal operation. The freezing point of brine, which has a mass fraction of ethylene glycol of 25%, drops to 10°F, and a mass fraction of propylene glycol of 25% drops to 15°F. Glycol-water, when glycol is dissolved in water, is another coolant widely used in ice-storage systems. Ice crystals are formed in glycol-water when its temperature drops below its freezing point during normal operation.

In an ice-storage system, *ice making* or *charging* is a process in which compressors are operated to produce ice. *Ice burning*, or *discharging*, is a process in which ice is melted to cool the brine or glycol-water to offset refrigeration load.

Brine-Coil Ice-Storage Systems. Currently used ice-storage systems include brine-coil, ice-harvester, and ice-on-coil systems. According to Knebel (1995), the brine-coil ice-storage system is most widely used today because of its simplicity, flexibility, and reliability as well as using modular ice-storage tanks.

In a typical brine-coil ice-storage system, ice is charged in multiple modular factory-fabricated storage tanks as shown in Figure 9.12.1(a). In each storage tank, closely spaced polyethylene or plastic tubes are surrounded by water. Brine, a mixture of 25 to 30% of ethylene glycol by mass and 70 to 75% water, circulates inside the tubes at an entering temperature of 24°F during the ice-making process. The water surrounding the tubes freezes into ice up to a thickness of about 1/2 in. as shown in Figure 9.12.1(c). Brine leaves the storage tank at about 30°F. Water usually at atmospheric pressure is separated from brine by a plastic tube wall. Plastic tubes occupy about one tenth of the tank volume, and another one tenth remains empty for the expansion of ice. Multiple modular storage tanks are always connected in parallel.

During the ice-melting or -burning process, brine returns from the cooling coils in the air-handling units (AHUs) at a temperature of 46°F or higher. It melts the ice on the outer surface of the tubes and is thus cooled to a temperature of 34 to 36°F, as shown in Figure 9.12.1(d). Brine is then pumped to the AHUs to cool and dehumidify the air again.

Case Study of a Brine-Coil Ice-Storage System. In Tackett (1989), a brine-coil ice-storage system cools a 550,000-ft^2 office building near Dallas, TX, as shown in Figure 9.12.1(a). Ethylene glycol water is used as the brine. There are two centrifugal chillers, each of them having a refrigeration capacity of 568 tons when 34°F brine is produced with a power consumption of 0.77 kW/ton for direct cooling. The refrigeration capacity drops to 425 tons if 24°F brine leaves the chiller with a power consumption of 0.85 kW/ton. A demand-limited partial storage is used, as shown in Figure 9.12.1(b). During on-peak hours, ice is melted; at the same time one chiller is also operating. The system uses 90 brine-coil modular storage tanks with a full-charged ice-storaged capacity of 7500 ton.hr.

For summer cooling, the weekdays' 24-hr day-and-night cycle is divided into three periods:

- Off peak lasts from 8 p.m. to AHU's start the next morning. Ice is charged at a maximum capacity of 650 tons. The chillers also provide 200 tons of direct cooling for refrigeration loads that operate 24 hr continuously.
- Direct cooling lasts from AHU's start until noon on weekdays. Chillers are operated for direct cooling. If required refrigeration load exceeds the chillers' capacity, some ice storage will be melted to supplement the direct cooling.
- On peak lasts from noon to 8 p.m. Ice is burning with one chiller in operation. During ice burning, water separates the tube and ice. Water has a much lower thermal conductivity (0.35 Btu/hr.ft.°F) than ice (1.3 Btu/hr.ft.°F). Therefore, the capacity of a brine-coil ice-storage system is dominated by the ice burning.

Ice-Harvester Ice-Storage Systems. In an *ice-harvester* system, glycol-water flows on the outside surface of the evaporator and forms ice sheets with a thickness of 0.25 to 0.40 in. within 20 to 30 min. Ice is harvested in the form of flakes when hot gas is flowing inside the tubes of the evaporator during a time interval of 20 to 30 sec. Ice flakes fall to the glycol-water in the storage tank below. The glycol-water at a temperature of 34°F is then supplied to the cooling coils in AHUs for conditioning. After cooling and dehumidifying the air, glycol-water returns to the storage tank at a temperature of 50 to 60°F and is again cooled to 34°F again.

Ice harvesting is an intermittent process. It has a cycle loss due to harvesting of about 15%. In addition, because of its operating complexity and maintenance, its applications are more suitable for large ice-storage systems.

Ice-on-Coil Ice-Storage Systems. In an ice-on-coil system, refrigerant is evaporated in the coils submerged in water in a storage tank. Ice of a thickness not exceeding 0.25 in. builds up on the outer surface of the coils. The remaining water in the storage tank is used for cooling in AHUs. Ice-on-coil systems need large amounts of refrigerant charge and are less flexible in operation. They are usually used in industrial applications.

Ice-in-Containers Ice-Storage Systems. Ice-in-containers ice-storage systems store ice in enclosed containers. Brine circulating over the containers produces the ice inside containers. Complexity in control of the ice inventory inside the containers limits the application of the ice-in-containers systems.

Chilled-Water Storage Systems

Basics

Chilled-water storage uses the same water chiller and a similar coolant distribution system, except for additional water storage tanks and corresponding piping, additional storage pumps, and controls. The larger the chilled-water storage system, the lower the installation cost per ton.hr storage capacity.

Various types of storage tanks had been used in chilled-water storage systems during the 1970s. A diaphragm tank uses a rubber diaphragm to separate the colder and warmer water. Baffles divide the tank into cells and compartments. Today, stratified tanks have become the most widely used chilled-water storage systems because of their simplicity, low cost, and negligible difference in loss of cooling capacity between stratified tanks and other types of storage tanks.

During the storage of chilled water, the loss in cooling capacity includes direct mixing, heat transfer between warmer return chilled water and colder stored chilled water, and also heat transfer between warmer ambient air and colder water inside the tank. An enthalpy-based easily measured index called *figure of merit* (*FOM*) is often used to indicate the loss in cooling capacity during chilled-water storage. FOM is defined as:

$$\text{FOM} = q_{\text{dis}} / q_{\text{ch}} \tag{9.12.1}$$

where q_{dis} = cooling capacity available in the discharge process, Btu/hr
q_{ch} = theoretical cooling capacity available during charging process, Btu/hr

Charging is the process of filling the storage tank with colder chilled water from the chiller. At the same time, warmer return chilled water is extracted from the storage tank and pumped to the chiller for cooling.

Discharging is the process of discharging the colder stored chilled water from the storage tank to AHUs and terminals. Meanwhile, the warmer return chilled water from the AHUs and terminals fills the tank.

Stratified Tanks. *Stratified tanks* utilize the buoyancy of warmer return chilled water to separate it from the colder stored chilled water during charging and discharging, as shown in Figure 9.12.2(a). Colder stored chilled water is always charged and discharged from bottom diffusers, and the warmer return chilled water is introduced to and withdrawn from the top diffusers.

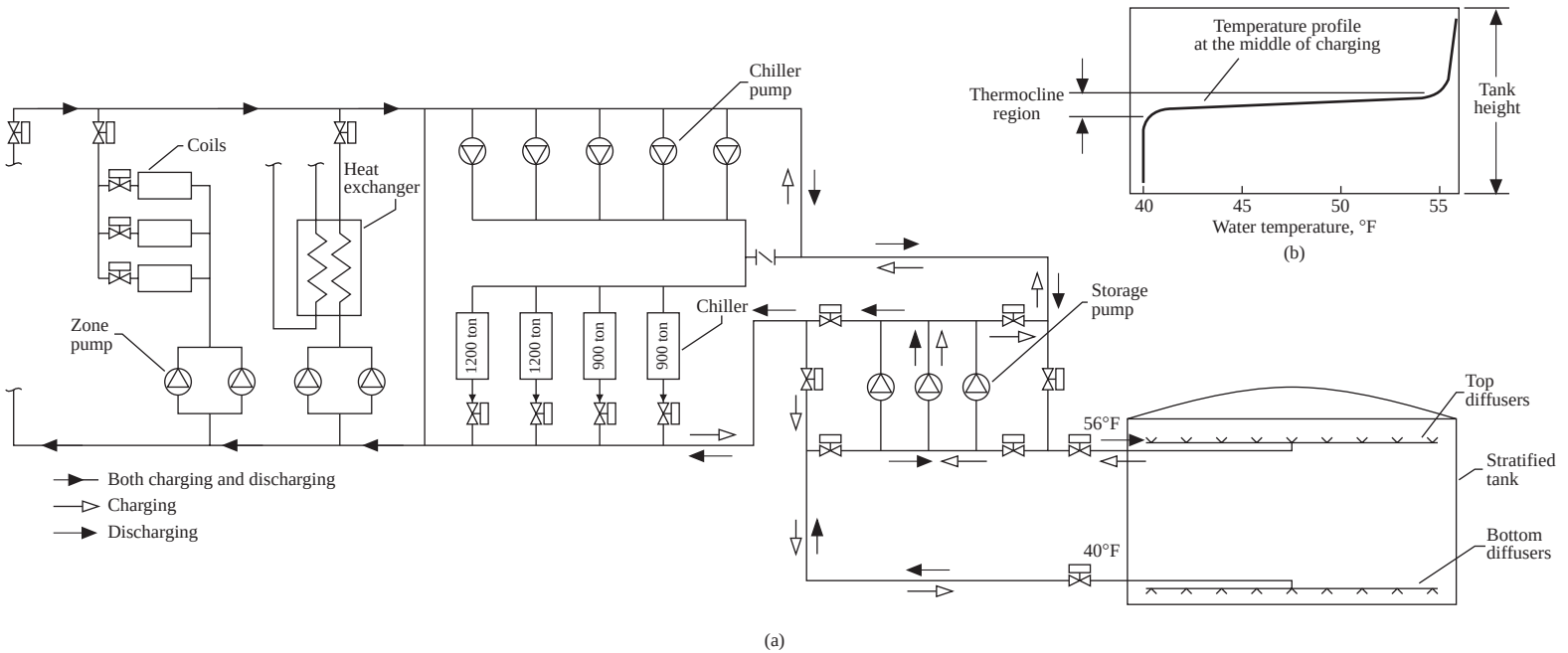


FIGURE 9.12.2 A chilled-water storage system using stratified tanks: (a) schematic diagram of a chilled-water storage system and (b) thermocline at the middle of charging process.

Chilled-water storage tanks are usually vertical cylinders and often have a height-to-diameter ratio of 0.25:0.35. Steel is the most commonly used material for above-grade tanks, with a 2-in.-thick spray-on polyurethane foam, a vapor barrier, and a highly reflective coating. Concrete, sometimes precast, prestressed tanks are widely used for underground tanks.

A key factor to reduce loss in cooling capacity during chilled water storage is to reduce mixing of colder and warmer water streams at the inlet. If the velocity pressure of the inlet stream is less than the buoyancy pressure, the entering colder stored chilled water at the bottom of tank will stratify. Refer to Wang's handbook (1993) and Knebel (1995) for details.

A *thermocline* is a stratified region in a chilled-water storage tank of which there is a steep temperature gradient as shown in Figure 9.12.2(b). Water temperature often varies from 42°F to about 60°F. Thermocline separates the bottom colder stored chilled water from the top warmer return chilled water. The thinner the thermocline, the lower the mixing loss.

Diffusers and symmetrical connected piping are used to evenly distribute the incoming water streams with sufficient low velocity, usually lower than 0.9 ft/sec. Inlet stream from bottom diffusers should be downward and from the top diffusers should be upward or horizontal.

Field measurements indicate that stratified tanks have a FOM between 0.85 to 0.9.

9.13 Air System Basics

Fan-Duct Systems

Flow Resistance

Flow resistance is a property of fluid flow which measures the characteristics of a flow passage resisting the fluid flow with a corresponding total pressure loss Δp, in in. WG, at a specific volume flow rate $\mathring{V}$, in cfm:

$$\Delta p = R\mathring{V}^2 \tag{9.13.1}$$

where R = flow resistance (in. WG/(cfm)^2).

For a duct system that consists of several duct sections connected in series, its flow resistance R_s, in in. WG/(cfm)^2, can be calculated as

$$R_s = R_1 + R_2 + \ldots + R_n \tag{9.13.2}$$

where R_1, R_2, … R_n = flow resistance of duct section 1, 2, … n in the duct system (in. WG/(cfm)^2).

For a duct system that consists of several duct sections connected in parallel, its flow resistance R_p, in in. WG/(cfm)^2, is:

$$1/\sqrt{R_p} = 1/\sqrt{R_1} + 1/\sqrt{R_2} + \ldots + 1/\sqrt{R_n} \tag{9.13.3}$$

Fan-Duct System

In a *fan-duct system*, a fan or several fans are connected to ductwork or ductwork and equipment. The volume flow and pressure loss characteristics of a duct system can be described by its performance curve, called *system curve*, and is described by $\Delta p = R\,\mathring{V}^2$.

An *air system* or an air handling system is a kind of fan-duct system. In addition, an outdoor ventilation air system to supply outdoor ventilation air, an exhaust system to exhaust contaminated air, and a smoke control system to provide fire protection are all air systems, that is, fan-duct systems.

Primary, Secondary, and Transfer Air

Primary air is the conditioned air or makeup air. Secondary air is often the induced space air, plenum air, or recirculating air. Transfer air is the indoor air that moves to a conditioned space from an adjacent area.

System-Operating Point

A *system-operating point* indicates the operating condition of an air system or fan-duct system. Since the operating point of a fan must lie on the fan performance curve, and the operating point of a duct system on the system curve, the system operating point of an air system must be the intersection point P_s of the fan performance curve and system curve as shown in Figure 9.13.1(a).

Fan-Laws

For the same air system operated at speeds n_1 and n_2, both in rpm, their relationship of $\mathring{V}$ volume flow rate, in cfm, system total pressure loss, in in. WG, and fan power input, in hp, can be expressed as

$$\begin{aligned} \mathring{V}_2/\mathring{V}_1 &= n_2/n_1 \\ \Delta p_{t2}/\Delta p_{t1} &= (n_2/n_1)^2(\rho_2/\rho_1) \\ P_2/P_1 &= (n_2/n_1)^3(\rho_2/\rho_1) \end{aligned} \tag{9.13.4}$$

where ρ = air density (lb/ft^3). Subscripts 1 and 2 indicate the original and the changed operating conditions. For air systems that are geometrically and dynamically similar:

$$\begin{aligned} \mathring{V}_2 / \mathring{V}_1 &= \left(D_2/D_1\right)^3 \left(n_2/n_1\right) \\ \Delta p_{t2}/\Delta p_{t1} &= \left(D_2/D_1\right)^2 \left(n_2/n_1\right)^2 \left(\rho_2/\rho_1\right) \\ P_2/P_1 &= \left(D_2/D_1\right)^5 \left(n_2/n_1\right)^3 \left(\rho_2/\rho_1\right) \end{aligned} \tag{9.13.5}$$

where D = diameter of the impeller (ft).

Geometrically similar means that two systems are similar in shape and construction. For two systems that are dynamically similar, they must be geometrically similar, and in addition, their velocity distribution or profile of fluid flow should also be similar. When fluid flows in the air systems are at high Reynolds number, such as Re > 10,000, their velocity profiles can be considered similar to each other.

System Effect

The system effect Δp_{se}, in in. WG, of an air system is its additional total pressure loss caused by uneven or nonuniform velocity profile at the fan inlet, or at duct fittings after fan outlet, due to the actual inlet and outlet connections as compared with the total pressure loss of the fan test unit during laboratory ratings. The selected fan total pressure which includes the system effect Δp_{ts}, in in. WG, as shown in Figure 9.13.1(a), can be calculated as

$$\begin{aligned} \Delta p_{ts} &= \Delta p_{sy} + \Delta p_{se} = \Delta p_{sy} + \Delta p_{s.i} + \Delta p_{s.o} \\ &= \Delta p_{sy} + C_{s.i}\left(v_{fi}/4005\right)^2 + C_{s.o}\left(v_{fo}/4005\right)^2 \end{aligned} \tag{9.13.6}$$

where Δp_{sy} = calculated total pressure loss of the air system, in WG
$\Delta p_{s.i}$, $\Delta p_{s.o}$ = fan inlet and outlet system effect loss, in WG
$C_{s.i}$, $C_{s.o}$ = local loss coefficient of inlet and outlet system effect, in WG
v_{fi}, v_{fo} = velocity at fan inlet (fan collar) and fan outlet, fpm

Both $C_{s.i}$ and $C_{s.o}$ are mainly affected by the length of connected duct between the duct fitting and fan inlet or outlet, by the configuration of the duct fitting, and by the air velocity at the inlet or outlet. Because v_{fi} and v_{fo} are often the maximum velocity of the air system, system effect should not be overlooked. According to AMCA Fan and Systems (1973), a square elbow (height to turning radius ratio R/H = 0.75) connected to a fan inlet with a connected duct length of 2 D_e (equivalent diameter) and v_{fi} = 3000 fpm may have a 0.67 in. WG $\Delta p_{i.o}$ loss. Refer to AMCA Fan and Systems (1973) or Wang's handbook (1993) for details.

Modulation of Air Systems

Air systems can be classified into two categories according to their operating volume flow: constant volume and variable-air-volume systems. The volume flow rate of a *constant volume system* remains constant during all the operating time. Its supply temperature is raised during part load. For a *variable-air-volume* (*VAV*) *system*, its volume flow rate is reduced to match the reduction of space load at part-load operation. The system pressure loss of a VAV system can be divided into two parts: variable part Δp_{var} and fixed part Δp_{fix}, which is the set point of the duct static pressure control as shown in Figure 9.13.1(b) and (c). The *modulation curve* of a VAV air system its its operating curve, or the locus of system operating points when its volume flow rate is modulated at full- and part-load operation.

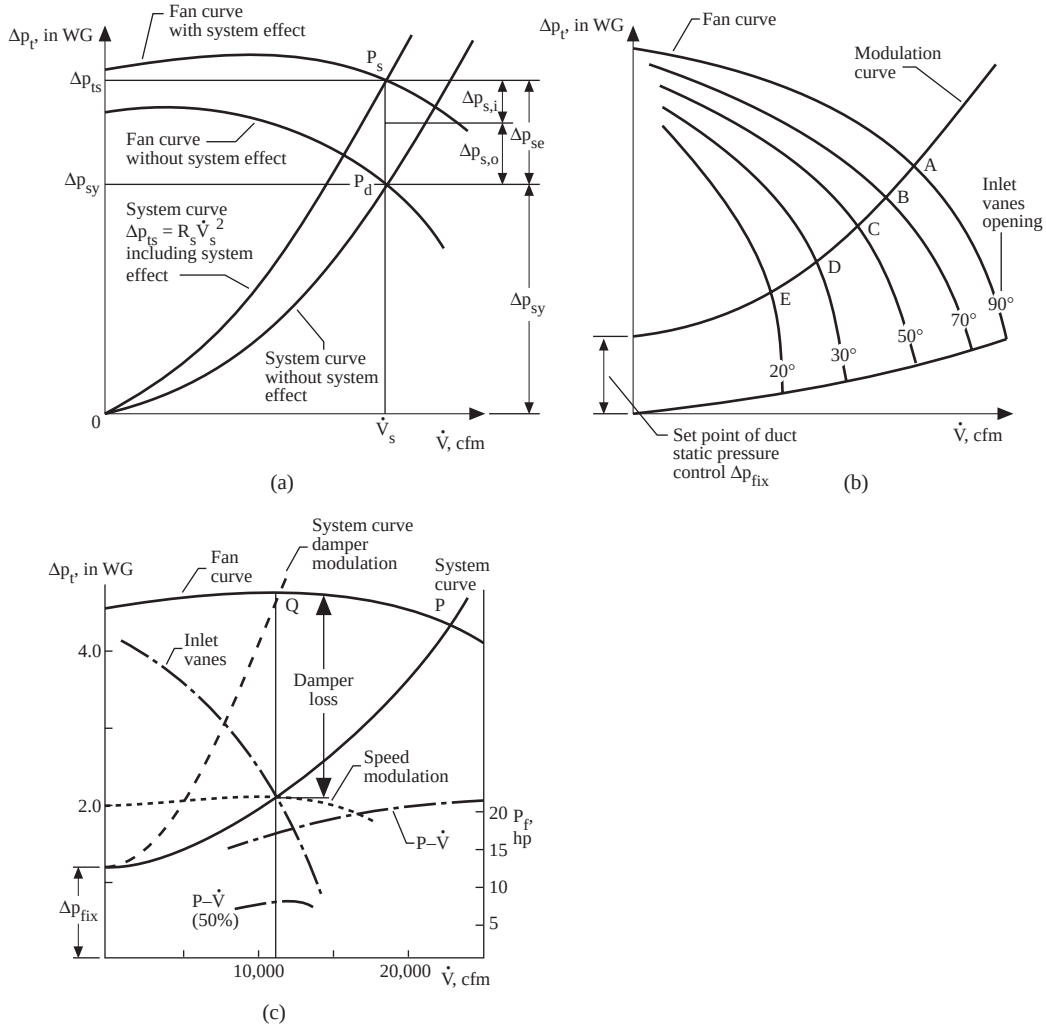


FIGURE 9.13.1 Air system $\dot{V}$-Δp_t performance: (a) system operating point and system effect, (b) modulation curve, and (c) damper, inlet vanes, and fan speed modulation.

The volume flow and system pressure loss of an air system can be modulated either by changing its fan characteristics or by varying its flow resistance of the duct system. Currently used types of modulation of volume flow rate of VAV air systems are

1. *Damper modulation* uses an air damper to vary the opening of the air flow passage and therefore its flow resistance.
2. *Inlet vanes modulation* varies the opening and the angle of inlet vanes at the centrifugal fan inlet and then gives different fan performance curves.
3. *Inlet cone modulation* varies the peripheral area of the fan impeller and therefore its performance curve.
4. *Blade pitch modulation* varies the blade angle of the axial fan and its performance curve.
5. *Fan speed modulation* using *adjustable frequency AC drives* varies the fan speed by supplying a variable-frequency and variable-voltage power source. There are three types of AC drives: adjustable voltage, adjustable current, and pulse width modulation (PWM). The PWM is universally applicable.

Damper modulation wastes energy. Inlet vanes are low in cost and are not so energy efficient compared with AC drives and inlet cones. Inlet cone is not expensive and is suitable for backward curved centrifugal fans. Blade pitch modulation is energy efficient and is mainly used for vane and tubular axial fans. AC drive is the most energy-efficient type of modulation; however, it is expensive and often considered cost effective for air systems using large centrifugal fans.

Example 9.13.1

A multizone VAV system equipped with a centrifugal fan has the following characteristics:

$\mathring{V}$ (cfm)	5,000	10,000	15,000	20,000	25,000
Δp_t, in. WG	4.75	4.85	4.83	4.60	4.20
P, hp		17.0	18.6	20.5	21.2

At design condition, it has a volume flow rate of 22,500 cfm and a fan total pressure of 4.45 in. WG. The set point of duct static pressure control is 1.20 in. WG.

When this VAV system is modulated by inlet vanes to 50% of design flow, its fan performance curves show the following characteristics:

$\mathring{V}$ (cfm)	5,000	10,000	11,250
Δp_t, in. WG	3.6	2.5	2.1
P, hp		7.5	7.8

Determine the fan power input when damper, inlet vanes, or AC drive fan speed modulation is used. Assume that the fan total efficiency remains the same at design condition when the fan speed is reduced.

Solutions

1. At 50% design flow, the volume flow of this VAV system is $0.5 \times 22{,}500 = 11{,}250$ cfm. The flow resistance of the variable part of this VAV system is

$$R_{var} = \Delta p_{va} = R\mathring{V}^2 = (4.45 - 1.20)/(22{,}500)^2 = 6.42 \times 10^{-9} \text{ in. WG}/(\text{cfm})^2$$

   When damper modulation is used, the system operating point Q must be the intersection of the fan curve and the system curve that has a higher flow resistance and a $\mathring{V}$ = 11,250 cfm. From Figure 9.13.1(c), at point Q, the fan power input is 17.0 hp.
2. From the given information, when inlet vane modulation is used, the fan power input is 7.8 hp.
3. The total pressure loss of the variable part of the VAV system at 50% volume flow is

$$\Delta p_{var} = R_{var}\mathring{V}^2 = 6.42 \times 10^{-9}(11{,}250)^2 = 0.81 \text{ in. WG}$$

   From Figure 9.13.1(c), since the fan power input at design condition is 21.2 hp, then its fan total efficiency is:

$$\eta_f = \mathring{V}\Delta p_{tf}/(6356P_f) = 22{,}500 \times 4.45/(6356 \times 21.2) = 74.3\%$$

   The fan power input at 50% design volume flow is:

$$P = \mathring{V}\Delta p_{tf}/(6356\eta_f) = 11{,}250(0.81 + 1.20)/(6356 \times 0.743) = 4.8 \text{ hp}$$

   Damper modulation has a energy waste of (17 – 4.8) = 12.2 hp

Fan Combinations in Air-Handling Units and Packaged Units

Currently used fan combinations in air-handling units (AHUs) and packaged units (PUs) (except dual-duct VAV systems) are shown in Figure 9.13.2(a), (b), and (c):

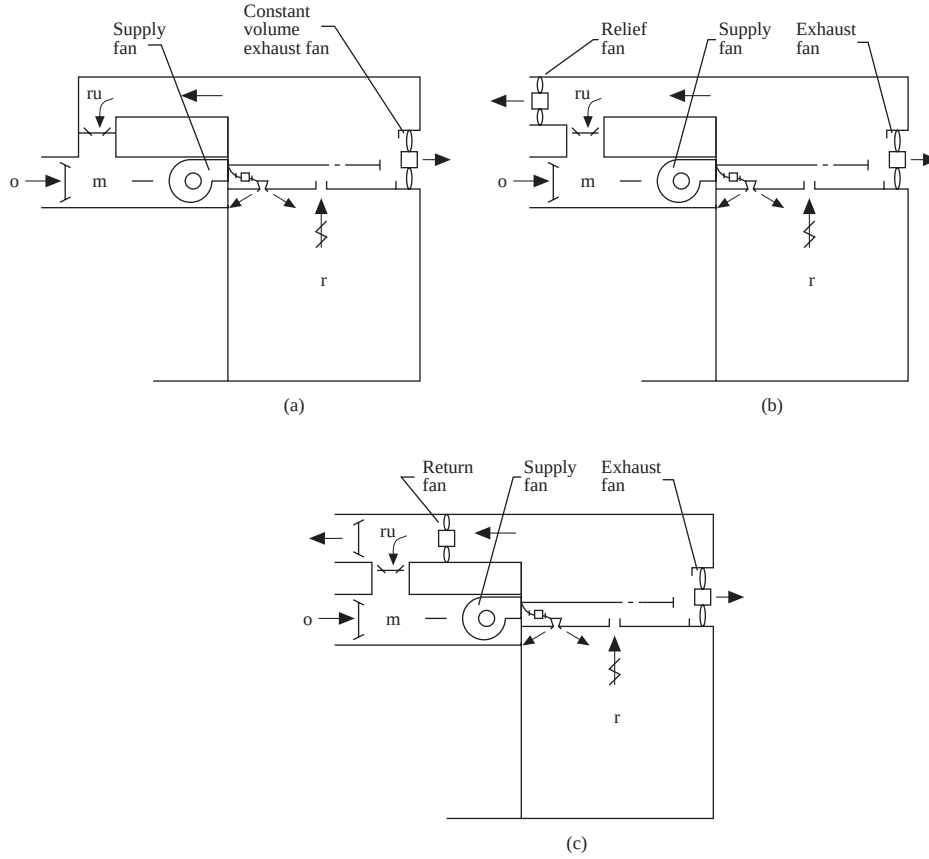


FIGURE 9.13.2 Fan combinations: (a) supply and exhaust fan, (b) supply and relief fan, and (c) supply and return fan.

Supply and Exhaust Fan/Barometric Damper Combination

An air system equipped with a single supply fan and a constant-volume exhaust fan, or a supply fan and a barometric damper combination as shown in Figure 9.13.2(a), is often used in buildings where there is no return duct or the pressure loss of the return duct is low. An all-outdoor air economizer cycle is usually not adopted due to the extremely high space pressure. A barometric relief damper is often installed in or connected to the conditioned space to prevent excessively high space pressure. When the space-positive pressure exerted on the barometric damper is greater than the weight of its damper and/or a spring, the damper is opened and the excessive space pressure is relieved.

Consider a VAV rooftop packaged system using a supply fan and a constant-volume exhaust system to serve a typical floor in an office building. This air system has the following design parameters:

During minimum outdoor ventilation air *recirculating mode* at summer design volume flow and at 50% of design volume flow, the outdoor damper is at its minimum opening. The recirculating damper is fully opened. The outdoor air intake at the PU must be approximately equal to the exfiltration at the conditioned space due to the positive space pressure p_r, in in. WG. By using the iteration method, the calculated pressure characteristics and the corresponding volume flow rates are shown below:

Supply volume flow rate:		20,000 cfm
Total pressure loss:	Across the recirculating damper	0.1 in. WG
	Filter and coils	2.5 in. WG
	Supply main duct	0.85 in. WG
	Return system between point r and ru	0.2 in. WG
Effective leakage area on the building shell		3.35 ft^2
Minimum outdoor ventilation air		3000 cfm
Space pressure at design flow		+0.05 in. WG

Point	**o**	**m**	**r**	**ru**
Design flow:				
p_o, in. WG	0	–0.20	+0.05	–0.15
$\dot{V}$, cfm	3,000	20,000	20,000	17,000
50% design:				
p_o, in. WG	0	–0.092	+0.0225	–0.052
$\dot{V}$, cfm	2,025	10,000	10,000	7,975

Here, o represents outdoor, m the mixing box, r the space, and ru the recirculating air inlet to the PU.

When the supply volume flow is reduced from the design volume flow to 50% of design flow during the recirculating mode, the total pressure in the mixing box increases from –0.20 to –0.092 in. WG and the outdoor air intake reduces from 3000 to 2025 cfm. Refer to Wang's (1993) *Handbook of Air Conditioning and Refrigeration* for details.

Supply and Relief Fan Combination

Figure 9.13.2(b) shows the schematic diagrams of an air system of supply fan and relief fan combination. A relief fan is used to relieve undesirable high positive space pressure by extracting space air and relieving it to the outside atmosphere. A relief fan is always installed in the relief flow passage after the junction of return flow, relief flow, and recirculating flow passage, point ru. It is usually energized only when the air system is operated in air economizer mode. A relief fan is often an axial fan. Since the relief fan is not energized during recirculating mode operation, the volume flow and pressure characteristics of a supply fan and relief fan combination are the same as that in a single supply fan and barometric damper combination when they have the same design parameters.

During air economizer mode, the outdoor air damper(s) are fully opened and the recirculating damper closed. The space pressure p_r = +0.05 in. WG is maintained by modulating the relief fan speed or relief damper opening. The pressure and volume flow characteristics of a supply and relief fan combination at design volume flow and 50% of design flow are as follows:

Point	**o**	**m**	**r**	**ru**
Design flow:				
p_o, in. WG	0	–0.20	+0.05	–0.15
$\dot{V}$, cfm	20,000	20,000	20,000	17,000

Point	**o**	**m**	**r**	**ru**
50% design:				
p_o, in. WG	0	–0.057	+0.05	+0.007
$\dot{V}$, cfm	10,000	10,000	10,000	7,000

Supply Fan and Return Fan Combination

A *return fan* is always installed at the upstream of the junction of return, recirculating, and exhaust flow passage, point ru as shown in Figure 9.13.2(c). A supply and return fan combination has similar pressure and volume flow characteristics as that of a supply and relief fan combination, except a higher total

pressure at point ru. If the return fan is improperly selected and has an excessive fan total pressure, total pressure at point m may be positive. There will be no outdoor intake at the PU or AHU, and at the same time there will also be a negative space pressure and an infiltration to the space.

Comparison of These Three Fan Combination Systems

A supply fan and barometric damper combination is simpler and less expensive. It is suitable for an air system which does not operate at air economizer mode and has a low pressure drop in the return system.

For those air systems whose pressure drop of return system is not exceeding 0.3 in. WG, or there is a considerable pressure drop in relief or exhaust flow passage, a supply and relief fan combination is recommended. For air systems whose return system has a pressure drop exceeding 0.6 in. WG, or those requiring a negative space pressure, a supply and return fan combination seems more appropriate.

Year-Round Operation and Economizers

Consider a typical single-duct VAV reheat system to serve a typical floor whose indoor temperature is 75°F with a relative humidity of 50%, as shown in Figure 9.13.3(a). During summer, the off-coil temperature is 55°F. The year-round operation of this air system can be divided into four regions on the psychrometric chart, as shown in Figure 9.13.3(b):

- *Region I — Refrigeration/evaporative cooling.* In this region, the enthalpy of the outdoor air h_o is higher than the enthalpy of the recirculating air h_{ru}, $h_o > h_{ru}$. It is more energy efficient to condition the mixture of recirculating air and minimum outdoor air.
- *Region II — Free cooling and refrigeration.* In this region, $h_o \le h_{ru}$. It is more energy efficient and also provides better indoor air quality to extract 100% outdoor air.
- *Region III — Free cooling evaporative cooling, and refrigeration.* In this region, extract 100% outdoor air for free cooling because $h_o \le h_{ru}$. Use evaporative cooling and refrigeration to cool and humidify if necessary.
- *Region IV — Winter heating.* Maintain a 55°F supply temperature by mixing the recirculating air with the outdoor air until the outdoor air is reduced to a minimum value. Provide heating if necessary.

An economizer is a device consisting of dampers and control that uses the free cooling capacity of either outdoor air or evaporatively cooled water from the cooling tower instead of mechanical refrigeration. An air economizer uses outdoor air for free cooling. There are two kinds of air economizers: enthalpy-based, in which the enthalpy of outdoor and recirculating air is compared, and temperature-based, in which temperature is compared. A water economizer uses evaporatively cooled water.

Fan Energy Use

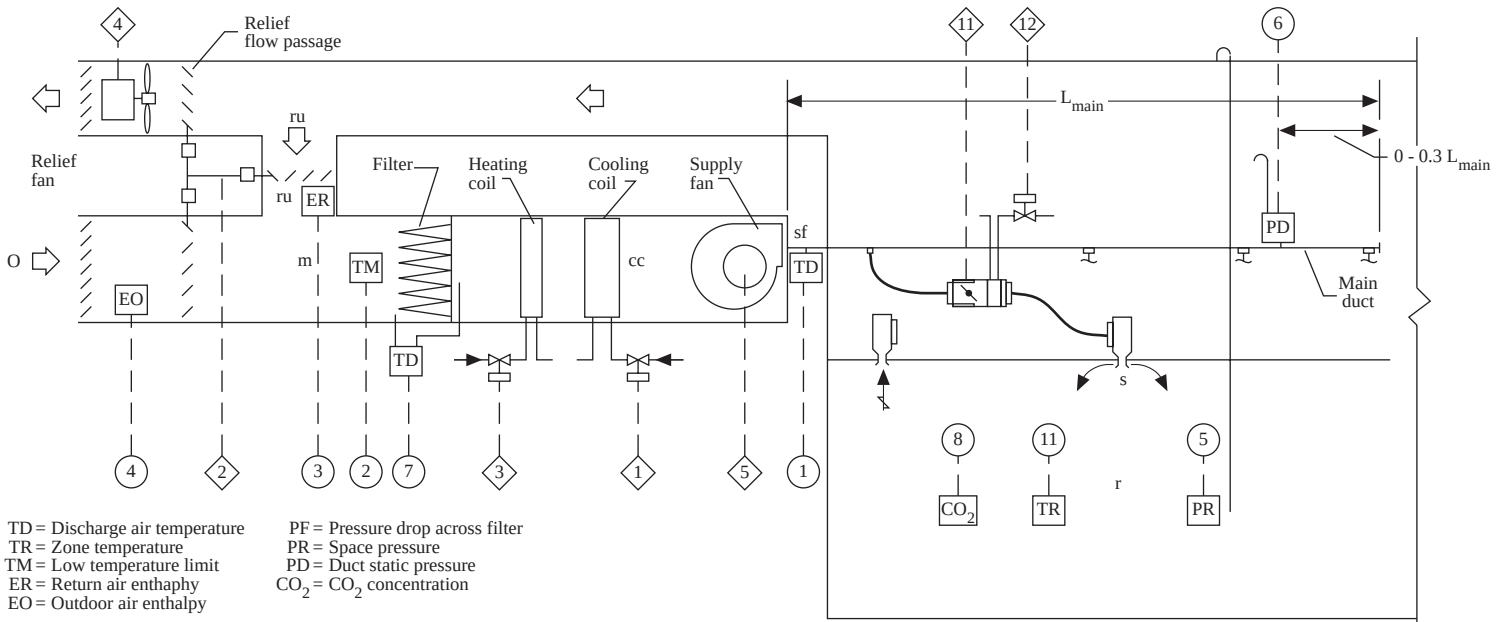
For an air system, fan energy use for each cfm of conditioned air supplied from the AHUs and PUs to the conditioned space within a certain time period, in W/cfm, can be calculated as

$$\text{W/cfm} = 0.1175\Delta p_{sy} / (\eta_f \eta_m) \tag{9.13.7}$$

where Δp_{sy} = mean system total pressure loss during a certain time period, in. WG
η_f = fan total efficiency
η_m = combined motor and drive (direct drive or belt drive) efficiency

For an air system using a separate outdoor ventilation system, its fan energy use, in W/cfm, is then calculated as

$$\text{W/cfm} = (1 + R_{o.s})\left[0.1175\Delta p_{sy} / (\eta_f \eta_m)\right] \tag{9.13.8}$$



(a)

FIGURE 9.13.3 Year-round operation, discharge air temperature, and duct static pressure control for a VAV reheat system: (a) control diagram, (b) year-round operation, and (c) discharge air temperature control output diagram.

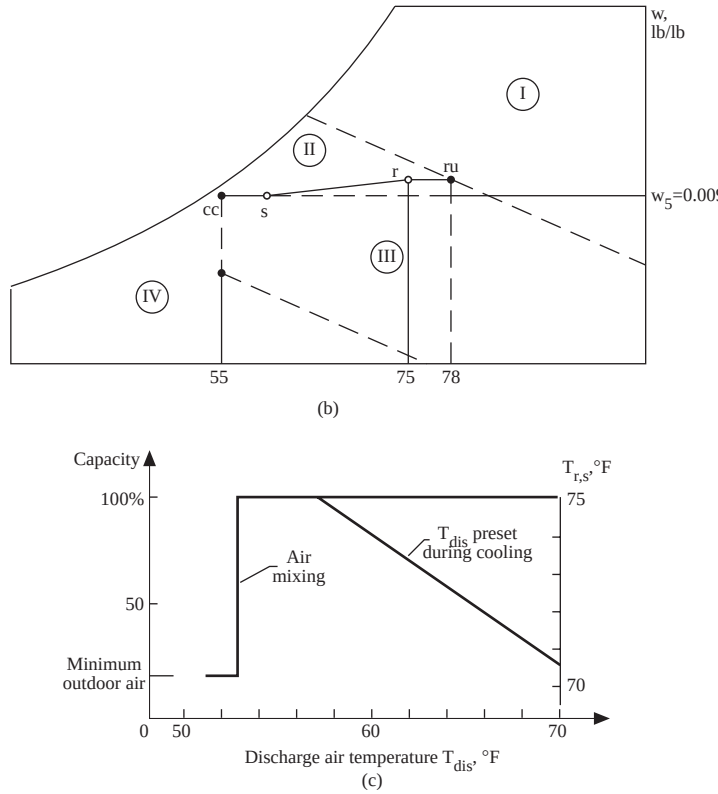


FIGURE 9.13.3bc

where $R_{o.s}$ = ratio of outdoor air volume flow rate to supply volume flow rate.

Outdoor Ventilation Air Supply

Basics

- An adequate amount of outdoor ventilation air supply is the key factor to provide acceptable indoor air quality (IAQ) for a conditioned space. Although an inadequate amount of outdoor ventilation air supply causes poor IAQ, an oversupply of outdoor ventilation air other than in an air economizer cycle is often a waste of energy.
- According to local codes and ANSI/ASHRAE Standard 62-1989, the minimum outdoor ventilation rate for each person must be provided at the outdoor air intake of AHU or PU, or by an outdoor air ventilation system. If the minimum outdoor ventilation air rate is reduced by using higher efficiency filters to remove air contaminants in the recirculating air, then indoor air contaminant concentration must be lower than the specified level in ANSI/ASHRAE Standard 62-1989.
- For a multizone air system, although the ratio of outdoor ventilation air rate to supply air volume flow rate required may be varied from zone to zone, the excessive outdoor air supply to a specified zone will increase the content of unused outdoor air in the recirculating air in AHU or PU. This helps to solve the problem in any zone that needs more outdoor air.
- Since the occupancy in many buildings is often variable and intermittent, a demand-based variable amount of outdoor ventilation air control should be used instead of time-based constant volume outdoor ventilation air control, except during the air economizer cycle.

- Carbon dioxide (CO_2) is a gaseous body effluent. CO_2 is an indicator of representative odor and an indicator of adequate outdoor ventilation rate at specific outdoor and indoor air concentration in a control zone at steady state. For most of the comfort air-conditioning systems, it is suitable to use CO_2 as a key parameter to control the intake volume flow rate of outdoor ventilation air to maintain an indoor CO_2 concentration not exceeding 800 to 1000 ppm in a critical or representative zone. As mentioned in Section 9.5, Persily (1993) showed that the actual measured indoor daily maximum CO_2 concentration levels in five buildings were all within 400 to 820 ppm.

If a field survey finds that a specific indoor air contaminant exceeds a specified indoor concentration, then a gas sensor for this specific contaminant or a mixed gas sensor should be used to control this specific indoor concentration level.

Types of Minimum Outdoor Ventilation Air Control. There are four types of minimum outdoor ventilation air control that are currently used:

- Type I uses a CO_2 sensor or a mixed gas sensor and a DDC controller to control the volume flow rate of outdoor ventilation air for a separate outdoor ventilation air system on the demand-based principle.
- Type II uses a CO_2 or mixed gas sensor and a DDC controller to control the ratio of the openings between outdoor and recirculating dampers and, therefore, the volume flow rates of outdoor air and recirculating air in AHUs or PUs on the demand-based principle.
- Type III uses a flow sensor or a pressure sensor and a DDC controller to control the openings of outdoor and recirculating dampers to provide a nearly constant volume outdoor air intake in VAV AHUs or VAV PUs.
- Type IV adjusts the opening of the outdoor damper manually to provide a constant volume of outdoor air in constant-volume AHUs and PUs. If the outdoor intake is mounted on the external wall without a windshield, the volume flow of outdoor ventilation air intake will be affected by wind force and direction.

Type I is the best minimum outdoor ventilation air control for the air system. For a VAV system, it is expensive. Type II is a better choice. Type III is more complicated and may cause energy waste. Type IV has the same result as Type III and is mainly used in constant-volume systems.

Outdoor intake must be located in a position away from the influence of exhaust outlets. Fans, control dampers, and filters should be properly operated and maintained in order to provide a proper amount of outdoor ventilation air as well as an acceptable IAQ.

9.14 Absorption System

Absorption systems use heat energy to produce refrigeration as well as heating if it is required. Water is the refrigerant and aqueous lithium bromide (LiBr) is widely used as the carrier to absorb the refrigerant and provide a higher coefficient of performance.

The mixture of water and anhydrous LiBr is called *solution*. The composition of a solution is usually expressed by its mass fraction, or percentage of LiBr, often called *concentration*. When the water vapor has boiled off from the solution, it is called *concentration solution*. If the solution has absorbed the water vapor, it is called *diluted solution*.

Absorption systems can be divided into the following categories:

- *Absorption chillers* use heat energy to produce refrigeration.
- *Absorption chiller/heaters* use direct-fired heat input to provide cooling or heating separately.
- *Absorption heat pumps* extract heat energy from the evaporator, add to the heat input, and release them both to the hot water for heating.
- *Absorption heat transformers* raise the temperature of the waste heat source to a required level.

Most recently installed absorption chillers use direct-fired natural gas as the heat source in many locations in the United States where there are high electric demand and electric rate at on-peak hours. Absorption chillers also are free from CFC and HCFC. An energy cost analysis should be done to determine whether an electric chiller or a gas-fired absorption chiller is the suitable choice.

Absorption heat pumps have only limited applications in district heating. Most absorption heat transformers need industrial waste heat. Both of them will not be covered here.

Double-Effect Direct-Fired Absorption Chillers

Figure 9.14.1(a) shows a double-effect direct-fired absorption chiller. *Double effect* means that there are two generators. *Direct fired* means that gas is directly fired at the generator instead of using steam or hot water. A single-effect absorption chiller using steam as the heat input to its single generator has a COP only from 0.7 to 0.8, whereas a double-effect direct-fired absorption chiller has a COP approximately equal to 1 and therefore is the mot widely used absorption chiller in the United States for new and retrofit projects today. The refrigeration capacity of double-effect direct-fired absorption chillers varies from 100 to 1500 tons.

A double-effect direct-fired absorption chiller mainly consists of the following components and controls:

- Evaporator — An *evaporator* is comprised of a tube bundle, spray nozzles, a water trough, a refrigerant pump, and an outer shell. Chilled water flows inside the tubes. A refrigerant pump sprays the liquid refrigerant over the outer surface of the tube bundle for a higher rate of evaporation. A water trough is located at the bottom to maintain a water level for recirculation.
- Absorber — In an *absorber*, there are tube bundles in which cooling water flows inside the tubes. Solution is sprayed over the outer surface of the tube bundle to absorb the water vapor. A solution pump is used to pump the diluted solution to the heat exchanger and low-temperature generator.
- Heat exchangers — There are two heat exchangers: *low-temperature heat exchanger* in which the temperature of hot concentrated solution is lower, and *high-temperature heat exchanger* in which the temperature of hot concentrated solution is higher. In both heat exchangers, heat is transferred from the hot concentrated solution to the cold diluted solution. Shell-and-tube or plate-and-frame heat exchangers are most widely used for their higher effectiveness.
- Generators — *Generators* are also called *desorbers*. In the *direct-fired generator*, there are the fire tube, flue tube, vapor/liquid separator, and flue-gas economizer. Heat is supplied from the gas burner or other waste heat source. The *low-temperature generator* is often of the shell-and-tube type. The water vapor vaporized in the direct-fired generator is condensed inside the tubes. The

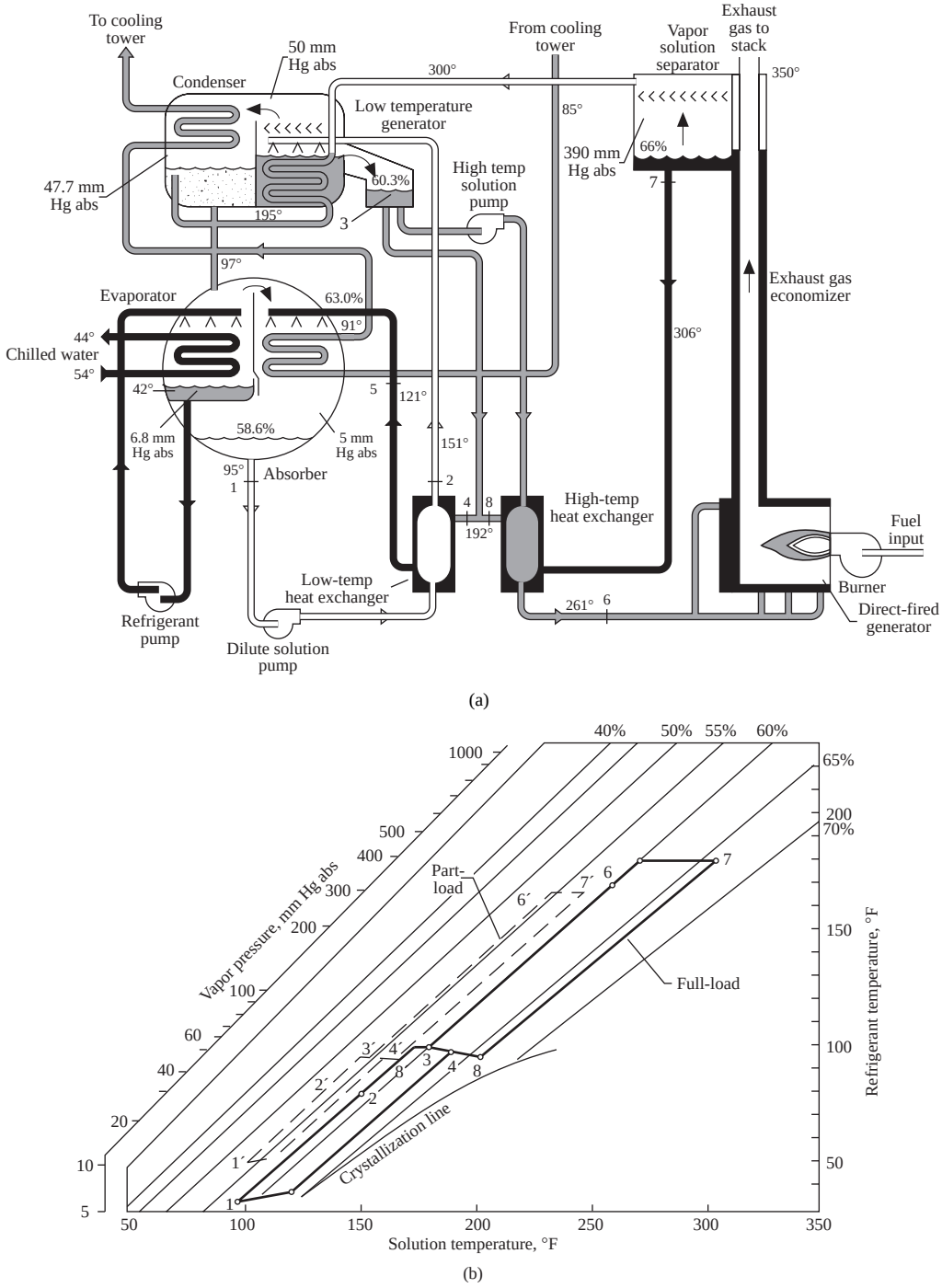


FIGURE 9.14.1 A double-effect direct-fired reverse-parallel-flow absorption chiller: (a) schematic diagram (reprinted by permission from the Trane catalog) and (b) absorption cycle.

latent heat of condensation thus released is used to vaporize the dilute solution in the low-temperature generator.

- Condenser — A condenser is usually also of the shell-and-tube type. Cooling water from the absorber flows inside the tubes.
- Throttling devices — Orifices and valves are often used as throttling devices to reduce the pressure of refrigerant and solution to the required values.
- Air purge unit — Since the pressure inside the absorption chiller is below atmospheric pressure, air and other noncondensable gases will leak into it from the ambient air. An *air purge unit* is used to remove these noncondensable gases from the chiller. A typical air purge unit is comprised of a pickup tube, a purge chamber, a solution spray, cooling water tubes, a vacuum pump, a solenoid valve, and a manual shut-off valve.

When noncondensable gases leak into the system, they tend to migrate to the absorber where pressure is lowest. Noncondensable gases and water vapor are picked from the absorber through the pickup tube. Water vapor is absorbed by the solution spray and returned to the absorber through a liquid trap at the bottom of the purge chamber. Heat of absorption is removed by the cooling water inside the tubes. Noncondensable gases are then evacuated from the chamber periodically by a vacuum pump to the outdoor atmosphere.

Palladium cells are used to continuously remove a small amount of hydrogen that is produced due to corrosion. Corrosion inhibitors like lithium chromate are needed to protect the machine parts from the corrosive effect of the absorbent when air is present.

Absorption Cycles, Parallel-, Series-, and Reverse-Parallel Flow

An *absorption* cycle shows the properties of the solution and its variation in concentrations, temperature, and pressure during absorbing, heat exchanging, and concentration processes on an equilibrium chart as shown in Figure 9.14.1(b). The ordinate of the equilibrium chart is the saturated temperature and pressure of water vapor, in °F and mm Hg abs. The abscissa is the temperature of the solution, in °F. Concentration lines are incline lines. At the bottom of the concentration lines, there is a *crystallization line* or *saturation line*. If the mass of fraction of LiBr in a solution which remains at constant temperature is higher than the saturated condition, that part of LiBr exceeding the saturation condition tends to form solid crystals.

Because there are two generators, the flow of solution from the absorber to generators can be in series flow, parallel flow, or reverse-parallel flow. In a series-flow system, the diluted solution from the absorber is first pumped to the direct-fired generator and then to the low-temperature generator. In a parallel-flow system, diluted solution is pumped to both direct-fired and low-temperature generators in parallel. In a reverse-parallel-flow system as shown in Figure 9.14.1(a), diluted solution is first pumped to the low-temperature generator. After that, the partly concentrated solution is then sent to the direct-fired generator as well as to the intermediate point 4 between high- and low-temperature heat exchangers in parallel. At point 4, partly concentrated solution mixes with concentrated solution from a direct-fired generator. A reverse-parallel-flow system is more energy efficient.

Solution and Refrigerant Flow

In a typical double-effect direct-fired reverse-parallel-flow absorption chiller operated at design full load, water is usually evaporated at a temperature of 42°F and a saturated pressure of 6.8 mm Hg abs in the evaporator. Chilled water returns from the AHUs or fan coils at a temperature typically 54°F, cools, and leaves the evaporator at 44°F. A refrigeration effect is produced due to the vaporization of water vapor and the removal of latent heat of vaporization from the chilled water.

Water vapor in the evaporator is then extracted to the absorber due to its lower vapor pressure. It is absorbed by the concentrated LiBr solution at a pressure of about 5 mm Hg abs. After absorption, the solution is diluted to a concentration of 58.6% and its temperature increases to 95°F (point 1). Most of the heat of absorption and the sensible heat of the solution is removed by the cooling water inside the

tube bundle. Diluted solution is then pumped by a solution pump to the low-temperature generator through a low-temperature heat exchanger.

In the low-temperature generator, the dilute solution is partly concentrated to 60.3% at a solution temperature of 180°F (point 3). It then divides into two streams: one of them is pumped to the direct-fired generator through a high-temperature heat exchanger, and the other stream having a slightly greater mass flow rate is sent to the intermediate point 4. In the direct-fired generator, the concentrated solution leaves at a concentration of 66% and a solution temperature of 306°F (point 7).

The mixture of concentrated and partly concentrated solution at point 4 has a concentration of 63% and a temperature of 192°F. It enters the low-temperature heat exchanger. Its temperature drops to 121°F before entering the absorber (point 5).

In the direct-fired generator, water is boiled off at a pressure of about 390 mm Hg abs. The boiled-off water vapor flows through the submerged tube in the low-temperature generator. The release of latent heat of condensation causes the evaporation of water from the dilution solution at a vapor pressure of about 50 mm Hg abs. The boiled-off water vapor in the low-temperature generator flows to the condenser through the top passage and is condensed into liquid water at a temperature of about 99°F and a vapor pressure of 47.7 mm Hg abs. This condensed liquid water is combined with the condensed water from the submerged tube at the trough. Both of them return to the evaporator after its pressure is throttled by an orifice plate.

Part-Load Operation and Capacity Control

During part-load operation, a double-effect direct-fired reverse-parallel-flow absorption chiller adjusts its capacity by reducing the heat input to the direct-fired generator through the burner. Lower heat input results at less water vapor boiled off from the solution in the generators. This causes the drop in solution concentration, the amount of water vapor extracted, the rate of evaporation, and the refrigeration capacity. Due to less water vapor being extracted, both evaporating pressure and temperature will rise. Since the amount of water vapor to be condensed is greater than that boiled off from the generators, both the condensing pressure and condensing temperature decrease.

Coefficient of Performance (COP)

The COP of an absorption chiller can be calculated as

$$\text{COP} = 12{,}000/q_{1g} \tag{9.14.1}$$

where q_{lg} = heat input to the direct-fired generator per ton of refrigeration output (Btu/hr.ton).

Safety Controls

Safety controls in an absorption chiller include the following:

- Crystallization controls are devices available to prevent crystallization and dissolve crystals. Absorption chillers are now designed to operate in a region away from the crystallization line. It is no longer a serious problem in newly developed absorption systems. One such device uses a bypass valve to permit refrigerant to flow to the concentration solution line when crystallization is detected. Condenser water temperature is controlled by using a three-way bypass valve to mix the recirculating water with the evaporated cooled water from the tower to avoid the sudden drop of the temperature of concentrated solution in the absorber.
- Low-temperature cut-out control shuts down the absorption chiller if the temperature of the refrigerant in the evaporator falls below a preset limit to protect the evaporator from freezing.
- Chilled and cooling water flow switches stop the absorption chiller when the mass flow rate of chilled water or the supply of cooling water falls below a preset value.
- A high-pressure relief valve is often installed on the shell of the direct-fired generator to prevent its pressure from exceeding a predetermined value.

- Monitoring of low and high pressure of gas supply and flame ignition are required for direct-fired burner(s).
- Interlocked controls between absorption chiller and chilled water pumps, cooling water pumps, and cooling tower fans are used to guarantee that they are in normal operation before the absorption chiller starts.

Absorption Chiller/Heater

A double-effect direct-fired reverse-parallel-flow absorption chiller/heater has approximately the same system components, construction, and flow process as the absorption chiller. The cooling mode operation is the same as in an absorption chiller. During the heating mode of an absorption chiller/heater, its evaporator becomes the condenser and is used to condense the water vapor that has been boiled off from the direct-fired generator. At design condition, hot water is supplied at a temperature of 130 to 140°F. The condenser and the low-temperature generator are not in operation. (See Figure 9.14.1.)

In order to increase the coefficient of performance of the absorption chiller, a triple-effect cycle with three condensers and three desorbers has been proposed and is under development. A triple-effect absorption chiller is predicted to have a coefficient of performance of around 1.5. Its initial cost is also considerably increased due to a greater number of condensers and desorbers.

9.15 Air-Conditioning Systems and Selection

Basics in Classification

The purpose of classifing air-conditioning or HVAC&R systems is to distinguish one type from another so that an optimum air-conditioning system can be selected according to the requirements. Proper classification of air-conditioning systems also will provide a background for using knowledge-based expert systems to help the designer to select an air-conditioning system and its subsystems.

Since air system characteristics directly affect the space indoor environmental parameters and the indoor air quality, the characteristics of an air system should be clearly designated in the classification.

The system and equipment should be compatible with each other. Each system has its own characteristics which are significantly different from others.

Individual Systems

As described in Section 9.1, air conditoning or HVAC&R systems can be classified as individual, space, packaged, and central systems.

Individual systems usually have no duct and are installed only in rooms that have external walls and external windows. Individual systems can again be subdivided into the following.

Room Air-Conditioner Systems

A room air conditioner is the sole factory-fabricated self-contained equipment used in the room air-conditioning system. It is often mounted on or under the window sill or on a window frame as shown in Figure 9.1.1. A room air-conditioner consists mainly of an indoor coil, a small forward-curved centrifugal fan for indoor coil, a capillary tube, a low-efficiency dry and reusable filter, grilles, a thermostat or other controls located in the indoor compartment, and a rotary, scroll, or reciprocating compressor, an outdoor coil, and a propeller fan for the outdoor coil located in the outdoor compartment. There is an outdoor ventilation air intake opening and a manually operated damper on the casing that divides the indoor and outdoor compartments. Room air-conditioners have a cooling capacity between 1/2 to 2 tons.

The following are system characteristics of a room air-conditioner system:

Room heat pump system is a room air-conditioner plus a four-way reversing valve which provides both the summer cooling and winter heating.
Air system: single supply fan
Fan, motor, and drive combined efficiency: 25%
Fan energy use: 0.3 to 0.4 W/cfm
Fan speed: HI-LO 2-speed or HI-MED-LO 3-speed
Outdoor ventilation air system: type IV
Cooling system: DX system, air-cooled condenser
EER 7.5 to 9.5 Btu/hr.W
Evaporating temperature T_{ev} at design load: typically 45°F
Heating system: electric heating (if any)
Part-load: on–off of refrigeration compressor
Sound level: indoor NC 45 to 50
Maintenance: More maintenance work is required.

Summer and winter mode air-conditioning cycles of a room air-conditioning system are similar to that shown in Figure 9.3.4.

All fan, motor, and drive combined efficiencies for various air-conditioning systems are from data in ASHRAE Standard 90.1-1989.

Packaged Terminal Air-Conditioner (PTAC) Systems

A packaged terminal air-conditioner is the primary equipment in a PTAC system. A PTAC system is similar to a room air-conditioner system. Their main differences are

- A PTAC uses a wall sleeve and is intended to be mounted through the wall.
- Heating is available from hot water, steam, heat pump, electric heater, and sometimes even direct-fired gas heaters.

PTACs are available in cooling capacity between 1/2 to 1 1/2 tons and a heating capacity of 2500 to 35,000 Btu/hr.

Space (Space-Conditioning) Systems

Most space conditioning air-conditioning systems cool, heat, and filtrate their recirculating space air above or in the conditioned space. Space conditioning systems often incorporate heat recovery by transferring the heat rejected from the interior zone to the perimeter zone through the condenser(s). Space systems often have a separate outdoor ventilation air system to supply the required outdoor ventilation air.

Space systems can be subdivided into four-pipe fan-coil systems and water-source heat pump systems.

Four-Pipe Fan-Coil Systems

In a four-pipe fan-coil unit system, space recirculating air is cooled and heated at a fan coil by using four pipes: one chilled water supply, one hot water supply, one chilled water return, and one hot water return. Outdoor ventilation air is conditioned at a make-up AHU or primary AHU. It is then supplied to the fan coil where it mixes with the recirculating air, as shown in Figure 9.15.1(a), or is supplied to the conditioned space directly.

A *fan-coil unit* or a *fan coil* is a *terminal* as shown in Figure 9.15.1(b). Fan-coil units are available in standard sizes 02, 03, 04, 06, 08, 10, and 12 which correspond to 200 cfm, 400 cfm, and so on in volume flow.

The following are system characteristics of a four-pipe fan-coil system:

A *two-pipe fan-coil system* has a supply and a return pipe only. Because of the problems of changeover from chilled water to hot water and vice versa, its applications are limited.

A *water-cooling electric heating fan-coil system* uses chilled water for cooling and an electric heater for heating as shown in Figure 9.1.2. This system is often used in a location that has a mild winter.

Air system:

Fan-coil, space air recirculating
Fan, motor, and drive combined efficiency: 25%
Fan speed: HI-LO 2-speed and HI-MED-LO 3-speed
External pressure for fan coil: 0.06 to 0.2 in. WG
System fan(s) energy use: 0.45 to 0.5 W/cfm
No return air and return air duct
Outdoor ventilation air system: type I
An exhaust system to exhaust part of the outdoor ventilation air

Cooling system: chilled water from centrifugal or screw chiller

Water-cooled chiller energy use: 0.4 to 0.65 kW/ton

Heating system: hot water from boiler, electric heater

Part load: control the flow rate of chilled and hot water supplied to the coil. Since air leaving coil temperature T_{cc} rises during summer mode part load, space relative humidity will be higher.

Sound level: indoor NC 40 to 45

Maintenance: higher maintenance cost

System fan(s) energy use: 0.45 to 0.55 W/cfm (includes all fans in the four-pipe fan-coil system)

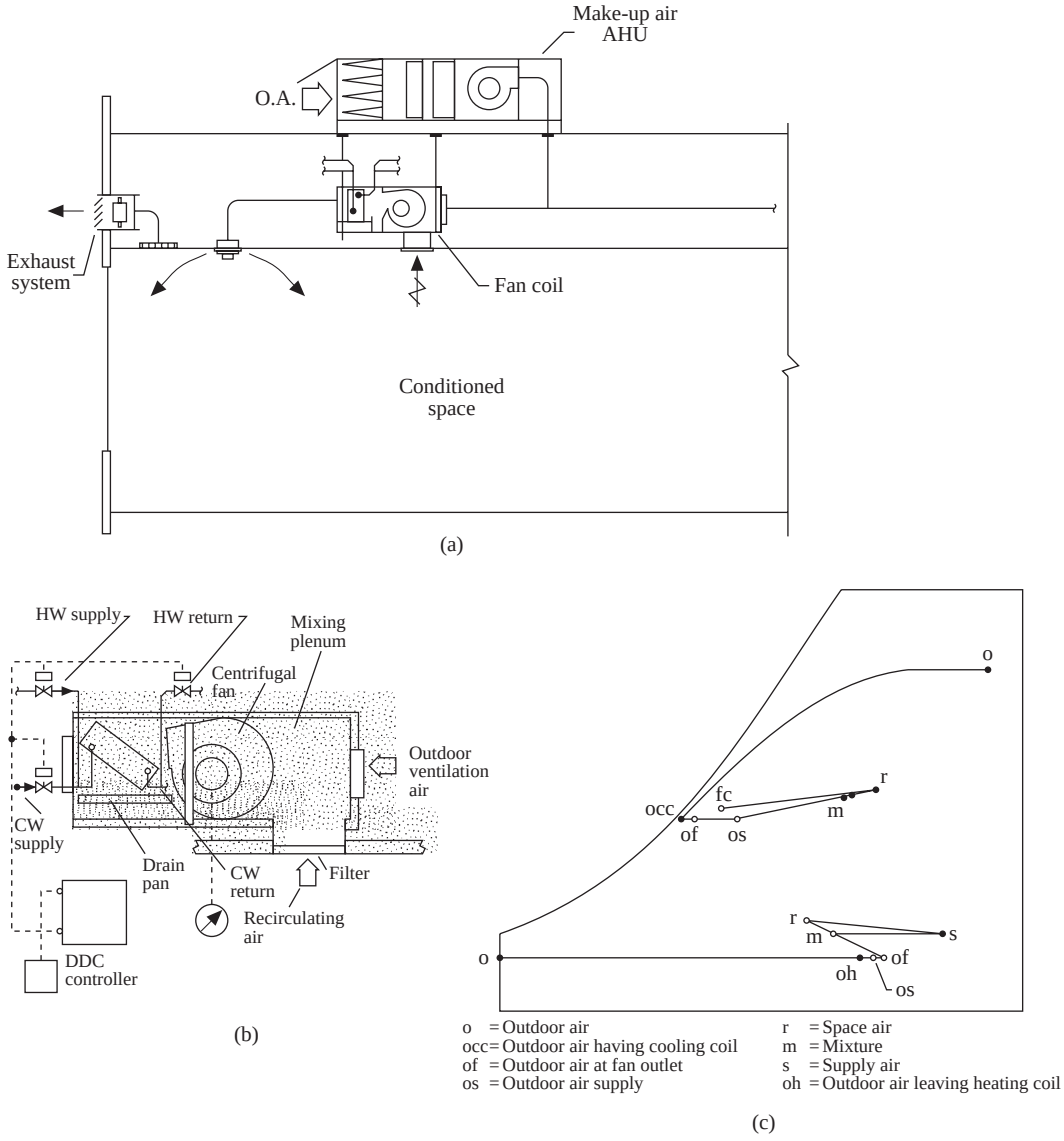


FIGURE 9.15.1 A four-pipe fan-coil system: (a) schematic diagram, (b) fan-coil unit, and (c) air-conditioning cycle.

An air-conditioning cycle for a four-pipe fan-coil system with outdoor ventilation air delivered to the suction side of the fan coil is shown in Figure 9.15.1(c). A part of the space cooling and dehumidifying load is usually taken care by the conditioned outdoor ventilation air from the make-up AHU. A double-bundle condenser is often adopted in a centrifugal chiller to incorporate heat recovery for providing winter heating.

Water-Source Heat Pump Systems

Water-source heat pumps (WSHPs) are the primary equipment in a water-source heat pump system as shown in Figure 9.15.2(a). A *water-source heat pump* usually consists of an air coil to cool and heat the air; a water coil to reject and extract heat from the condenser water; a forward-curved centrifugal fan; reciprocating, rotary, or scroll compressor(s); a short capillary tube; a reversing valve; controls; and an outer casing. WSHPs could be either a horizontal or vertical unit. WSHPs usually have cooling

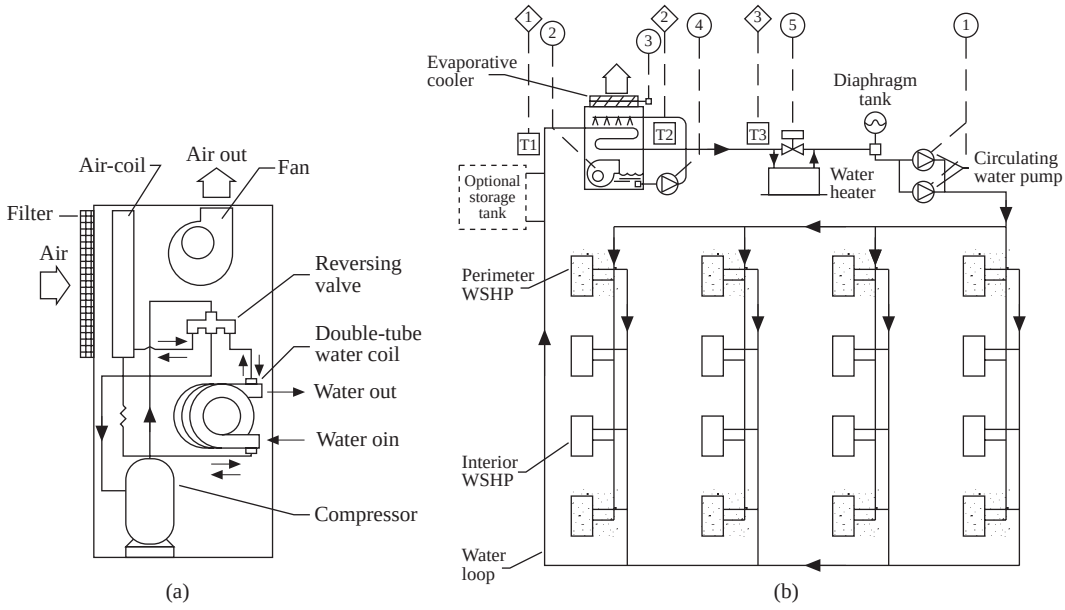


FIGURE 9.15.2 A water-source heat pump system: (a) vertical system and (b) system schematic diagram..

capacities between 1/2 to 26 tons. Small-capacity WSHPs of 3 tons or less without ducts are used in perimeter zones, whereas large-capacity WSHPs with ductwork are used only in interior zones.

In addition to the WSHPs, a WSHP system usually is also comprised of an evaporative cooler or cooling tower to cool the condenser water; a boiler to provide the supplementary heat for the condenser water if necessary; two circulating pumps, one of them being standby; and controls, as shown in Figure 9.15.2(b). A separate outdoor ventilation air system is required to supply outdoor air to the WSHP or directly to the space.

During hot weather, such as outdoor wet bulb at 78°F, all the WSHPs are operated in cooling mode. Condenser water leaves the evaporative cooler at a temperature typically 92°F and absorbs condensing heat rejected from the condensers — the water coils in WSHPs. Condenser water is then raised to 104°F and enters the evaporative cooler. In an evaporative cooler, condenser water is evaporatively cooled indirectly by atmospheric air, so that it would not foul the inner surface of water coils in WSHPs.

During moderate weather, the WSHPs serving the shady side of a building may be in heating mode, and while serving the sunny side of the building and the interior space in cooling mode. During cold weather, most of the WSHPs serving perimeter zones are in heating mode, while serving interior spaces are in cooling mode except morning warm-up. Cooling WSHPs reject heat to the condenser water loop; meanwhile heating WSHPs absorb heat from the condenser water loop. The condenser water is usually maintained at 60 to 90°F. If its temperature rises above 90°F, the evaporative cooler is energized. If it drops below 60°F, the boiler or electric heater is energized. A WSHP system itself is a combination of WSHP and a heat recovery system to transfer the heat from the interior space and sunny side of the building to the perimeter zone and shady side of building for heating in winter, spring, and fall.

System characteristics of air, cooling, and heating in a WSHP system are similar to a room conditioner heat pump system. In addition:

Outdoor ventilating air system: type I and IV
Water system: two-pipe, close circuit
Centrifugal water circulating pump
Water storage tank is optional

To prevent freezing in locations where outdoor temperature may drop below 32°F, isolate the outdoor portion of the water loop, outdoor evaporative cooler, and the pipe work from the indoor portion by using a plate-and-frame heat exchanger. Add ethylene or propylene glycol to the outdoor water loop for freezing protection.

There is another space system called a panel heating and cooling system. Because of its higher installation cost and dehumidification must be performed in the separate ventilation systems, its applications are very limited.

A space conditioning system has the benefit of a separate demand-based outdoor ventilation air system. A WSHP system incorporates heat recovery automatically. However, its indoor sound level is higher; only a low-efficiency air filter is used for recirculating air, and more space maintenance is required than central and packaged systems. Because of the increase of the minimum outdoor ventilation air rate to 15 cfm/person recently, it may gain more applications in the future.

Packaged Systems

In packaged systems, air is cooled directly by a DX coil and heated by direct-fired gas furnace or electric heater in a packaged unit (PU) instead of chilled and hot water from a central plant in a central system. Packaged systems are different from space conditioning systems since variable-air-volume supply and air economizer could be features in a packaged system. Packaged systems are often used to serve two or more rooms with supply and return ducts instead of serving individual rooms only in an individual system.

As mentioned in Section 9.7, packaged units are divided according to their locations into rooftop, split, or indoor units. Based on their operating characteristics, packaged systems can be subdivided into the following systems:

Single-Zone Constant-Volume (CV) Packaged Systems

Although a single-zone CV packaged system may have duct supplies to and returns from two or more rooms, there is only a single zone sensor located in the representative room or space. A CV system has a constant supply volume flow rate during operation except the undesirable reduction of volume flow due to the increase of pressure drop across the filter.

A single-zone CV packaged system consists mainly of a centrifugal fan, a DX coil, a direct-fired gas furnace or an electric heater, a low or medium efficiency filter, mixing box, dampers, DDC controls, and an outer casing. A relief or a return fan is equipped for larger systems.

A single-zone CV packaged system serving a church is shown in Figure 9.1.3. This system operates on basic air-conditioning cycles as shown in Figure 9.3.4 during cooling and heating modes.

The system characteristics of a single-zone CV packaged system are

Air system: single supply fan, a relief or return fan for a large system
- Fan, motor, and drive combined efficiency: 40 to 45%
- Fan total pressure: 1.5 to 3 in. WG
- Fan(s) energy use: 0.45 to 0.8 W/cfm
- Outdoor ventilation air system: type IV and II
- Enthalpy or temperature air economizer

Cooling systems: DX system, air cooled
- Compressor: reciprocating or scroll
- EER: 8.9 to 10.0 Btu/hr.W

Heating system: direct-fired gas furnace, air-source heat pump, or electric heating

Part load: on–off or step control of the compressor capacity, DX-coil effective area, and the gas flow to the burner

Sound level: indoor NC 35 to 45

Maintenance: higher maintenance cost than central systems

Single-zone, CV packaged systems are widely used in residences, small retail stores, and other commercial buildings.

Constant-Volume Zone-Reheat Packaged Systems

System construction and system characteristics of a CV zone-reheat system are similar to the single-zone CV packaged systems except:

1. It serves multizones and has a sensor and a DDC controller for each zone.
2. There is a reheating coil or electric heater in the branch duct for each zone.

A CV zone-reheat packaged system cools and heats simultaneously and therefore wastes energy. It is usually used for the manufacturing process and space needs control of temperature and humidity simultaneously.

Variable-Air-Volume Packaged Systems

A variable-air-volume (VAV) system varies its volume flow rate to match the reduction of space load at part load. A VAV packaged system, also called a *VAV cooling packaged system*, is a multizone system and uses a VAV box in each zone to control the zone temperature at part load during summer cooling mode operation, as shown in Figure 9.15.3(a).

A *VAV box* is a terminal in which the supply volume flow rate of the conditioned supply air is modulated by varying the opening of the air passage by means of a single blade damper, as shown in Figure 9.15.3(b), or a moving disc against a cone-shaped casing.

The following are the system characteristics of a VAV packaged system:

Single-zone VAV packaged system which serves a single zone without VAV boxes. A DDC controller modulates the position of the inlet vanes or the fan speed according to the signal of the space temperature sensor.

Air system: a supply/relief fan or supply/return fan combination. Space pressurization control by a relief/return fan
Fan, motor, and drive combined efficiency: 45%
Supply fan total pressure: 3.75 to 4.5 in. WG
Fan(s) energy use at design condition: 1 to 1.25 W/cfm
VAV box minimum setting: 30% of peak supply volume flow
Outdoor ventilation air system: type II and III
Economizer: enthalpy air economizer or water economizer
Cooling system: DX coil, air-, water-, or evaporative-cooled condenser
Compressor: reciprocating, scroll, and screw
EER: 8.9 to 12 Btu/hr.W
Capacity: 20 to 100 tons
Part load: zone volume flow modulation by VAV box; step control of compressor capacity; modulation of gas flow to burner; and discharge air temperature reset
Smoke control: exhausts smoke on the fire floor, and supplies air and pressurizes the floors immediately above or below the fire floor
Diagnostics: a diagnostic module displays the status and readings of various switches, dampers, sensors, etc. and the operative problems by means of expert system
Maintenance: higher than central system
Sound level: indoor NC 30 to 45

Heating system characteristics as well as the air-conditioning cycles are similar as that in a single-zone CV packaged system.

VAV Reheat Packaged Systems

A VAV reheat packaged system has its system construction and characteristics similar to that in a VAV packaged system except in each VAV box there is an additional reheating coil. Such a VAV box is called

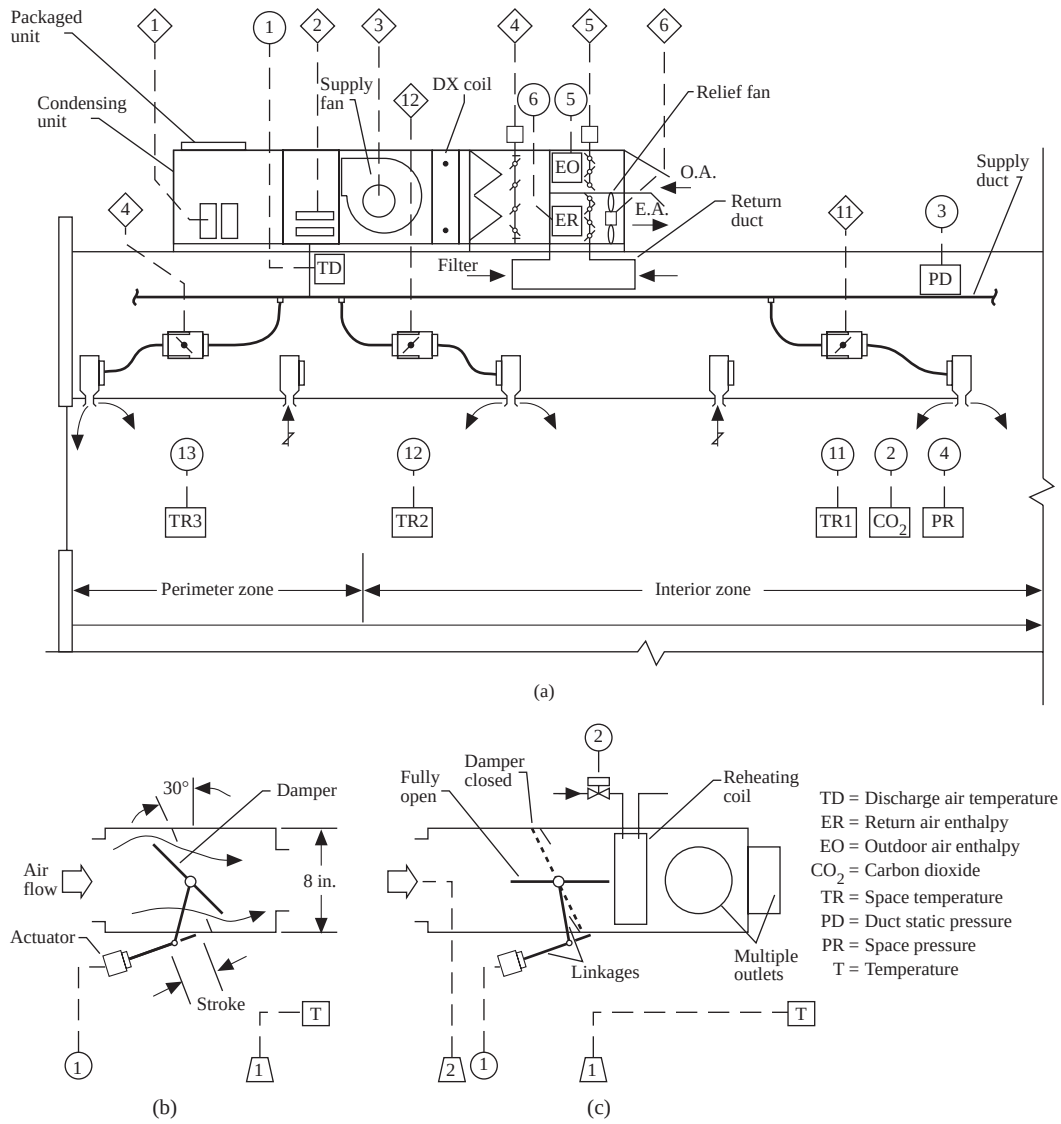


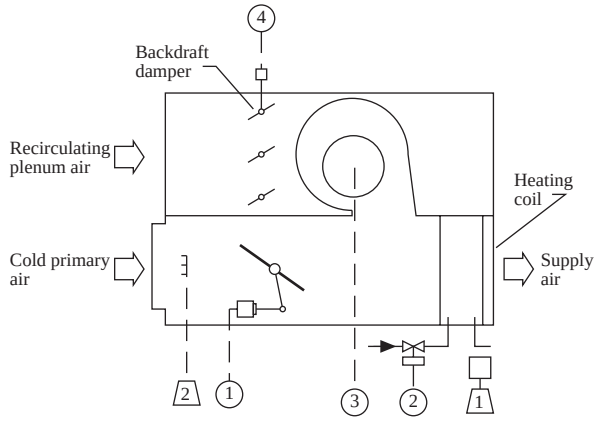
FIGURE 9.15.3 A variable-air-volume (VAV) package system: (a) schematic diagram, (b) VAV box, (c) reheating box, (d) parallel-flow fan-powered VAV box.

a *reheating VAV box*, as shown in Figure 9.15.2(a) and 9.15.3(c). VAV reheat packaged systems are used to serve perimeter zones where winter heating is required.

Fan-Powered VAV Packaged Systems

A fan-powered VAV packaged system is similar to that of a VAV packaged system except *fan-powered VAV boxes* as shown in Figure 9.15.3(d) are used instead of VAV boxes.

There are two types of fan-powered VAV boxes: parallel-flow and series-flow boxes. In a *parallel-flow* fan-powered box, the plenum air flow induced by the fan is parallel with the cold primary air flow through the VAV box. These two air streams are then combined and mixed together. In a *series-flow* box, cold primarily from the VAV box is mixed with the induced plenum air and then flows through the small fan. The parallel-flow fan-powered VAV box is more widely used.



(d)

FIGURE 9.15.3d

In a fan-powered VAV box, volume flow dropping to minimum setting, extracting of ceiling plenum air, and energizing of reheating coil will actuate in sequence to maintain the space temperature during part-load/heating mode operation. A fan-powered VAV box can also mix the cold primary air from cold air distribution with the ceiling plenum air and provides greater space air movements during minimum space load.

Packaged systems are lower in installation cost and occupy less space than central systems. During the past two decades, DDC-controlled packaged systems have evolved into sophisticated equipment and provide many features that only a built-up central system could provide before.

Central Systems

Central systems use chilled and hot water that comes from the central plant to cool and heat the air in the air-handling units (AHUs). Central systems are built-up systems. The most clean, most quiet thermal-storage systems, and the systems which offer the most sophisticated features, are always central systems. Central systems can be subdivided into the following.

Single-Zone Constant-Volume Central Systems

A single-zone CV central system uses a single controller to control the flow of chilled water, hot water, or the opening of dampers to maintain a predetermined indoor temperature, relative humidity, or air contaminants. They are often used in manufacturing factories. The system characteristics of a single-zone CV central system are

Single-zone CV air washer central system uses air washer to control both space relative humidity and temperature. This system is widely used in textile mills. The reason to use constant volume is to dilute the fiber dusts produced during manufacturing. A rotary filter is often used for high dust-collecting capacity.

Air system: supply and return fan combination
- Fan, motor, and drive combined efficiency: 45 to 50%
- Outdoor ventilation air system: type II and IV

Economizer: air or water economizer
Smoke control: exhaust smoke on the fire floor, and pressurize adjacent floor(s) or area
Cooling system: centrifugal or screw chiller, water-cooled condenser
- Cooling energy use: 0.4 to 0.65 kW/ton

Heating system: hot water from boiler or from heat recovery system

Part load: modulate the water mass flow to cooling and heating coils in AHUs, and discharge air temperature reset
Sound level: indoor NC 30 to 45. Silencers are used both in supply and return air systems if they are required
Maintenance: in central plant and fan rooms, lower maintenance cost

Single-Zone CV Clean Room Systems

This is the central system which controls the air cleanliness, temperature, and relative humidity in Class 1, 10, 100, 1000, and 10,000 clean rooms for electronic, pharmaceutical, and precision manufacturing and other industries. Figure 9.15.4(a) shows a schematic diagram of this system. The recirculating air unit (RAU) uses prefilter, HEPA filters, and a water cooling coil to control the space air cleanliness and required space temperature, whereas a make-up air unit (MAU) supplies conditioned outdoor air, always within narrow dew point limits to the RAU at any outside climate, as shown in Figure 9.15.4(b). A unidirectional air flow of 90 fpm is maintained at the working area. For details, refer to *ASHRAE Handbook 1991 HVAC Applications* and Wang's *Handbook of Air Conditioning and Refrigeration.*

CV Zone-Reheat Central Systems

These systems have their system construction and characteristics similar to that for a single-zone CV central system, except they serve multizone spaces and there is a reheating coil, hot water, or electric heating in each zone. CV zone-reheat central systems are often used for health care facilities and in industrial applications.

VAV Central Systems

A VAV central system is used to serve multizone space and is also called *VAV cooling central system.* Its schematic diagram is similar to that of a VAV packaged system (Figure 9.15.3) except air will be cooled or heated by water cooling or heating coils in the AHUs. The same VAV box shown in Figure 9.15.3(b) will be used in a VAV central system. The system characteristics of VAV central systems are as follows:

Single-zone VAV central system differs from a VAV central system only because it serves a single zone, and therefore there is no VAV box in the branch ducts. Supply volume flow is modulated by inlet vanes and AC inverter.

Air system: supply and relief/return fan combination
Fan, motor, and drive combined efficiency for airfoil centrifugal fan with AC inverter fan speed modulation: 55%
Fan(s) energy use: 0.9 to 1.2 W/cfm
VAV box: minimum setting 30% of peak supply volume flow
Outdoor ventilation air system: type I, II, and III
Cooling system: centrifugal, screw, and reciprocating chillers, water-cooled condenser, with energy use 0.4 to 0.65 kW/ton; or sometimes absorption chiller
Heating system: hot water from boiler or electric heating at the terminals
Economizer: air and water economizer
Part load: zone volume flow modulation, discharge air temperature reset, and chilled water temperature reset
Smoke control: exhausts smoke from the fire floor and pressurizes the immediate floors above and below
Sound level: indoor NC 20 to 40. Silencers are often used both in supply and return systems.
Maintenance: less space maintenance

VAV central systems are widely used for interior zone in buildings.

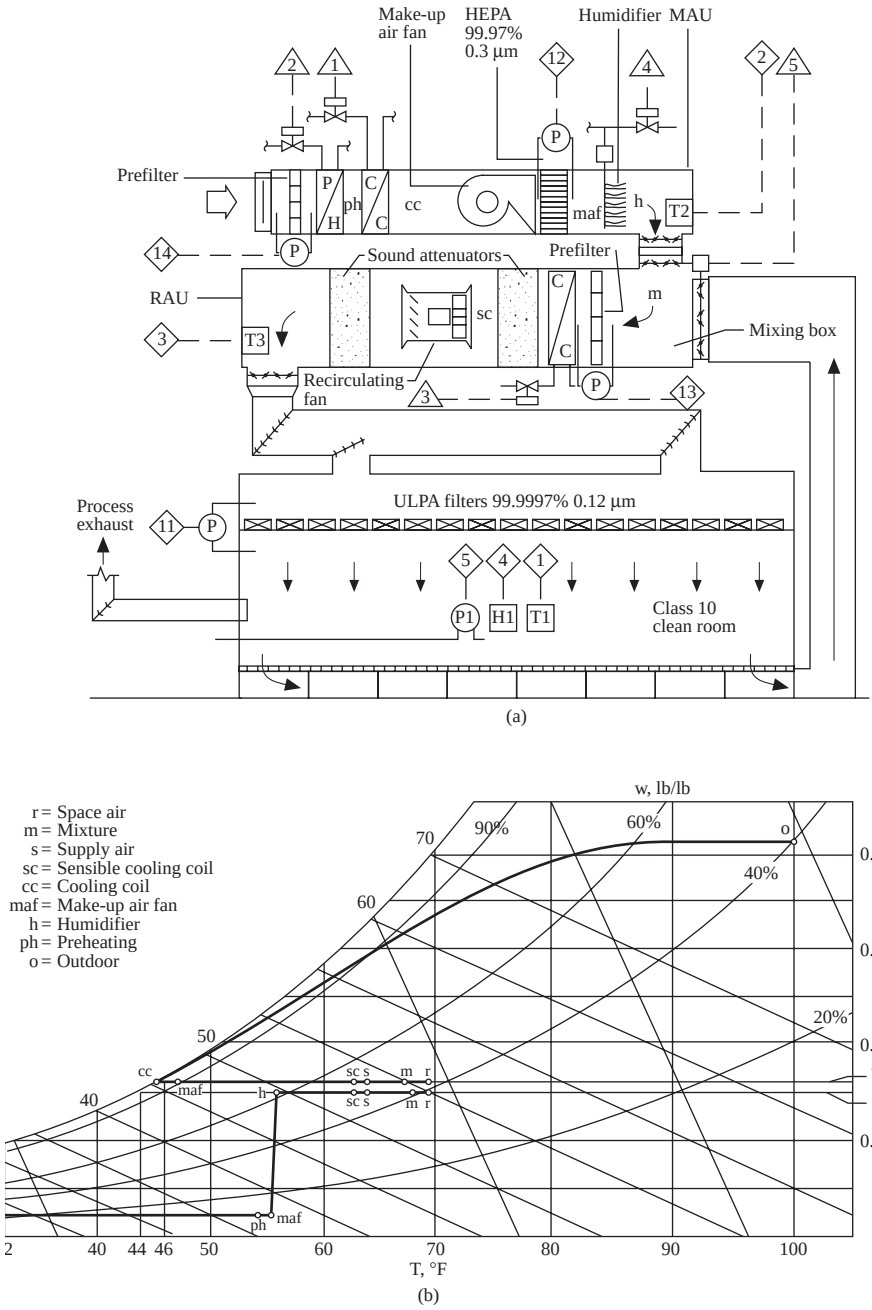


FIGURE 9.15.4 A single-zone CV clean room system: (a) schematic diagram and (b) air-conditioning cycle.

VAV Reheat Central Systems

A VAV reheat system is similar in system construction and characteristics to that in a VAV central system except that reheating boxes are used instead of VAV boxes in a VAV central system.

Fan-Powered VAV Central Systems

A fan-powered VAV central system is similar in system construction and characteristics to that in a VAV central system except that fan-powered VAV boxes are used instead of VAV boxes.

Dual-Duct VAV Central Systems

A dual-duct VAV system uses a warm air duct and a cold air duct to supply both warm and cold air to each zone in a multizone space, as shown in Figure 9.15.5(a). Warm and cold air are first mixed in a mixing VAV box, and are then supplied to the conditioned space. Warm air duct is only used for perimeter zones.

A *mixing VAV box* consists of two equal air passages, one for warm air and one for cold air, arranged in parallel. Each of them has a single blade damper and its volume flow rate is modulated. Warm and cold air are then combined, mixed, and supplied to the space.

A dual-duct VAV system is usually either a single supply fan and a relief/return fan combination, or a warm air supply fan, a cold air supply fan, and a relief/return fan. A separate warm air fan and cold air supply fan are beneficial in discharge air temperature reset and fan energy use.

During summer cooling mode operation, the mixture of recirculating air and outdoor air is used as the warm air supply. The heating coil is not energized. During winter heating mode operation, mixture of outdoor and recirculating air or 100% outdoor air is used as the cold air supply; the cooling coil is not energized.

Because there is often air leakage at the dampers in the mixing VAV box, more cold air supply is needed to compensate for the leaked warm air or leaked cold air.

Other system characteristics of a dual-duct VAV central system are similar to a VAV central system.

Dual-Duct CV Central System

This is another version of a dual-duct VAV central system and is similar in construction to a dual-duct VAV system, except that a mixing box is used instead of a mixing VAV box. The supply volume flow rate from a mixing box is nearly constant. Dual-duct CV central systems have only limited applications, like health care facilities, etc.

An ice- or chilled-water storage system is always a central system plus a thermal storage system. The thermal storage system does not influence the system characteristics of the air distribution, and air cooling and heating — except for a greater head lift for a refrigeration compressor — is needed for ice-storage systems. Therefore, the following central systems should be added:

- VAV ice-storage or chilled-water systems
- VAV reheat ice-storage or chilled-water storage systems
- Fan-powered VAV ice-storage systems

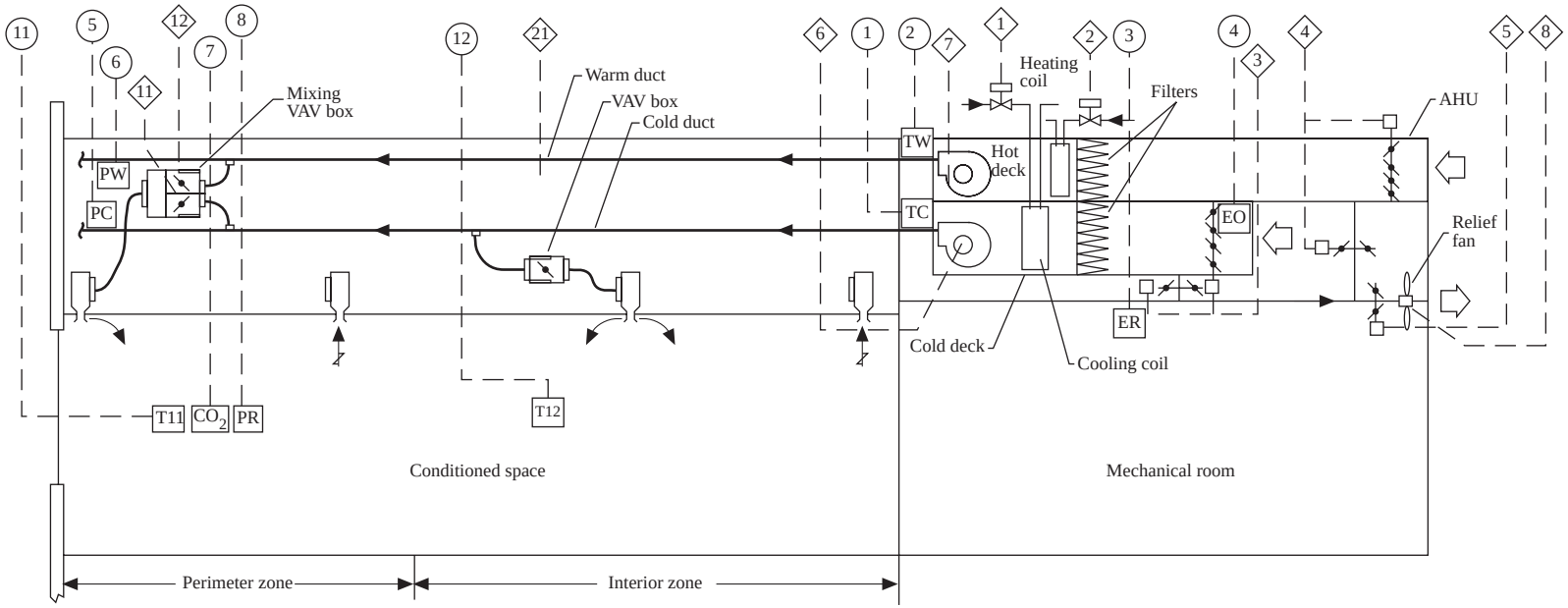
Some of the air-conditioning systems are not listed because they are not effective or are a waste of energy, and therefore rarely used in new and retrofit projects such as:

- High-velocity induction space conditioning systems which need a higher pressure drop primary air to induce recirculating air in the induction unit and use more energy than fan-coil systems
- Multizone central systems which mix warm and cool air at the fan room and use a supply duct from fan room to each control zone
- Air skin central systems which use a warm air heating system to offset transmission loss in the perimeter zone and overlook the effect of the solar radiation from variation building orientations

In the future, there will be newly developed systems added to this classification list.

Air-Conditioning System Selection

As described in Section 9.1, the goal of an air-conditioning or HVAC&R system is to provide a healthy, comfortable, manufacturable indoor environment at acceptable indoor air quality, keeping the system



(a)

FIGURE 9.15.5 A dual-duct VAV central system: (a) schematic diagram.

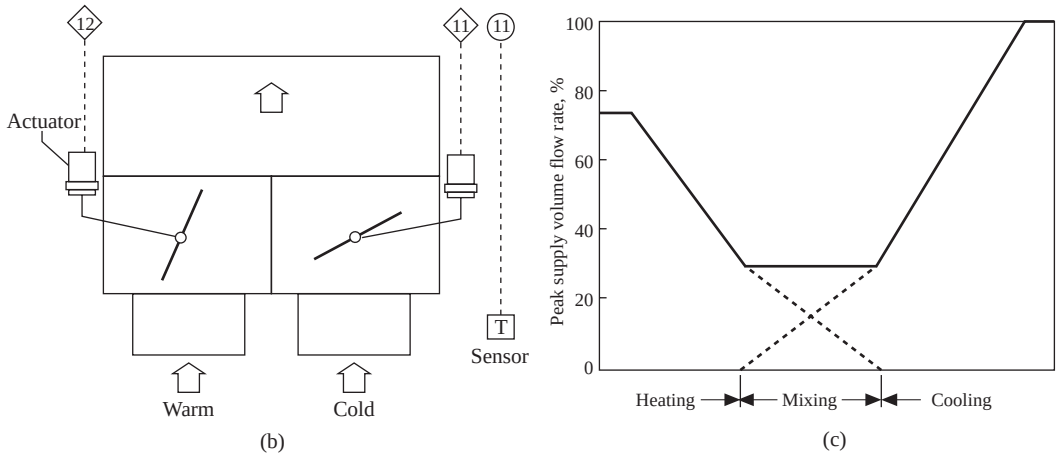


FIGURE 9.15.5 A dual-duct VAV central system: (b) mixing VAV box and (c) volume flow-operating mode diagram.

energy efficient. Various occupancies have their own requirements for their indoor environment. The basic considerations to select an air-conditioning system include:

1. The selection of an air-conditioning system must satisfy the required space temperature, relative humidity, air cleanliness, sound level, and pressurization. For a Class 100 clean room, a single-zone CV clean room system is always selected. A four-pipe fan-coil space conditioning system is usually considered suitable for guest rooms in hotels for operative convenience, better privacy, and a guaranteed outdoor ventilation air system. A concert hall needs a very quiet single-zone VAV central system for its main hall and balcony.
2. The size of the project has a considerable influence on the selection. For a small-size residential air-conditioning system, a single-zone constant-volume packaged system is often the first choice.
3. Energy-efficient measures are specified by local codes. Comparison of alternatives by annual energy-use computer programs for medium and large projects is often necessary. Selection of energy source includes electricity or gas, and also using electrical energy at off-peak hours, like thermal storage systems is important to achieve minimum energy cost.

   For a building whose sound level requirement is not critical and conditioned space is comprised of both perimeter and interior zones, a WSHP system incorporating heat recovery is especially suitable for energy saving.
4. First cost or investment is another critical factor that often determines the selection.
5. Selection of an air-conditioning system is the result of synthetical assessment. It is difficult to combine the effect of comfort, reliability, safety, and cost. Experience and detailed computer program comparisons are both important.

The selection procedure usually begins whether an individual, space conditioning, packaged, central system, or CV, VAV, VAV reheat, fan-powered VAV, dual-duct VAV, or thermal storage system is selected. Then the air, refrigeration, heating, and control subsystems will be determined. After that, choose the option, the feature, the construction, etc. in each subsystem.

Comparison of Various Systems

The sequential order of system performance — excellent, very good, good, satisfactory — regarding temperature and relative humidity control (T&HC), outdoor ventilation air (OA), sound level, energy use, first cost, and maintenance for individual, space conditioning (SC), packaged, and central systems is as follows:

	Excellent (low or less)	Very good	Good	Satisfactory
T&HC	Central	Packaged	Space	Individual
IAQ	Space	Central	Packaged	Individual
Sound	Central	Packaged	Space	Individual
Energy use	Individual	Space	Packaged	Central
First cost	Individual	Packaged	Space	Central
Maintenance	Central	Packaged	Space	Individual

Among the packaged and central systems, VAV cooling systems are used only for interior zones. VAV reheat, fan-powered VAV, and dual-duct VAV central systems are all for perimeter zones. VAV reheat systems are simple and effective, but have a certain degree of simultaneous cooling and heating when their volume flow has been reduced to minimum setting. Fan-powered VAV systems have the function of mixing cold primary air with ceiling plenum air. They are widely used in ice-storage systems with cold air distribution. Fan-powered VAV is also helpful to create a greater air movement at minimum cold primary air flow. Dual-duct VAV systems are effective and more flexible in operation. They are also more complicated and expensive.

Subsystems

Air Systems

The economical size of an air system is often 10,000 to 25,000 cfm. A very large air system always has higher duct pressure loss and is more difficult to balance. For highrise buildings of four stories and higher, floor-by-floor AHU(s) or PU(s) (one or more AHU or PU per floor) are often adopted. Such an arrangement is beneficial for the balance of supply and return volume flow in VAV systems and also for fire protection. A fan-powered VAV system using a riser to supply less cold primary to the fan-powered VAV box at various floors may have a larger air system. Its risers can be used as supply and exhaust ducts for a smoke-control system during a building fire.

In air systems, constant-volume systems are widely used in small systems or to dilute air contaminants in health care facilities and manufacturing applications. VAV systems save fan energy and have better operating characteristics. They are widely used in commercial buildings and in many factories.

Refrigeration Systems

For comfort air-conditioning systems, the amounts of required cooling capacity and energy saving are dominant factors in the selection of the refrigeration system. For packaged systems having cooling capacity less than 100 tons, reciprocating and scroll vapor compression systems with air-cooled condensers are most widely used. Evaporative-cooled condensers are available in many packaged units manufactured for their lower energy use. Scroll compressors are gradually replacing the reciprocating compressors for their simple construction and energy saving. For chillers of cooling capacity of 100 tons and greater, centrifugal chillers are still most widely used for effective operation, reliability, and energy efficiency. Screw chillers have become more popular in many applications, especially for ice-storage systems.

Heating Systems

For locations where there is a cold and long winter, a perimeter baseboard hot water heating system or dual-duct VAV systems are often a suitable choice. For perimeter zones in locations where winter is mild, winter heating is often provided by using warm air supply from AHU or PU from terminals with electric or hot water heaters. Direct-fired furnace warm air supply may be used for morning warm-up. For interior or conditioned zones, a cold air supply during occupied periods in winter and a warm air supply from the PUs or AHUs during morning warm-up period is often used.

Control Systems

Today, DDC microprocessor-based control with open data communication protocol is often the choice for medium- and large-size HVAC&R projects. For each of the air, cooling, and heating systems, carefully select the required generic and specific control systems. If a simple control system and a more complicated control system can provide the same required results, the simple one is always the choice.

Energy Conservation Recommendations

1. Turn off electric lights, personal computers, and office appliances when they are not needed. Shut down AHUs, PUs, fan coils, VAV boxes, compressors, fans, and pumps when the space or zone they serve is not occupied or working.
2. Provide optimum start and stop for the AHUs and PUs and terminals daily.
3. Temperature set point should be at its optimum value. For comfort systems, provide a dead band between summer and winter mode operation. Temperature of discharged air from the AHU or PU and chilled water leaving the chiller should be reset according to space or outdoor temperature or the system load.
4. Reduce air leakages from ducts and dampers. Reduce the number of duct fittings and pipe fittings and their pressure loss along the design path if this does not affect the effectiveness of the duct system. The maximum design velocity in ducts for comfort systems should not exceed 3000 fpm, except that a still higher velocity is extremely necessary.
5. Adopt first the energy-efficient cooling methods: air and water economizer, evaporative cooler, or ground water instead of refrigeration.
6. Use cost-effective high-efficiency compressors, fans, pumps, and motors as well as evaporative-cooled condensers in PUs. Use adjustable-frequency fan speed modulation for large centrifugal fans. Equipment should be properly sized. Over-sized equipment will not be energy efficient.
7. Use heat recovery systems and waste heat for winter heating or reheating. Use a heat-pump system whenever its COP_{hp} is greater than 1.
8. For medium- and large-size air-conditioning systems, use VAV systems instead of CV systems except for health care or applications where dilution of air contaminant is needed. Use variable flow for building-loop and distribution-loop water systems.
9. Use double- and triple-pane windows with low emissive coatings. Construct low U-value roofs and external walls.

References

AMCA. 1973. Fan and Systems Publication 201. AMCA, Arlington Heights, IL.

Amistadi, H. 1993. Design and drawing software review, *Eng. Syst.* 6:18–29.

ANSI/ASHRAE. 1992. ANSI/ASHRAE Standard 34-1992, *Numbering Designation and Safety Classification of Refrigerants*. ASHRAE, Atlanta, GA.

ASHRAE. 1991. *ASHRAE Handbook, HVAC Applications*. ASHRAE, Atlanta, GA.

ASHRAE. 1992. *ASHRAE Handbook, HVAC Systems and Equipment*. ASHRAE, Atlanta, GA.

ASHRAE. 1993. *ASHRAE Handbook, Fundamentals*. ASHRAE, Atlanta, GA.

ASHRAE. 1994. *ASHRAE Handbook, Refrigeration*. ASHRAE, Atlanta, GA.

Bayer, C.W. and Black, M.S. 1988. IAQ evaluations of three office buildings. *ASHRAE J.* 7:48–52.

Bushby, S.T. and Newman, H.M. 1994. BACnet: a technical update, *ASHRAE J.* 1:S72–84.

Carlson, G.F. 1968. Hydronic systems: analysis and evaluation, I. *ASHRAE J.* 10:2–11.

DOE. 1981. DOE-2 Reference Material (Version 2.1A). National Technical Information Service, Springfield, VA.

Dorgan, C.E. and Elleson, J.S. 1988. Cold air distribution. *ASHRAE Trans.* I:2008–2025.

Durkin, J. 1994. *Expert Systems Design and Development*. Macmillan, New York.

EIA. 1994. Commercial Buildings Characteristics 1992. U.S. Government Printing Office, Washington, D.C.

Elyashiv, T. 1994. Beneath the surface: BACnet™ data link and physical layer options. *ASHRAE J.* 11:32–36.

EPA/CPSC. 1988. The Inside Story: A Guide to Indoor Air Quality. Environmental Protection Agency, Washington, D.C.

Fanger, P.O., Melikow, A.K., Hanzawa, H., and Ring, J. 1989. Turbulence and draft. *ASHRAE J.* 4:18–25.

Fiorino, D.P. 1991. Case study of a large, naturally stratified, chilled-water thermal storage system. *ASHRAE Trans.* II:1161–1169.

Gammage, R.B., Hawthorne, A.R., and White, D.A. 1986. Parameters Affecting Air Infiltration and Air Tightness in Thirty-One East Tennessee Homes, Measured Air Leakage in Buildings, ASIM STP 904. American Society of Testing Materials, Philadelphia.

Goldschmidt, I.G. 1994. A data communucations introduction to BACnet™. *ASHRAE J.* 11:22–29.

Gorton, R.L. and Sassi, M.M. 1982. Determination of temperature profiles and loads in a thermally stratified air-conditioning system. I. Model studies. *ASHRAE Trans.* II:14–32.

Grimm, N.R. and Rosaler, R.C. 1990. *Handbook of HVAC Design.* McGraw-Hill, New York.

Hartman, T.B. 1989. TRAV — a new HVAC concept. *Heating/Piping/Air Conditioning.* 7:69–73.

Hayner, A.M. 1994. Engineering in quality. *Eng. Syst.* 1:28–33.

Heyt, H.W. and Diaz, M.J. 1975. Pressure drop in spiral air duct. *ASHRAE Trans.* II:221–232.

Huebscher, R.G. 1948. Friction equivalents for round, square, and rectangular ducts. *ASHRAE Trans.* 101–144.

Hummel, K.E., Nelson, T.P., and Tompson, P.A. 1991. Survey of the use and emissions of chlorofluorocarbons from large chillers. *ASHRAE Trans.* II:416–421.

Jakob, F.E., Locklin, D.W., Fisher, R.D., Flanigan, L.G., and Cudnik, L.A. 1986. SP43 evaluation of system options for residential forced-air heating. *ASHRAE Trans.* IIB:644–673.

Kimura, K. 1977. *Scientific Basis of Air Conditioning*. Applied Science Publishers, London.

Knebel, D.E. 1995. Current trends in thermal storage. *Eng. Syst.* 1:42–58.

Korte, B. 1994. The health of the industry. *Heating/Piping/Air Conditioning.* 1:111–112.

Locklin, D.W., Herold, K.E., Fisher, R.D., Jakob, F.E., and Cudnik, R.A. 1987. Supplemental information from SP43 evaluation of system options for residential forced-air heating. *ASHRA Trans.* II:1934–1958.

Lowe, R. and Ares, R. 1995. From CFC-12 to HFC-134a: an analysis of a refrigerant retrofit project. *Heating/Piping/Air Conditioning.* 1:81–89.

McQuiston, F.C. and Spitler, J.D. 1992. *Cooling and Heating Load Calculating Manual*, 2nd ed. ASHRAE, Atlanta, GA.

Mitalas, G.P. 1972. Transfer function method of calculating cooling loads, heat extraction rate and space temperature, *ASHRAE J.* 12:52–56.

Mitalas, G.P. and Stephenson, D.G. 1967. Room thermal response factors. *ASHRAE Trans.* 2, III.2.1.

Modera, M.P. 1989. Residential duct system leakage: magnitude, impact, and potential for reduction. *ASHRAE Trans.* II:561–569.

Molina, M.J. and Rowland, S. 1974. Stratospheric sink for chloromethanes: chlorine atom catalyzed destruction of ozone. *Nature*. 249:810–812.

NIOSH. 1989. Congressional Testimony of J. Donald Miller, M.D., before the Subcommittee of Superfund, Ocean, and Water Protection, May 26, 1989. NIOSH, Cincinnati, Cleveland.

Parsons, B.K., Pesaran, A.A., Bharathan, D., and Shelpuk, B. 1989. Improving gas-fired heat pump capacity and performance by adding a desiccant dehumidification subsystem. *ASHRAE Trans.* I:835–844.

Persily, A.K. 1993. Ventilation, carbon dioxide, and ASHRAE Standard 62-1989. *ASHRAE J.* 7:40–44.

Reynolds, S. 1994. CFD modeling optimizes contaminant elimination. *Eng. Syst.* 2:35–37.

Rowland, S. 1992. The CFC controversy: issues and answers. *ASHRAE J.* 12:20–27.

Rudoy, W. and Duran, F. 1975. Development of an improved cooling load calculation method. *ASHRAE Trans.* II:19–69.

Scofield, C.M. and DesChamps, N.H. 1984. Indirect evaporative cooling using plate-type heat exchangers. *ASHRAE Trans.* I B:148–153.

Shinn, K.E. 1994. A specifier's guide to BACnet™. *ASHRAE J.* 4:54–58.

Sowell, E.F. 1988. Classification of 200,640 parametric zones for cooling load calculations. *ASHRAE Trans.* II:754–777.

Spitler, J.D., McQuiston, F.C., and Lindsey, K.L. 1993. The CLTD/SCL/CLF Cooling Calculation Method. *ASHRAE Trans.* I:183–192.

Straub, H.E. and Cooper, J.G. 1991. Space heating with ceiling diffusers. *Heating/Piping/Air Conditioning.* May:49–55.

Tackett, R.K. 1989. Case study: office building use ice storage, heat recovery, and cold air distribution. *ASHRAE Trans.* I:1113–1121.

Threlkeld, J.L. 1970. *Thermal Environmental Engineering.* Prentice-Hall, Englewood Cliffs, NJ.

The Trane Company. 1992. TRANE TRACE 600, Engineering Manual. The Trane Co., Lacrosse, WI.

Tinsley, W.E., Swindler, B., and Huggins, D.R. 1992. Rooftop HVAC system offers optimum energy efficiency. *ASHRAE J.* 3:24–28.

Tsal, R.J., Behls, H.F., and Mangel, R. 1988. T-method duct design. I. Optimizing theory. *ASHRAE Trans.* II:90–111.

Tsal, R.J., Behls, H.F., and Mangel, R. 1988. T-method duct design. II. Calculation procedure and economic analysis. *ASHRAE Trans.* II:112–150.

Vaculik, F. and Plett, E.G. 1993. Carbon dioxide concentration-based ventilation control. *ASHRAE Trans.* I:1536–1547.

Van Horn, M. 1986. *Understanding Expert Systems.* Bantam Books, Toronto.

Wang, S.K. 1993. *Handbook of Air Conditioning and Refrigeration.* McGraw-Hill, New York.

Wang, S.K., Leung, K.L., and Wong, W.K. 1984. Sizing a rectangular supply duct with transversal slots by using optimum cost and balanced total pressure principle. *ASHRAE Trans.* II A:414–429.

Williams, P.T., Baker, A.J., and Kelso, R.M. 1994. Numerical calculation of room air motion. III. Three-dimensional CFD simulation of a full scale experiment. *ASHRAE Trans.* I:549–564.

Wong, S.P.W. and Wang, S.K. 1990. Fundamentals of simultaneous heat and moisture transfer between the building envelope and the conditioned space air. *ASHRAE Trans.* II:73–83.

Wright, D.K. 1945. A new friction chart for round ducts. *ASHRA Trans.* 303–316.

9.16 Desiccant Dehumidification and Air-Conditioning

Zalman Lavan

Introduction

Desiccant air-conditioning is a promising emerging technology to supplement electrically driven vapor compression systems that rely almost exclusively on R22 refrigerant that causes depletion of the ozone layer. To date, this technology has only a limited market, e.g., in supermarkets where the latent heat loads are very high, in specialized manufacturing facilities that require very dry air, and in hospitals where maximum clean air is required. However, recent emphasis on increased air change requirements (see ASHRAE standards, ANSI 62-1989), improved indoor air quality, and restriction on use of CFC refrigerants (see The Montreal Protocol Agreement, as amended in Copenhagen in 1992, United Nations Environmental Programme, 1992) may stimulate wider penetration of desiccant-based air-conditioning which can be used as stand-alone systems or in combination with conventional systems. (See Table 9.4.1 for properties of some refrigerants.)

Sorbents and Desiccants

Sorbents are materials which attract and hold certain vapor or liquid substances. The process is referred to **absorption** if a chemical change takes place and as **adsorption** if no chemical change occurs. **Desiccants**, in both liquid and solid forms, are a subset of sorbents that have a high affinity to water molecules. Liquid desiccants *absorb* water molecules, while solid desiccants *adsorb* water molecules and hold them on their vast surfaces (specific surface areas are typically hundreds of square meters per gram).

While desiccants can sorb water in both liquid and vapor forms, the present discussion is limited to **sorption** of water vapor from adjacent air streams. The sorption driving force for both liquid and solid desiccants is a vapor pressure gradient. Adsorption (in solid desiccants) and absorption (in liquid desiccants) occur when the water vapor partial pressure of the surrounding air is larger than that at the desiccant surface. When an air stream is brought in contact with a desiccant, water vapor from the air is attracted by the desiccant, the air is **dehumidified**, and the water content of the desiccant rises. As the water sorbed by the desiccant increases, the sorption rate decreases and finally stops when *sorption equilibrium* is reached. For dehumidification to be resumed, water must be removed from the desiccant by heating. This process is referred to as **desorption**, **reactivation**, or **regeneration**. The heat of sorption (or desorption) is generally higher than the latent heat of vaporization of water; it approaches the latter as sorption equilibrium is reached.

Some typical *liquid desiccants* are water solutions of calcium chloride (CaCl), lithium chloride (LiCl), lithium bromide (LiBr), and triethylene glycol. The equilibrium water vapor pressure at the solution surface as a function of temperature and water content is shown in Figure 9.16.1 for water-lithium chloride solution. The surface vapor pressure (and dew point) increases with increasing solution temperature and decreases with increasing moisture content.

Common *solid desiccants* are silica gel, molecular sieves (zeolites), activated alumina, and activated carbon. The equilibrium sorption capacity (or moisture content) at a constant temperature, referred to as an **isotherm**, is usually presented as percent water (mass of water divided by mass of dry desiccant) vs. percent relative humidity (vapor pressure divided by saturation vapor pressure). Sorption capacity decreases with increasing temperature, but the spread of isotherms is relatively small (especially for concave down isotherms). Figure 9.16.2 shows normalized loading (sorption capacity divided by sorption capacity at 100% relative humidity) vs. relative humidity for silica gel, molecular sieve, and a generic desiccant, type 1 (modified) or simply 1-M (Collier et al., 1986).

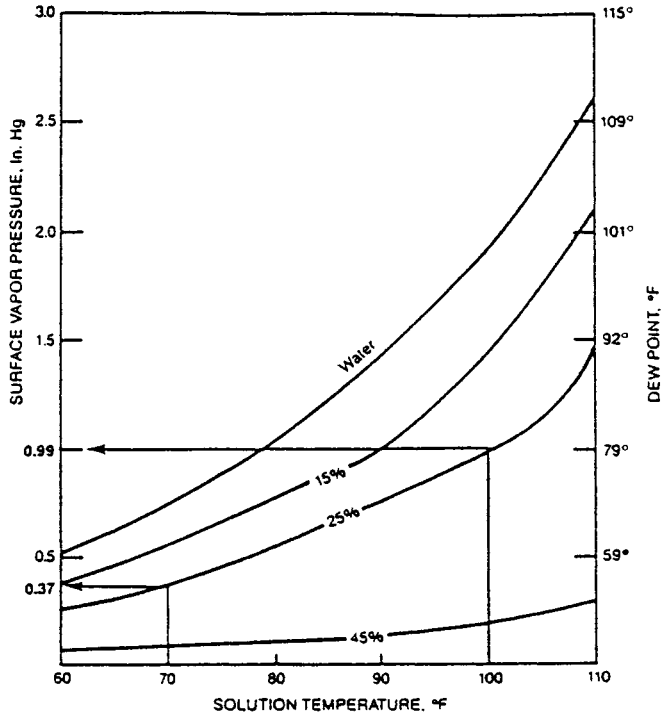


FIGURE 9.16.1 Surface vapor pressure of water-lithium chloride solutions. (Source: ASHRAE 1993, *Fundamentals Handbook*, chap. 19. With permission.)

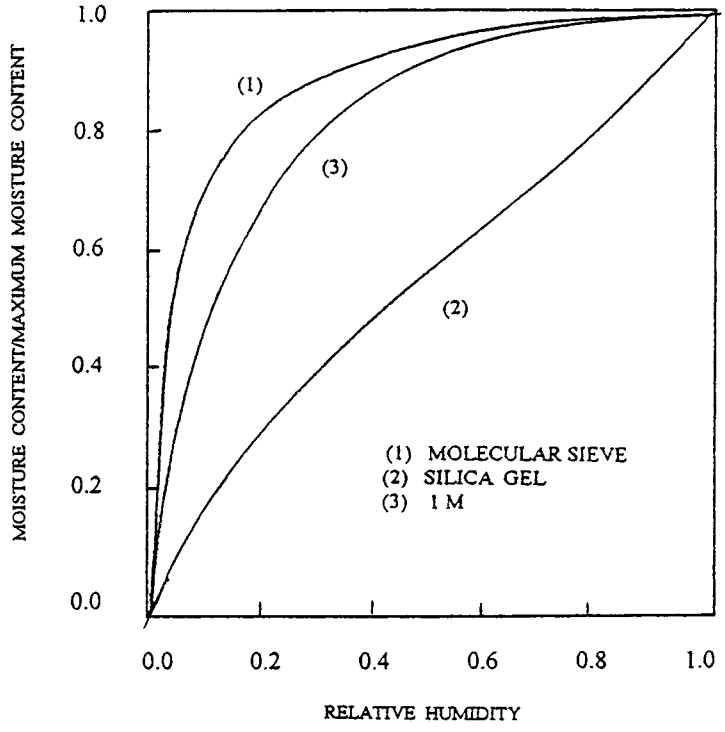


FIGURE 9.16.2 Normalized solid desiccant isotherms.

Dehumidification

Dehumidification by vapor compression systems is accomplished by cooling the air below the dew point and then reheating it. The performance is greatly hindered when the desired outlet dew point is below 40°F due to frost formation on the cooling coils (*ASHRAE, Systems and Equipment Handbook*, 1992).

Desiccant dehumidification is accomplished by direct exchange of water vapor between an air stream and a desiccant material due to water vapor pressure difference. Figure 9.16.3 shows the cyclic operation of a desiccant dehumidification system.

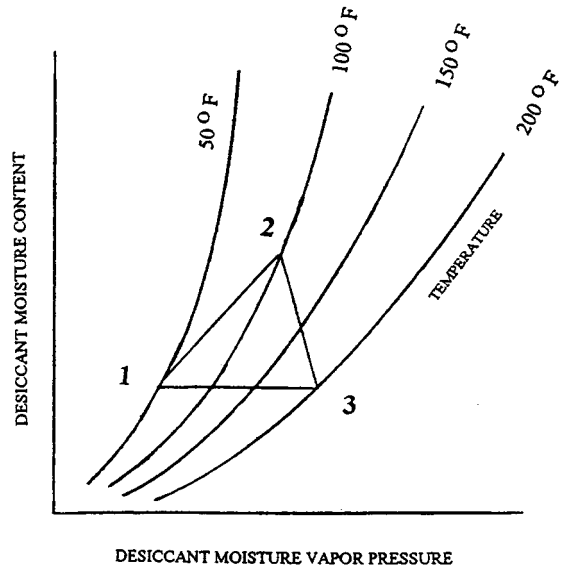


FIGURE 9.16.3 Cyclic dehumidification processes.

In *sorption (1–2)*, dry and cold desiccant (point 1) sorbs moisture since the vapor pressure at the surface is lower than that of the air stream. During this process the moisture content (loading or uptake) increases, the surface vapor pressure increases, and the liberated heat of sorption raises the desiccant temperature. During *desorption (2–3)*, the desiccant is subjected to a hot air stream, and moisture is removed and transferred to the surrounding air. The surface vapor pressure is increased and the desiccant temperature rises due to the added heat. The cycle is closed by *cooling (3–1)*. The desiccant is cooled while its moisture content is constant and the surface vapor pressure is lowered. The above cycle of sorption, desorption, and cooling can be modified by combining the sorption process with cooling to approach isothermal rather than adiabatic sorption.

Desirable Characteristics for High-Performance Liquid and Solid Desiccant Dehumidifiers

High equilibrium moisture sorption capacity
High heat and mass transfer rates
Low heat input for regeneration
Low pressure drop
Large contact transfer surface area per unit volume
Compatible desiccant/contact materials
Inexpensive materials and manufacturing techniques
Minimum deterioration and maintenance

Additional Requirements for Liquid Desiccant Dehumidifiers

Small liquid side resistance to moisture diffusion

Minimum crystallization

Additional Requirements for Solid Desiccant Dehumidifiers

The desiccant should not deliquesce even at 100% relative humidity.
The airflow channels should be uniform.
The desiccant should be bonded well to the matrix.
The material should not be carciogenic or combustible.

Liquid Spray Tower

Figure 9.16.4 is a schematic of a liquid spray tower. A desiccant solution from the sump is continuously sprayed downward in the absorber, while air, the process stream, moves upward. The air is dehumidified and the desiccant solution absorbs moisture and is weakened. In order to maintain the desired solution concentration, a fraction of the solution from the sump is passed through the regenerator, where it is heated by the heating coil and gives up moisture to the desorbing air stream. The strong, concentrated solution is then returned to the sump. The heat liberated in the absorber during dehumidification is removed by the cooling coil to facilitate continuous absorption (see Figures 9.16.1 and 9.16.3). The process air stream exits at a relatively low temperature. If sufficiently low water temperature is available (an underground well, for example), the process stream could provide both sensible and latent cooling.

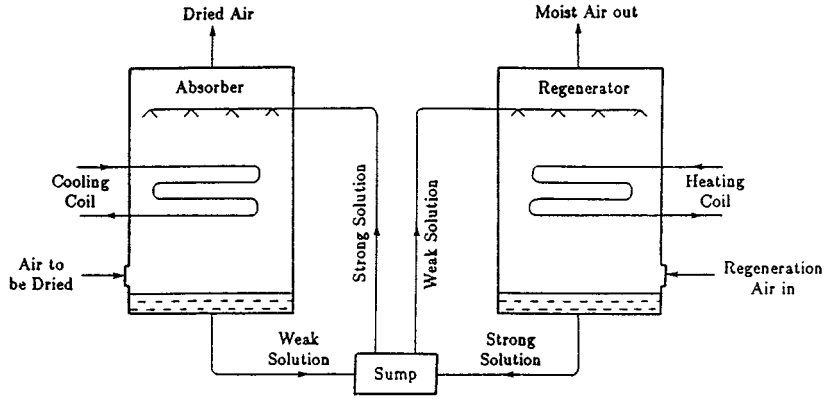


FIGURE 9.16.4 Liquid desiccant dehumidifier with heating and cooling coils.

The heating and cooling coils, shown in Figure 9.16.4, are often eliminated and the liquid solutions are passed through heating and cooling heat exchangers before entering the spray towers.

Advantages

The system is controlled to deliver the desired level of dry air by adjusting the solution concentration.

Uniform exit process stream conditions can be maintained.
A concentrated solution can be economically stored for subsequent drying use.
The system can serve as a humidifier when required by simply weakening the solution.
When used in conjunction with conventional A/C systems, humidity control is improved and energy is conserved.

Disadvantages

Some desiccants are corrosive.
Response time is relatively large.
Maintenance can be extensive.
Crystallization may be a problem.

Solid Packed Tower

The dehumidification system, shown in Figure 9.16.5, consists of two side-by-side cylindrical containers filled with solid desiccant and a heat exchanger acting as a desiccant cooler. The air stream to be processed is passed through dry desiccant in one of the containers, while a heated air stream is passed over the moist desiccant in the other. Adsorption (1–2) takes place in the first container, desorption (2–3) in the other container, and cooling (3–1) occurs in the desiccant cooler. The function of the two containers is periodically switched by redirecting the two air streams.

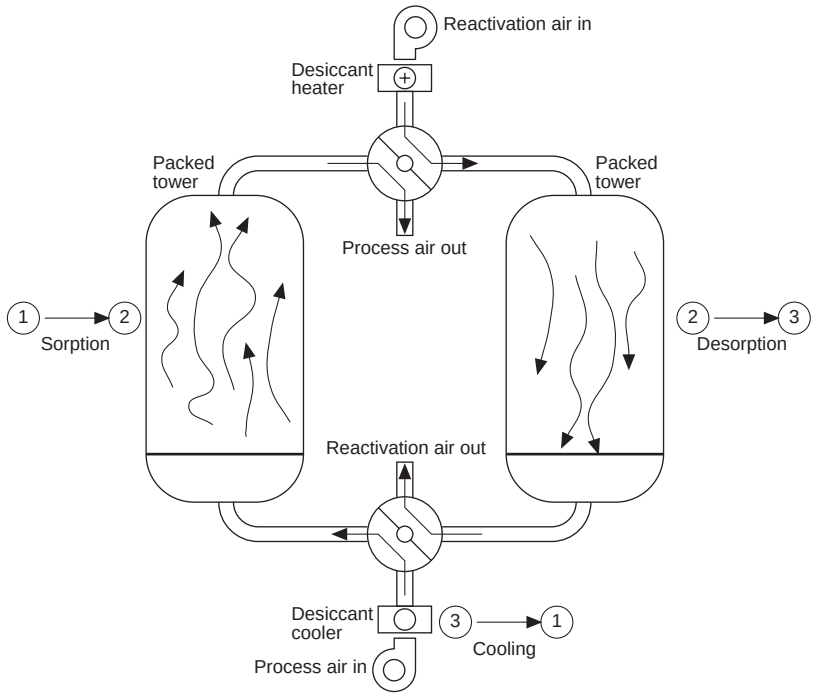


FIGURE 9.16.5 Solid packed tower dehumidification.

Advantages

No corrosion or crystallization
Low maintenance
Very low dew point can be achieved

Disadvantages

The air flow velocity must be low in order to maintain uniform velocity through the containers and to avoid dusting.
Uniform exit process stream dew point cannot be maintained due to changing moisture content in the adsorbing desiccant.

Rotary Desiccant Dehumidifiers

A typical rotary solid desiccant dehumidifier is shown in Figure 9.16.6. Unlike the intermittent operation of packed towers, rotary desiccant dehumidifiers use a wheel (or drum) that rotates continuously and delivers air at constant humidity levels.

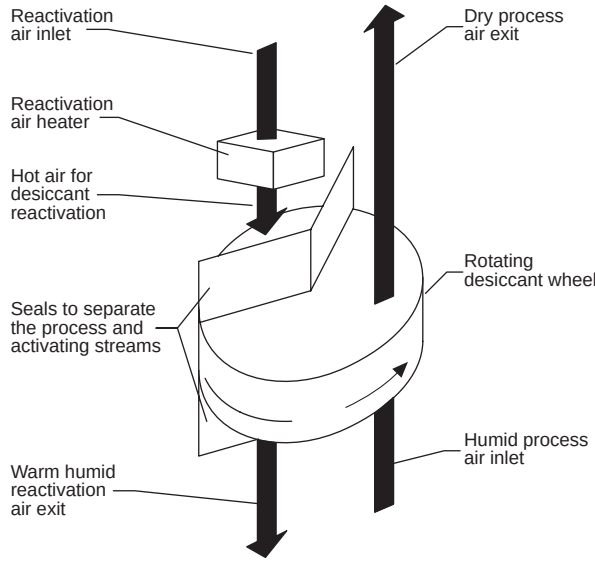


FIGURE 9.16.6 Rotary desiccant dehumidification wheel.

Desiccant wheels typically consist of very fine desiccant particles dispersed and impregnated with a fibrous or ceramic medium shaped like a honeycomb or fluted corrugated paper. The wheel is divided into two segments. The process stream flows through the channels in one segment, while the regenerating (or reactivating) stream flows through the other segment.

Desiccant Material

The desired desiccant properties for optimum dehumidification performance are a suitable isotherm shape and a large moisture sorption capacity. The isotherms of silica gel are almost linear. The moisture sorption capacity is high; the desiccant is reactivated at relatively low temperatures and is suitable for moderate dehumidification. Molecular sieves have very steep isotherms at low relative humidity. The desiccant is reactivated at relatively high temperatures and is used for deep dehumidification. The isotherm of the type 1-M yields optimum dehumidification performance (Collier et al., 1986), especially when used in conjunction with high regeneration temperatures.

The Desiccant Wheel

Some considerations for selection of desiccant wheels are:

Appropriate desiccant materials
Large desiccant content
Wheel depth and flute size (for large contact surface area and low pressure drop)
Size and cost

The actual performance depends on several additional factors that must be addressed. These include:

Inlet process air temperature and humidity
Desired exit process air humidity
Inlet reactivating air temperature and humidity
Face velocity of the two air streams
Size of reactivation segment

It should be noted that:

Higher inlet process air humidity results in higher exit humidity and temperature (more heat of sorption is released).
Lower face velocity of the process stream results in lower exit humidity and higher temperature.
Higher regeneration temperatures result in deeper drying, hence lower exit process air humidity and higher temperature.
When lower exit air temperature is required, the exit process air should be cooled by a heat exchanger.
Final cooling of the exit process air can be achieved by partial humidification (this counteracts in part previous dehumidification).

The following is a range of typical parameters for rotary desiccant wheels:

Rotation speed: 4 to 10 rpm
Desiccant fraction: 70 to 80%
Flute size: 1 to 2 mm
Reactivation segment: 25 to 30% of wheel
Face velocity: 300 to 700 fpm
Reactivating temperature: 100 to 300°F

Hybrid Cycles

A limited number of hybrid systems consisting of desiccant dehumidifiers and electrically driven vapor compression air-conditioners are presently in use in supermarkets. This application is uniquely suited for this purpose since the latent heat loads are high due to the large number of people and frequent traffic through doors. Also, low relative humidity air is advantageous for open-case displays.

Vapor compression systems are inefficient below a dew point of 45 to 50°F. When used in supermarkets, they require high airflow rates, the air must be reheated for comfort, and the evaporator coils must be defrosted frequently. Hybrid systems offer improved performance and lower energy cost in these cases.

Figure 9.16.7 shows a typical hybrid air-conditioning system for supermarkets. A mixture of outdoor and recirculated air is first passed through the desiccant and sensible heat exchanger wheels, where it is dehumidified and precooled. It then enters the conventional chiller before it is introduced to the interior of the supermarket. The sensible heat exchanger wheel is cooled by outdoor air and the desiccant wheel is regenerated by air heated with natural gas. Energy cost can be further reduced by preheating the reactivating air stream with waste heat rejected from the condenser of the refrigeration and/or air-conditioning systems.

The advantages of these hybrid systems are

Air-conditioning requirement is reduced by up to 20%.
The vapor compression system operates at a higher coefficient of performance (COP) since the evaporator coils are at a higher temperature.
Airflow requirements are reduced; electric fan energy is saved and duct sizes are reduced.
The refrigeration cases run more efficiently since the frequency of defrost cycles is greatly reduced.

Solid Desiccant Air-Conditioning

Several stand-alone desiccant air-conditioning systems were suggested and extensively studied. These systems consist of a desiccant wheel, a sensible heat exchanger wheel, and evaporating pads. Sorption can be adiabatic or cooled (if cooling is combined with sorption). When room air is dehumidified and recirculated, the system is said to operate in the **recirculating** mode. When 100% outside air is used as the process stream, the system operates in the **ventilating mode**.

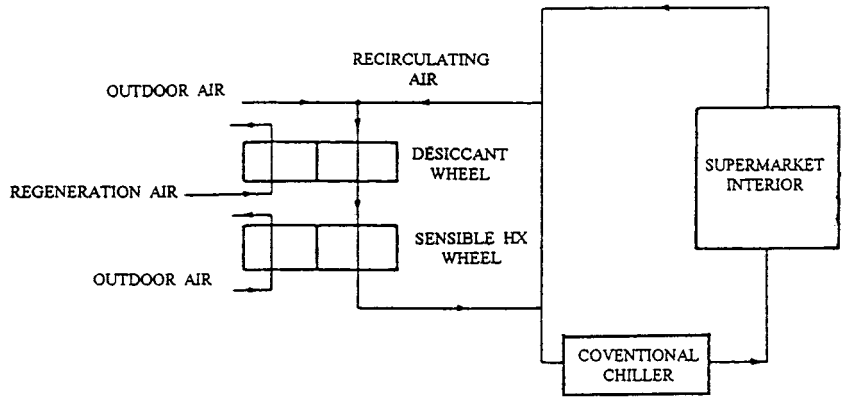


FIGURE 9.16.7 Hybrid air-conditioning system for supermarkets.

Ventilation Mode

In the *adsorption path* the process air stream drawn from the outdoors is passed through the dry section of the desiccant wheel where it is dehumidified and heated by the liberated heat of sorption. It then passes through the sensible heat exchanger wheel and exits as dry but slightly warm air. The hot and dry air leaving the dehumidifier enters the heat exchanger, where it is sensibly cooled down to near room temperature. It is then passed through the evaporative cooler, where it is further cooled and slightly humidified as it enters the conditioned space.

In the *desorption path*, air is drawn from the conditioned space; it is humidified (and thus cooled) in the evaporative cooler. The air stream enters the sensible heat exchanger, where it is preheated, and it is then heated to the desired regeneration temperature by a suitable heat source (natural gas, waste heat, or solar energy), passed through the desiccant wheel (regenerating the desiccant material), and discharged out of doors.

Performance. In order to achieve high performance, the maximum moisture content of the desiccant should be high and the isotherm should have the optimum shape (1 M). In addition, Zheng et al. (1993) showed that the optimum performance is very sensitive to the rotational speed of the desiccant wheel. Glav (1966) introduced stage regeneration. He showed that performance is improved when the reactivation segment of the wheel is at a temperature which increases in the direction of rotation. Collier (Collier et al., 1986) showed that well-designed open-cycle desiccant cooling systems can have a thermal COP of 1.3. This, however, would require the use of high-effectiveness sensible heat exchangers, which would be large and expensive. Smaller and more affordable heat exchangers should yield system COPs in the order of unity. An extensive review of the state-of-the-art assessment of desiccant cooling is given by Pesaran et al. (1992).

Conclusions

Desiccant-based air-conditioning offers significant advantages over conventional systems. Desiccant systems are already successfully used in some supermarkets. It is expected that these systems will gradually attain wider market penetration due to environmental requirements and potential energy savings.

The advantages of desiccant air-conditioning are summarized below:

No CFC refrigerants are used.
Indoor air quality is improved.
Large latent heat loads and dry air requirements are conveniently handled.
Individual control of temperature and humidity is possible.

The energy source may be natural gas and/or waste heat.
Less circulated air is required.
Summer electric peak is reduced.

Defining Terms

Absorb, absorption: When a chemical change takes place during sorption.
Adsorb, adsorption: When no chemical change occurs during sorption.
Dehumidification: Process of removing water vapor from air.
Desiccant: A subset of sorbents that has a particular affinity to water.
Desorb, desorption: Process of removing the sorbed material from the sorbent.
Isotherm: Sorbed material vs. relative humidity at a constant temperature.
Reactivation: Process of removing the sorbed material from the sorbent.
Recirculation: Indoor air only is continuously processed.
Regeneration: Process of removing the sorbed material from the sorbent.
Sorbent: A material that attracts and holds other gases or liquids.
Sorption: Binding of one substance to another.
Staged regeneration: When the temperature of the regeneration segment of the desiccant wheel is not uniform.
Ventilation mode: 100% of outdoor air is processed.

References

ANSI/ASHRAE 62-1989. *Ventilation for Acceptable Indoor Air Quality.* American Society of Heating, Refrigeration and Air-Conditioning Engineers, Atlanta, GA.

ASHRAE. 1992. *HVAC Systems and Equipment Handbook*, Chap. 22. American Society of Heating, Refrigeration and Air Conditioning Engineers, Atlanta, GA.

ASHRAE. 1993. *Fundamentals Handbook*, Chap. 19. American Society of Heating, Refrigeration and Air Conditioning Engineers, Atlanta, GA.

Collier, R.K. 1989. Desiccant properties and their effect on cooling system performance. *ASHRAE Trans.* 95(1):823–827.

Collier, R.K, Cale, T.S., and Lavan, Z. 1986. *Advanced Desiccant Materials Assessment*, pb-87-172805/XAB. Gas Research Institute, Chicago, IL.

Glav, B.O. 1966. Air Conditioning Apparatus, U.S. Patent No. 3251402.

Harriman, L.G. III. 1990. *The Dehumidification Handbook Second Edition.* Munters Cargocaire, Amesbury, MA.

Pesaran, A.A., Penny, T.R., and Czanderna. 1992. *Desiccant Cooling: State-of-the-Art Assessment.* National Renewable Energy Laboratory, Golden, CO.

United Nations Environmental Programme. 1992. Report of the fourth meeting of the parties to the Montreal protocol on substances that deplete the ozone layer, November 23–25, 1992, Copenhagen.

Zheng, W., Worek, W.M., and Novosel, D. 1993. Control and optimization of rotational speeds for rotary dehumidifiers. *ASHRAE Trans.* 99(1).

Cluff, K.D., et. al. “Electronic Packaging Technologies”
Mechanical Engineering Handbook
Ed. Frank Kreith
Boca Raton: CRC Press LLC, 1999

10A.1 Electronic Packaging Technologies

Kevin D. Cluff
Advanced Packaging Technology
Honeywell Aerospace
Electronic Systems
Phoenix, Arizona

Michael G. Pecht
CALCE Electronic Products and Systems Center
University of Maryland
College Park, Maryland

Electronic packaging is the art and science of connecting circuitry to reliably perform some desired function in some application environment. Packaging also provides ease of handling and protection for assembly operations. This chapter defines chip, or die-level, and assembly-level packaging, with an emphasis on recent technologies. The relative advantages of each technology will be discussed in relation to performance, cost, reliability, and manufacturability. Extensive details on electronic packaging can be found in the references.[1–6]

10A.1.1 Packaging the Die

A semiconductor device, also known as a *die* or *chip*, is fragile and must be packaged for protection and for interfacing with the outside world. The chip package provides an electrical interconnection to the assembly, module, or display and protects the chip in the manufacturing and application environments. A number of different materials can be used in the die package, including ceramics, plastics, and metals. Single-chip, three-dimensional, and multichip module (MCM) packages are some of the packaging formats. Interconnection is the process and technique of making electrical connections between the bond pads of the chip and a leadframe, substrate, or even another chip.

Plastic Die Package

Plastic packages are very popular in commercial applications because of their low cost and small size compared to ceramic packages. The cost of plastic packages is typically one half to one tenth the cost of comparable ceramic or metal packages. More than 98% of all integrated circuits were packaged in plastic in 1992. In a plastic package, the chip is encapsulated by a polymer, usually referred to as the *encapsulant*. An encapsulant is generally an electrically insulating material formulation that protects an electronic device and die-leadframe assembly from the adverse effects of handling, storage, and operation.

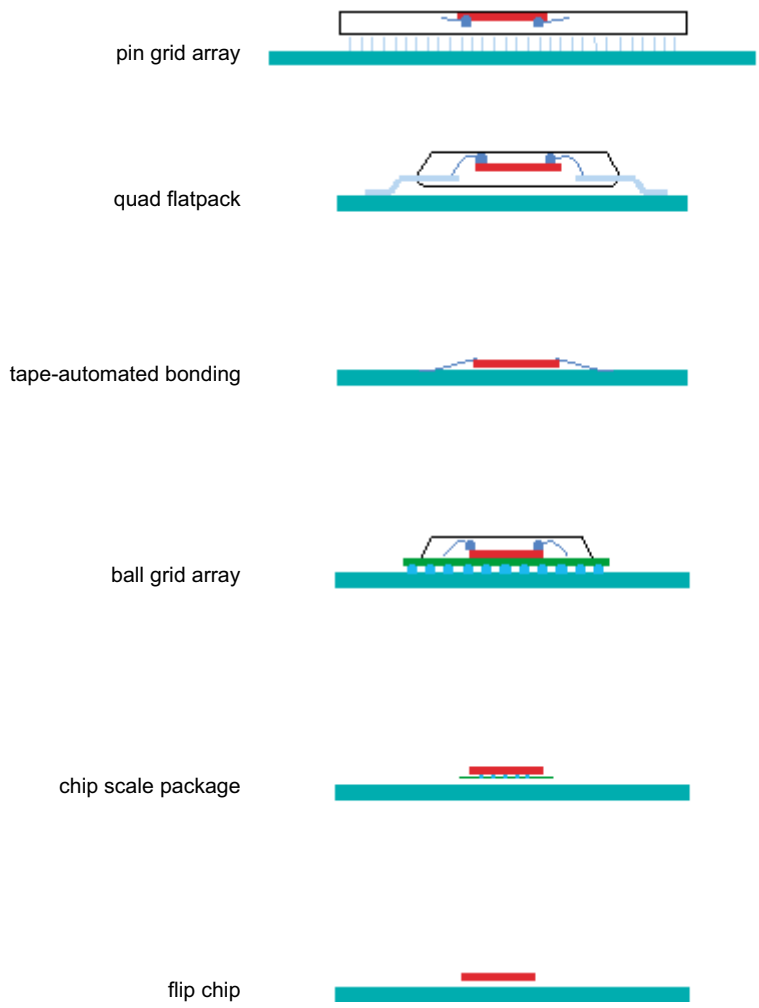


FIGURE 10A.1.1 Shrinking size of packaging.

Figure 10A.1.1 depicts the evolution of packaging shrinkage through the last three decades. Packaging size is being pressed by the shrinking feature size (approximately 15% per year) of CMOS integrated circuits. Package lead styles have changed dramatically from the 1970s when the dual in-line package (DIP) and pin grid array dominated the semiconductor industry. Even today, the DIP retains a sizable share of the market in which low-cost, rugged packaging meets the application need. Generally, there are at least three major lead types, with numerous variations of each: the through-hole, the surface-mount (J- or gull-lead), and the solder ball. The through-hole and leaded surface-mount packages usually have a leadframe to which the die is mounted. The plastic solder ball packages typically employ organic substrates that redistribute the die inputs-outputs (I/O) to an array or perimeter pattern of solder balls. (Substrate technologies are addressed in Section 10A.1.2.)

In the late 1980s, 1.27-mm pitch, surface-mount technology became prominent. As device complexity increased, the I/O count also increased. To keep electrical performance and package sizes reasonable, the lead pitch decreased, and fine-pitch leaded devices (lead pitch of less than or equal to 0.65 mm) were introduced.

Array packaging was first introduced in the mid-1970s with ceramic pin grid array packages. Surface-mount array packages evolved to plastic ball grid array (PBGA) and chip-scale packages (CSP). Solder

TABLE 10A.1.1 Typical Coplanarity Requirements for SMT Components[7]

Component	Pitch (mm)	Coplanarity Requirement (mm)
Leaded SMT	—	0.10
PBGA	1.27 –1.5	0.20
	1.0	0.15
	<1.0	0.08
CBGA	—	0.15
CSP	—	0.08

From Hwang, J.S., BGA and CSP solder spheres, *Surface Mount Technology,* April, p. 18, 1999. © 1999 by IHS Publishing Group (www.smtmag.com). With permission.

spheres — usually composed of 63% tin and 37% lead, 62/36/2 (% silver), or high-temperature 10/90 — form the connection between the package and the board.

The chip-scale package is typically defined as being not more than 20% larger than the die. The CSP interposer buffers the final assembly from die-shrink redesign. The interposer can be organic substrate, tape-automated bonded (TAB), or ceramic substrate. The interposer can also act as a stress buffer between the low thermal expansion of the semiconductor die and the higher expansion of the substrate. Tessera (San Jose, CA) adds a patented elastomer between the die and the interposer to further buffer the solder balls. While the majority of CSP used today have solder-ball interconnects, there are non-solder-ball variations that meet the CSP definition, such as the small-outline, no-lead (SON) and bottom-leaded plastic packages (BLP).

The coplanarity of surface-mount parts is an important mechanical parameter in assembly. Non-planar parts can lead to poor assembly yields (voids, electrical opens) and unreliable connections (misregistration). Table 10A.1.1 outlines typical coplanarity requirements.

Encapsulation techniques include molding, potting, glob-topping, and conformal coating, but the majority of encapsulated packages use the molding process. The molding compound is a proprietary multicomponent mixture of an encapsulating resin with various types of additives. The principal active and passive (inert) components in a molding compound include curing agents or hardeners, accelerators, inert fillers, coupling agents, flame retardants, stress-relief additives, coloring agents, and mold-release agents. Molding compounds have moved steadily toward all-epoxy systems.

Thermosetting encapsulation compounds are based on epoxy resins. In electrical and electronic applications, three types of epoxy resins are commonly used: the diglycidyl ethers of bisphenol A (DGEBA) or bisphenol F (DGEBF), the phenolic and cresol novolacs, and the cycloaliphatic epoxides. Novolac-based epoxies are supplied as molding bricks (preforms) that are ready for use in transfer-molding machines to produce single-in-line packages, dual-in-line packages, plastic leaded chip carriers, quad flatpacks, various types of small-outline integrated circuits (SOIC), and ball grid array packages.

Pin grid array carriers, some ball grid array packages, and carriers with metal cans frequently use multicomponent liquid epoxies or preforms. Flip chips mounted on organic substrates also use liquid epoxy underfills to protect the die surface environmentally as well as to provide stress relief to the solder balls. Polyurethanes, polyamides, and polyesters are used to encase modules and hybrids intended for use under low-temperature and high-humidity conditions. Modified polyimide encapsulants have the advantages of thermal and moisture stability, low coefficient of thermal expansion, and high material purity. Thermoplastics are rarely used because they generally require processing conditions of unacceptably high temperature and pressure, are low-purity materials, and can induce high moisture-induced stresses.

Several molding techniques (Table 10A.1.2) are available for the manufacture of plastic cases. The most widely used manufacturing process is the transfer molding process, used for molding thermosetting polymers. These polymers are plastic or fluid at low temperatures but, when heated, react irreversibly to

TABLE 10A.1.2 Comparison of Molding Processes

Molding type	Advantages	Disadvantages
Transfer molding	Multiple cavities, high yield Relative material savings Short cycle time Low tool maintenance costs	High molding pressure High viscosity Restricted to the packaging of leadframes
Injection molding	Good surface finish Good dimensional control	Poor material availability
Reaction-injection molding	Energy efficiency Low mold pressure Good wetting of chip surface Adaptability to TAB	Few resin systems available for electronic packaging Requires good mixing

form a cross-linked network no longer capable of being melted. Other processes include the injection molding process and the reaction-injection molding process.

Hermetic (Metal and Ceramic) Packages

Hermetic packages have been used predominantly for military and government applications because of their perceived reliability advantage over plastic packages. They have also been employed in some high-power applications when heat must be dissipated from the device. Metal packages are typically used for small integrated circuits with a low lead count and in applications that require electromagnetic shielding.

Both metal and ceramic packages can be made nearly impervious to moisture when hermetically sealed. A package is classified as hermetic if it has a minimal leak rate (the rate at which gases can diffuse into or out of the package). Typical acceptable helium leak rates depend on the package size and helium pressure in the package. Some metal packages (e.g., Olin Metal quad flatpacks) are sealed using epoxies and, therefore, are not hermetic.

Several major semiconductor companies (Intel, Motorola, and AMD) have announced their exit from military markets, and the availability of these packages is shrinking dramatically. Plastic components are finding their way into many applications that had once been dominated by hermetic packages.

Interconnection Technologies at the First Level of Packaging

The major interconnection technologies at the first level of packaging are wirebonding, flip chip, and TAB. Wirebonding, the mainstay of the microelectronics industry, still accounts for more than 98% of all semiconductor interconnections. Invented more than 30 years ago, flip chip was once available only to large, vertically integrated companies. Now, contract sources of wafer bumping are available, making flip chip one of the fastest growing interconnect methods. Tape-automated bonding (TAB) techniques are mature, although they are utilized only in limited, specialized applications because of high tooling costs. Chip stacking or three-dimensional (3-D) interconnection is typically used in systems where size, weight, and speed are critical. Comparisons of these technologies are provided in Tables 10A.1.3 and 10A.1.4.

Wirebond Interconnects

Wirebond interconnects are formed by metallurgically bonding a small-diameter wire from the semiconductor device to the leadframe, substrate, or other semiconductor. The wirebond forms a low-resistance path for signal propagation. Typical wire materials include aluminum, gold, and copper. Gold is usually bonded by thermocompression, while aluminum is usually bonded ultrasonically.

Common materials for pads on bondable surfaces include aluminum, gold, silver, nickel, and copper. Aluminum and gold are the most frequently used bondpad materials. Silver has been used as a bondable plating material on leadframes and as a bondable thick-film metallization in alloy form with platinum

TABLE 10A.1.3 Comparison of Interconnect Technologies

Attribute	Wirebonding	TAB	Flip TAB	Flip Chip	High-Density Interconnects
Minimum (and typical) I/O pitch (μm)	100 (150)	50 (100)	50 (100)	125 (250)	25/1 (50/2)
Maximum I/O range	256–500	400–700	400–700	>800	>1000
Lead inductance (nH)	1–2	1	0.1	0.05–0.1	<0.05
Mutual inductance between leads (pH)	100	5	5	1	<1
Typical effective diameter (μm)	25	50	50	125	N/A
Typical length (mm)	1	1	0.25	0.1	N/A
Connection technique	Perimeter; array area is difficult but possible, using multi-height loops	Perimeter; array area is moderately difficult but possible, using multilayer tape	Perimeter; array area is moderately difficult	Perimeter and array area	Perimeter and array area
Packaging efficiency	Medium	Medium	Medium-high	High	High
Pretestability in fine pitch	Difficult with available instrumentation	Good	Good	Difficult with available instrumentation	Difficult with available instrumentation
Ability to rework	Difficult	Fair	Good	Good	Good
Loop control	Fair	Good	Good	Good	Good
Flexibility of the manufacturing process	Excellent	Fair (gang bonding); good (single-point bonding)	Fair (gang bonding); good (single-point bonding)	Good	Good
Dominant failure mechanisms	Fatigue, bond-pad corrosion	Lead fatigue, interdiffusion	Lead fatigue, interdiffusion	Fatigue, intermetallics	Electromigration, corrosion
Heat dissipation	Good (die bonded to substrate)	Good (die bonded to substrate)	Poor, but excellent if heat sink attached to back side	Poor, but excellent if heat sink attached to back side	Good (die bonded to substrate)
Die availability	Excellent	Fair	Fair	Poor	Excellent
Tool availability	Excellent	Fair	Fair	Fair	Fair
Technology maturity	Excellent	Good	Good	Good	Fair
Market share	98%	<2%	<1%	<1%	<1%
Cost	Low	Medium	Medium	High (potentially low)	High (potentially low)

TABLE A10.1.4 Relative Advantages of Interconnect Technologies

	Advantages	Disadvantages
Wirebond	Low resistance interconnects Low capital investment Mature technology Highly stable manufacturing process	Potential for cross-talk High lead inductance Use of costly, precious metal wire Low interconnect density Limited chip size
Flip chip	Self-alignment during die joining Automatic die placement and testing Low lead inductance Less need for precious metals Simultaneous bonding	Potential for solder fatigue Thermal dissipation concerns
TAB	Lightweight packaging Small size/high I/O-count Simultaneous bonding	High capital investment
HDI	Power and signal distribution High frequency performance High clock-rate performance High wiring densities High device-to-substrate area ratio	Process not fully mature High cost of deposited HDI
3-D	Increased package density Better board utilization Increased electrical performance Shorter interchip connections Lower system power	Potential for cross-talk Thermal dissipation concerns Poor yields Process not fully mature

or palladium.[8] Nickel has been widely used as a substitute for gold in power devices. In addition, nickel or chromium serve as barrier layers to avoid unwanted intermetallic compounds and to facilitate adhesion.

Thermocompression and thermosonic bonds are typically used to produce a ball bond on the first bond and a wedge bond on the second. Thermocompression is used on gold and in plastic packages to prevent wire sweep. Thermocompression bonds occur when two metal surfaces are brought into intimate contact through a controlled time, temperature, and pressure cycle. Typical bonding temperatures during thermocompression bonding range from 300 to 400°C. Heat is generated during the manufacturing process, either by a heated pedestal on which the assembly is placed for bonding or by a heated capillary feeding the wire. Typically, bonding time is in the neighborhood of 40 ms, which translates to a machine speed of approximately two wires per second.

Thermosonic bonding combines ultrasonic energy with the ball-bonding capillary technique of thermocompression bonding. The processes are similar, but in thermosonic bonding the capillary is not heated and substrate temperatures are maintained between 100 and 150°C. Ultrasonic bursts of energy make the bond. Typically, at the second bond, the capillary leaves a characteristic circular pattern called a *crescent bond*. The bonding time is usually about 20 ms, or a machine speed of ten wires per second, for ball bonding and three wires per second for wedge bonding.

Ultrasonic bonding is a low-temperature process in which the source of energy for the metal welding is a transducer vibrating the bonding tool parallel to the bonding pad in a frequency range from 20 to 60 kHz. Ultrasonic bonding uses aluminum wire in ceramic/metal packages. Ultrasonic processes are used for wedge-wedge bonds, for which a common centerline and careful alignment during manufacture are critical to avoid large stresses and irregular deformation. Normally, the first bond is made to the die and the second to the substrate. Referred to as *forward bonding*, this is the preferred bonding procedure because it is less susceptible to edge shorts between the die and the wire. Typical bonding time for

ultrasonic bonding is about 20 ms, which translates to a machine speed of four wires per second. A disadvantage of wedge bonding over ball bonding is the necessity of maintaining the directional first-to-second alignment by rotating and aligning the die and substrate with the direction of the wire. Rotation is difficult in large MCMs (>7 cm on the side), as most equipment was designed for much smaller integrated circuit (IC) bonding. Thus, while ultrasonic bonding has the advantage of small bond pitch, the overall speed of the process is less than that for thermosonic bonds, due to machine rotational movements.

Some quality/infant-mortality failure mechanisms associated with wirebonding include cratering of the die or chip-out due to excessive bond stress on the die. It is important to maintain adequate bonding load and temperature to allow dispersion of contaminants. Cleanliness, proper atmosphere during bonding, and surface finish are also important assembly factors. Long-term failure mechanisms include differential thermal fatigue, electrical leakage, and Kirkendall voids — voids formed by different interdiffusion rates of gold and aluminum. Purple plague is a brittle, gold-aluminum intermetallic formed at high temperatures.

Flip Chip Bonding

Flip chip bonding is an interconnection technique in which the die is connected facedown onto the substrate. Solder bumps on area-array metallized terminals are reflowed to matching footprints of pads on a substrate. Flip chip bonding grew at a rate of about 18% compound annual growth (CAG) between 1991 and 1997. With wafer bumping services readily available, growth from now to the year 2002 is projected at nearly 50% CAG. One of the first products with extensive use of flip chip packaging was a Sony camcorder that contained sixteen flip chip die on chip-scale interposers.[9]

The three major steps involved in manufacturing flip chip bonds are (1) die or wafer bumping, (2) alignment of the die and substrate, and (3) assembly. Flip chip technology has several advantages over wirebonding:

- Self-alignment during die joining, which allows for automatic die placement and testing
- Low lead inductances, due to shorter interconnection lengths than wirebonding
- Reduced need for precious attach metals
- Increased productivity, due to the ability to make large numbers of bonds simultaneously

Evaporating and plating are the major methods of bump formation. Evaporated bumps (125 μm diameter and 100 μm high) can be smaller and more uniform than plated bumps (125–175 μm diameter and 25–100 μm high).

The flip chip assembly of a microelectronic component package is a composite of multiple, thin layers of metals. On the die bond pads, the pad structure is called *ball-limiting metallurgy* (BLM), *pad-limiting metallurgy* (PLM), or *under-bump metallurgy* (UBM). On the substrate bond pads, the connection structure is called *top-surface metallurgy* (TSM). Figure 10A1.2 shows a typical flip chip bond structure.

The BLM structure and solder bump manufacturing, called the *bumping process*, can be implemented using a variety of methods, including metal masking, photolithography, electroplating and ultrasonic soldering, maskless bumping, and copper bumping. Critical properties of the BLM adhesive layer include good electrical contact with the metallization pad on the die, strong adhesion to the pad and to the passivation layer surrounding it, and selective etchability of the metal layer, enabling the use of photolithography techniques during fabrication. The BLM barrier layer must also have good solderability and the capability to prevent interdiffusion between the solder and the pad metallization. Finally, the BLM bonding layer must provide an inert surface during bonding and protect the barrier layer from oxidation during storage.

The barrier layer of the TSM structure is made of a metal that increases wettability and has a surface suitable for solder reflow. The bonding layer uses metals that have the ability to retain wetting properties and provide adequate shelf life prior to die attachment. The melting point is the most important material property when selecting solder materials. High-melting-point solders are selected for die-level

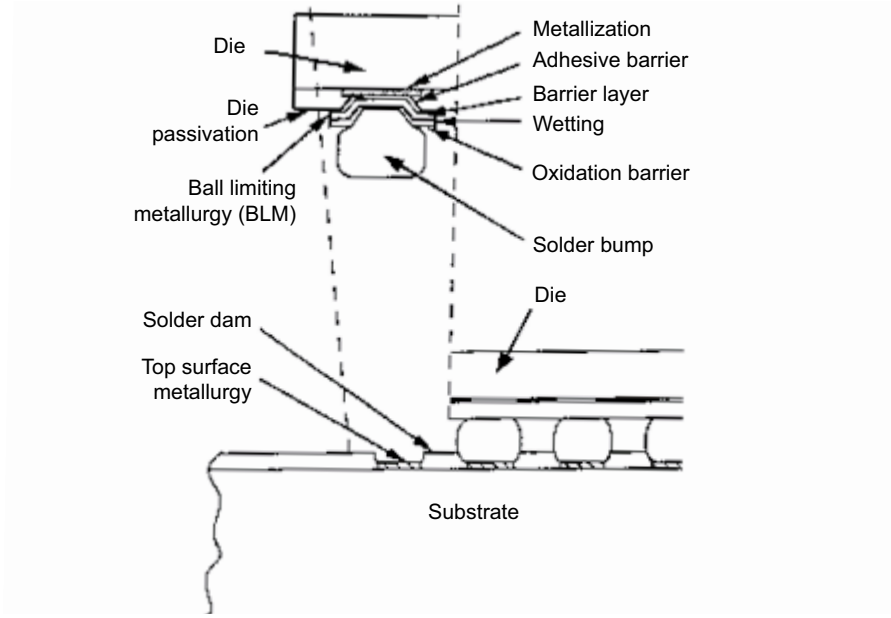


FIGURE 10A.1.2 Flip chip bond structure. (From Pecht, M. et al., *Quality Conformance and Qualification of Microelectronic Packages and Interconnects,* John Wiley & Sons, New York. With permission.)

interconnections to enable lower melting-point solders to be used for the board-level packaging. For example, 95Pb/5Sn solder (melting point around 315°C) is used with alumina ceramic substrates for die connections, while eutectic tin lead solder (melting point around 183°C) is used to attach dies to such organic substrates as Kapton film and glass/epoxy printed wiring boards.

Also important in solder selection is the cyclic strain hardening behavior of the solder, which describes the strain induced in the solder by a given stress condition. This material property strongly influences the thermal fatigue life of the solder. A slow interdiffusion rate that reduces intermetallic formation with the metal layers on the die and the substrate pads is required to enhance the thermal fatigue life of the solder. Intermetallics, though necessary to form a good bond, may reduce the thermal fatigue life if present in excessive quantities because of their brittle nature.

High corrosion resistance is required for solders used in non-hermetic packages to prevent corrosion-related failures. High oxidation resistance is also desirable, as the operational environment in non-hermetic packages accelerates solder oxidation, first on the surface and then in the bulk solder, which can eventually result in solder-die separation. High thermal conductivity is a benefit of metallic solders that allows the heat generated by the die to be dissipated to the substrate via solder bumps.

Tape-Automated Bonding (TAB)

Tape-automated bonding is particularly well suited for volume applications requiring small, lightweight, and high-I/O-count electronic packages. Furthermore, with hybrids and multichip modules, TAB has been used as a means to test and package very large-scale integration dies for a variety of products. (For TAB design guidelines and comparisons with other interconnect technologies, the interested reader is referred to Reference 2.)

The TAB assembly includes bumps, interface metallurgy, leads, lead plating, adhesive, polymeric tape, and encapsulant. In one-layer tape (all metal), the tape is made of metal foil with an etched lead pattern. In multiple-layer tape, the tape assembly consists of metal foil with an etched lead pattern and polymeric tape, with or without adhesive. A TAB assembly can be mounted onto the substrate face-up (conventional), facedown (flip-TAB), or in recessed pockets. Because of the geometry configuration, flip-TAB mounting offers options for either straight or formed leads.

TABLE 10A1.5 Dimensions of Tape-Automated Bonds[11–13]

Element of TAB	Dimensions
Bump	
Square	50–120 μm wide and 10–30 μm high
Truncated sphere	80- to 100-μm diameter and 25 μm high
Interface metallurgy	
Adhesion and barrier layers	0.08–1 μm
Bonding layer	0.1–0.3 μm
Lead	
Thickness	18–70 μm
Inner lead pitch	30–100 μm
Outer lead pitch	50–635 μm
Plating thickness	0.25–2 μm
Polymeric tape	
Width	8–70 mm
Thickness	25–127 μm
Polymeric adhesive thickness	12.7–25.4 μm
Die passivation thickness	1–1.5 μm

Common TAB dimensions are summarized in Table 10A.1.5. The bump configurations in TAB can be classified as bumped die, bumped leads (BTAB), transferred bumps (TTAB), or bumpless. The bumped-die approach requires bumping the die before bonding. One-layer tape typically consists of about 70-μm-thick etched copper foil. Two-layer tape consists of a polymeric film, approximately 50 μm thick, and a patterned metal layer of 20 to 40 μm. Three-layer tape consists of patterned metal, usually as sheets of electro-deposited or rolled copper (typically 35 μm thick), laminated with an adhesive to prepunched polymeric film (typically 125 μm thick). Two-metal-layer tapes with signal and ground planes and three-metal-layer tapes that provide signal, ground, and power planes are being used in some multichip modules.[10]

Ultrasonic bonding is typically used to create the inner-lead bond. The outer-lead bond can be ultrasonically bonded, adhesive bonded, or soldered to the board or substrate. The BTAB approach bonds an unbumped die to bumps on the tips of the inner leads. The TTAB approach bonds an unbumped die to bumps on a separate substrate, which are transferred to the inner leads. The bumpless TAB approach bonds the TAB inner leads directly to an unbumped die.

Tape-automated bonding is capable of accommodating dies with peripheral and area-array I/Os. In an area-array TAB (ATAB), a multilayer metal tape is usually used to provide interconnections between the die bumps and the substrate bonding pads. Area-array TAB increases the number of potential I/Os and improves integrated circuit design flexibility. The common inner lead pitch is 100 μm, with 50 μm wide lines and spacing.

The typical outer-lead pitch ranges from 200 to 625 μm. Tape widths can range from 8 to 70 mm to accommodate various die sizes and I/O counts. Multiple-conductor (multimetal) tapes have been developed to control impedance and reduce cross-talk.[10] An advantage of TAB is the ability to test the components at extremely fine pitches, prior to excision of the part from the tape. This practice prevents lead damage. TAB pitches of less than 50 μm are in production and perform electrically and thermally better than wirebonding.

Some disadvantages of TAB are tooling and nonrecurring costs. Each die design requires a new tape design. This reduces the flexibility to accommodate die shrinks.

Three-Dimensional Die Packaging

Three-dimensional (3-D) packaging and interconnection provide increased electrical performance and packaging density compared to traditional two-dimensional (2-D) approaches. Three-dimensional

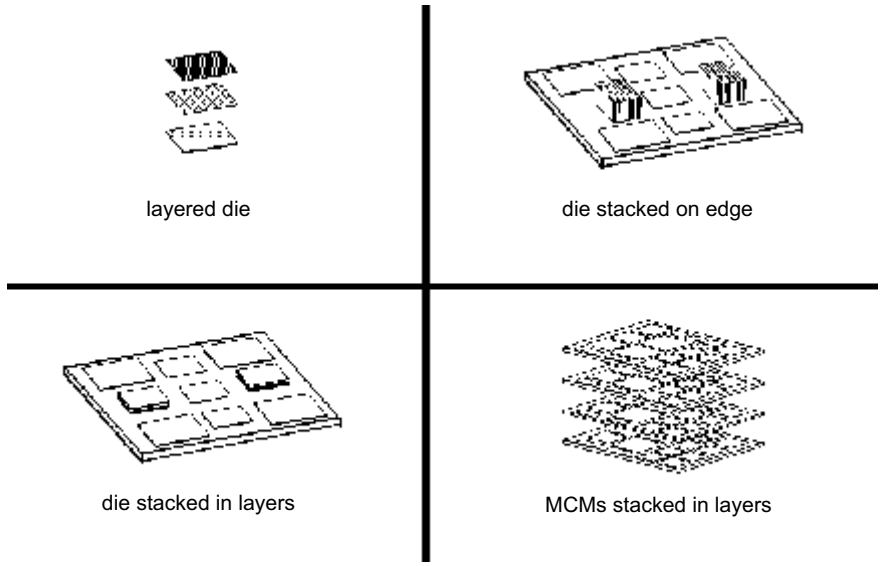


FIGURE 10A.1.3 Types of 3-D packaging. (From Pecht, M. et al., *Quality Conformance and Qualification of Microelectronic Packages and Interconnects,* John Wiley & Sons, New York. With permission.)

packaging and interconnection can increase density by a factor of more than 50 by stacking integrated circuits (ICs). Schematically illustrated in Figure 10A1.3 are four generic types of 3-D packaging: layered dies, dies stacked on edge, dies stacked in layers, and vertically stacked multichip modules.[3]

Cost-effective multiple IC packages have been in high volume production for several years. A polyimide tape interposer is used to mount mirrored memory die. One example of this is a Sharp (Osaka, Japan) device that integrates a static RAM with a flash memory. The result not only enhances access time but also reduces the area to about half of the size of conventional packaging.[14]

Three-dimensional packaging provides increased system performance over shorter interchip connections. Also, potentially lower system power is necessary for physically smaller and fewer drivers. To reduce the package thickness and bond length, the integrated circuits are often thinned. Three-dimensional techniques incorporate many of the manufacturing processes currently used in 2-D production, adapting standard semiconductor processing equipment. Most processing is accomplished simultaneously on multiple die.

The materials used in 3-D and their critical properties are the same as those used in advanced 2-D packaging and interconnection. Common materials found in both 2-D and 3-D structures include polyimides for dielectrics, copper or aluminum for conductors, solders for bonding, and a variety of substrates using alumina, silicon, and aluminum nitride. Common properties of these materials are discussed in the literature.[1,5] Much of the success of 3-D packaging will depend on adhesive technology. Thermosetting adhesives used to bond the dies together in a stack, including polyimides and epoxies, are selected for strength and stability. Thermoplastic adhesives used to bond stacked substrates are selected for compliance and reworkability. In some cases, a hierarchy of adhesives, based on melting or softening temperature, is needed to allow disassembly of the structure; this is unique to 3-D technology.

The manufacturing process for 3-D microelectronic devices depends on the format chosen for the final product. The package elements of layered die include the die themselves, the dielectric between the die, and interconnects. The elements of dies stacked on edge, layered dies, and vertically stacked modules include dies, attachments between dies or modules, wires and wirebonds, and interconnects between dies or modules. The package-to-substrate interconnects are usually solder bumps or pad arrays.

Table 10A1.6 compares the process attributes and interconnection techniques of some of the major 3-D package manufacturers. The manufacturing yield of 3-D packages depends on the number of directly

TABLE 10A.1.6 Comparison of Three-Dimensional Packaging Technology

	Supplier					
	Texas Instrument	**Irvine Sensor**	**Irvine Sensor**	**Thomson CSF**	**General Electric**	**Hughes**
Stack approach	Edge stack	Edge stack	Layered stack	Edge or layered stack	Stacked substrate	Stacked substrate
Full IC test and burn-in	Yes	No	No	Yes	Yes	Yes
On-cube reroute required	No	Yes	Yes	No	Yes	No
Logic IC integration	Yes	No	No	No	Yes	Yes
Interconnect chip-to-chip	Epoxy w/TAB	Polyimide	Polyimide	Epoxy resin	Polymer glues	—
Interconnect of chip cube to substrate	Butt tin-lead solder joint	Flip chip solder joint	Flip chip solder joint	Wirebond	N/A	Micro-bridge technology
Chip modification	No	Yes	Yes	No	No	No
Packaging efficiency	Medium	High	High	Medium	High	High
Heat dissipation	Good	Good	Poor	Good	Poor	Poor

connected dies built into the device. For example, if the yield of a particular type of die is 90% and the completed 3-D module comprises three layers, the final yield of the module can be as low as 73%. In some format types, screening can be conducted prior to module assembly to remove defective dies from the 3-D structure or to "turn off" the defective die. In some cases, the fabrication process incorporates redundancy to minimize the number of costly screens.

In 3-D packaging, both single-level, horizontal cross-talk and cross-talk between different layers can be potential problems for very high-speed applications. Furthermore, the approach may require the use of heat spreaders when using multiple high-power devices in order to meet thermal operating specifications as well as thermal and thermomechanical stress design limits.

10A.1.2 Printed Wiring Board Technology

The printed wiring board is a substrate on which electronic components are mounted for electrical interconnection and mechanical support. The interconnection is provided by patterned metal tracks either on the surface or on the inner layers of the printed wiring board. Based on the raw materials used for manufacture, printed wiring boards can be either organic or ceramic boards. The most common organic board material consists of a woven glass fabric bound by a resin. Ceramic board materials typically include alumina, aluminum nitride, and beryllia. Often, in ceramic boards, materials and structures are added to create embedded capacitors and resistors, freeing precious top-layer real estate. While not widely in production, similar techniques can be used in the inner layers of organic boards.

Conventional Printed Circuit Board Technology

Conventional printed wiring board (PWB) technology can be categorized as single layer, double layer, or multilayer. Single-layer boards have conductors etched on one side of the board, usually without hole plating. Double-layer boards have conductors on both sides of the board, but no inner layers. Multilayer organic boards typically consist of stacks of partially or fully cured fabric-reinforced epoxy sheets and copper foil conductors.

While there are many ways to fabricate a multilayer printed wiring board, a typical flow chart of the fabrication process is shown in Figure 10A.1.4.[1] In conventional printed circuit board technology, the

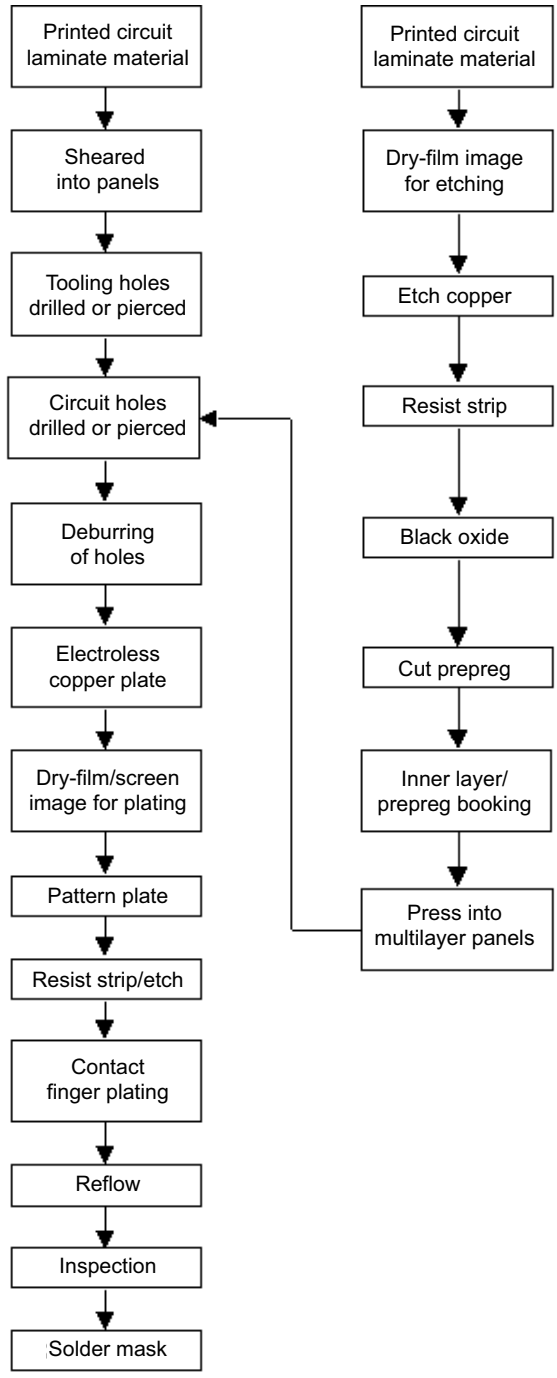


FIGURE 10A.1.4 Typical multilayer PWB fabrication process.

conductors are generally formed through both subtractive and additive processes. The partially cured laminate sheets, called *prepreg*, are cured and laminated with pressure and temperature.

The most common and least expensive resin system is FR-4 epoxy. This epoxy is a diglycidyl ether of tetrabromobisphenol A, with a glass transition temperature ranging from 135 to 180°C. Variations are

TABLE 10A.1.7 Comparison of Common Printed Wiring Board Materials

Substrate Material	Glass Transition Temp. (°C)	Dielectric Constant @ 1 MHz	Dissipation Factor @ 1 MHz	In-Plane CTE (ppm/K)	Relative Cost (FR-4 = 1)
FR-4 epoxy-glass	125–135	4.10–4.60	0.028–0.030	14–16	1
Polyfunctional FR-4	140–150	4.10–4.60	0.028–0.030	14–16	1.1–1.5
Multifunctional FR-4	150–170	4.10–4.60	0.028–0.030	14–16	1.1–1.5
Polyimide-glass	210–220	3.95–4.05	0.01–0.015	12–14	2–3
BT epoxy	180–190	3.85–4.10	0.011–0.013	13–17	1.1–1.5
Cyanate ester-glass	240–250	3.50–3.60	0.005–0.007	11–15	4–5
PPO epoxy	175–185	3.80–4.20	0.008–0.009	12–13	2–3
PTFE-glass	327 melting	2.45–2.55	0.001–0.003	70–120	15
Alumina	N/A	8.90	.0002–.0015	6.8	8–12

sometimes introduced in the resin to tailor desirable mechanical and electrical properties, including glass transition temperature, dielectric constant, coefficient of thermal expansion, moisture absorption, and thermal conductivity.

Holes in conventional PWBs are typically mechanically drilled. Those holes that allow conduction through all layers are called *plated through-holes* (PTH) for through-hole leads. For surface-mount devices, the holes are referred to as *plated through-vias* (PTV). Blind vias begin at a surface layer of the board and end at an internal layer. Buried vias are located in internal layers and do not include the external layers.

The explosion of commercial wireless (1–2 GHz) and high-performance processors (>1 GHz) is challenging the limits of FR-4. The edge rates of the signals and clock are actually more critical to application design than the clock frequency. The dissipation factor correlates with signal distortion and the dielectric constant with the signal delay. Thus, the relatively high dielectric constant of FR-4 (4.2) and high dissipation factor (0.030) make FR-4 use increasingly difficult for high-frequency (>150-MHz board frequencies) and fast-edge-rate (less than 1–2 ns) applications. While polytetrafluoroethylene (PTFE) materials have been used in military and high-end commercial applications for three decades because of their low dielectric constant of 2.4 to 2.6, these materials tend to be about 15 times costlier than FR-4. New materials such as polyphenylene oxide (PPO)/epoxy have reduced the dissipation factor significantly. Higher performance resins, such as polyfunctional FR-4 epoxy, bismaleimide triazine/epoxy (BT), cyanate esters (CE), and polyimide, are also used for a variety of special applications. Some key parameters and the relative costs of common materials are summarized in Table 10A.1.7.

Printed wiring boards can also be connected by flexible dielectric materials called flex print. The rigid portion of printed wiring boards can be either organic or ceramic, while flexible circuits use thinner, flexible organic base materials, including polyester and polyimide. Flex print provides an interconnection between boards that can be electrically superior to conventional connectors.

High-Density Interconnect Technology

Driven by increasing component I/O densities, board technology is migrating to the high-density interconnection (HDI). While there are numerous processes to form HDI layers, these can be classified into two general categories: microvia and thin-film. The microvia approach offers many advantages, including high frequency, high clock-rate performance, high wiring densities, and high ratios of active device area to substrate area. While the increased density comes at a higher cost, the per-function cost can actually be significantly lower. Thin-film HDI technology has been used in space applications and supercomputer applications.

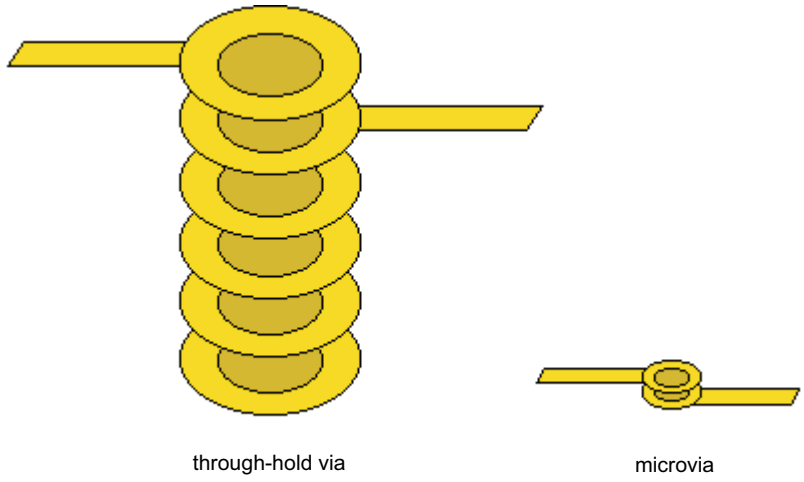


FIGURE 10A.1.5 Size advantage of microvia technology. (From IPC, *HDI and Microvia Technology,* Institute for Interconnecting and Packaging Electronic Circuits, Northbrook, IL, www.ipc.org/html/hdi.htm. With permission.)

Microvia HDI

Microvia HDI is becoming increasingly important with the advent of fine pitch array packaging. The Institute for Interconnecting and Packaging Electronic Circuits (IPC) defines a microvia as a blind or buried via of less than 150 μm, with a capture pad of less than 300 μm. Microvias can be used with conventional boards or chip-level packaging. In conventional boards, microvia technologies can be used selectively to redistribute difficult-to-route areas and enable components on both sides of the board. Ball grid arrays (BGAs) and CSPs with pitches of less than 1 mm typically require microvia redistribution layers. Figure 10A.1.5 shows the size advantage of microvias over conventional multilayer plated through-vias. Figure 10A.1.6 is a typical microvia stack-up; note that the microvia is actually included in the component solder pad. Current volume production yields exceed 99% in Japan and Taiwan. The microvias are generally formed by three basic approaches: plasma etch, photoimaging, and laser ablation. Laser ablation is a sequential technique, while plasma etch and photoimaging form vias simultaneously. Due to the high energy and difficulty of ablating glass fiber, the materials are typically not reinforced. Mechanical drilling can also be cost effective in material from 250 μm thick down to about 125 μm thick.[15]

Because there is no glass reinforcement, the dielectric constant of FR-4 epoxy is significantly lower than its reinforced counterpart. DYCOstrate®, developed by Dyconex of Zurich, Switzerland, is widely used in East Asia. An isotropic oxygen plasma is used to form vias that range from 75 to 150 μm in diameter. One advantage of this process is that the plasma removes all organic material, leaving the microvia clean.[17] The parallel processing and high yield make this process very cost effective.

A liquid- or dry-film photoimageable dielectric is generally an epoxy-acrylate material. Processing is similar to conventional soldermasking, followed by copper plating and patterning. Via diameters range from 50 to 250 μm.

Both CO_2 and ultraviolet (UV) lasers can be used to ablate the dielectric; UV can also burn through the copper layers. The CO_2 lasers are able to produce much smaller holes (10–25 μm) and are faster than the UV laser (25–200 μm). Laser vias can be formed in a variety of ways, including a two-pass process (first through copper, followed by a dielectric at a different frequency), a combination of conventional etching through the copper and laser ablation of the dielectric, and, finally, a deposited dielectric with laser via formation followed by an additive outer copper layer.[18] The remaining residue at the bottom of the ablated via can be removed with potassium permanganate or plasma etching for higher aspect ratio holes.

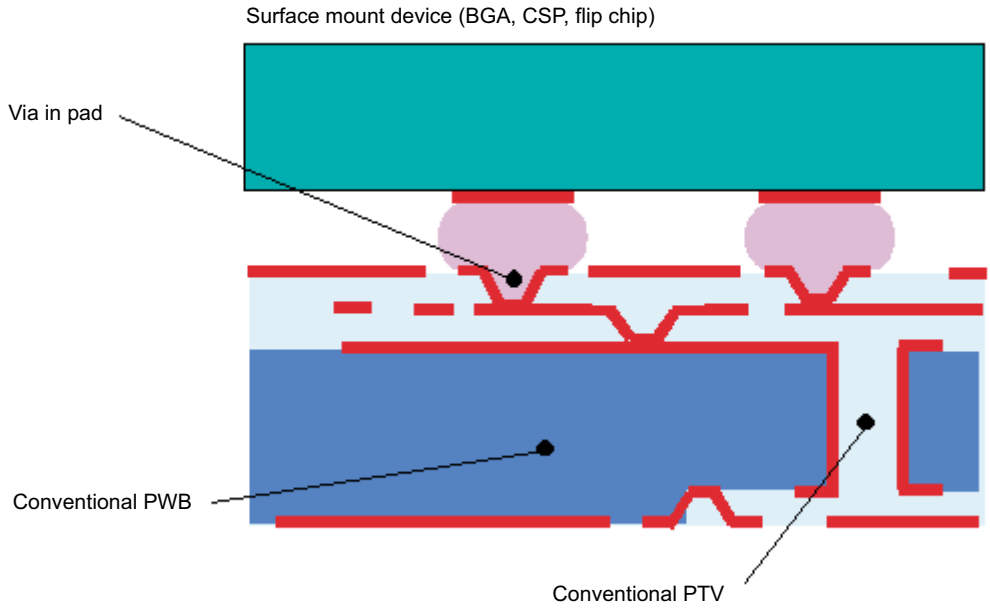


FIGURE 10A.1.6 Example of microvia technology. (From IPC, *HDI and Microvia Technology,* Institute for Interconnecting and Packaging Electronic Circuits, Northbrook, IL, www.ipc.org/html/hdi.htm. With permission.)

The reliability of the microvias is generally excellent, due to an aspect ratio of less than one. Other critical reliability factors are the hole diameter, plating thickness, hole cleanliness, and plating uniformity. In a sample of 15 different microvia test cases, only two test cases could not pass 2000 liquid-to-liquid thermal shock cycles (–55 to 125°C).[19] In another study of 18 U.S. and European fabricators, strong correlation was found between hole diameter and reliability. Holes of less than 150 μm were not reliable. Vias in solder pads were found to have no impact on solder-joint reliability.

Thin-Film HDI

All thin-film HDI approaches use deposited thin-film conductors with organic or inorganic interlayer dielectrics. The materials are sputtered with processes similar to semiconductor processing. Vias provide interconnection between conductor layers. Fully assembled modules using thin-film HDI are referred to as MCM-Ds.

Thin-film HDI can be classified into four generic types: overlay HDI, reach-through via HDI, beveled die-edge HDI, and conventional, chip-last, MCM-D, thin-film conductor and deposited dielectric HDI. A chip-last approach may employ wirebonding or other interconnects for die-to-substrate connectivity.

Table 10A.1.8 illustrates some of the many manufacturers and research institutions that make various types of HDI and MCM-D substrates. Typical design considerations, such as the number of metal layers, conductor line width, line spacing, via diameter, and via pitch, are listed in Table 10A.1.8 for each manufacturer. While many layers of dielectric and metal are possible, applications are often easily routed in two signal layers and a power-and-ground layer.

Ceramic Substrate Technology

Thick-film ceramic boards are formed by screen printing conductive, resistive, and dielectric paste onto a dielectric substrate and then firing. Typical thicknesses are in the range of 10 to 60 μm, depending on the materials. The classification "thin-film" refers to materials that are deposited in a sputtered vacuum process, similar to that used in semiconductor fabrication. Thin-film feature sizes are an order of

TABLE 10A.1.8 HDI and MCM-D Capabilities

Manufacturer	Maximum Number of Metal Layers	Minimum Line Width (μm)	Minimum Via Diameter (μm)	Minimum Line Spacing (μm)	Minimum Via Pitch (μm)
Rogers Corp.	9	50	75	125	150
AT&T	4	50	150	120	200
Toshiba	8	40	50	114	150
MMS	11	18	25	75	75
NEC	8	25	50	75	75
NTK	6	25	25	63	63
Hughes	10	15	20	60	60
Narumi	4	10	20	60	60
Kyocera	7	25	50	50	50
GE	5	37	44	50	114
HP	7	15	6	30	30
MCC	8	8	8	25	25
NChip	5	10	8	25	25
Midway	8	10	10	20	20
NTT	4	3	15	20	25
Polycon	8	5	5	10	10

magnitude less than the thick-film structures. These hybrid composites can have multiple conductor layers and can include both thick- and thin-film processes.

10A.1.3 Printed Circuit Assembly Processes

The printed circuit assembly processes consist of placing components on the printed wiring board, soldering the components, and post-solder processing steps such as cleaning and conformal coating. Classification of soldering processes is based on the method of applying the solder and the heating processes involved. Two processes in common use are wave soldering and reflow soldering. Wave soldering is used primarily for through-hole technology, while reflow soldering is used for surface-mount assembly.

Through-Hole Assembly

Through-hole components can be placed manually or with automated insertion equipment. High-speed equipment can place more than a thousand devices per minute, crimp the leads on the opposite side of the board, and dress the lead to an appropriate length. In wave soldering, the board is passed over a jet of molten solder. The hydrostatic pressure of the solder from the nozzle wets the underside of the PWB and pushes the solder slightly up the plated through-hole. Subsequent capillary action, resulting from the surface tension between the solder and the walls of the plated through-hole, forces the molten solder to rise through the hole and wet the upper lands of the through-hole, securing the part in place.

Before soldering, boards must be fluxed and preheated. Fluxing consists of applying a chemical compound with activators, solvents, and detergents to the solder-joint locations to chemically clean the surfaces to be joined. Preheating reduces thermal shock to the components and the PWB, activates the flux, and evaporates any flux solvents to prevent the formation of blowholes (voids) in the solder joint.

SMT Assembly

The reflow soldering process consists of remelting solder previously applied to a PWB joint site (pad) in the form of a preform or paste. No solder is added during reflow. The first step in the process is to apply solder paste and any necessary adhesives to the land patterns on the printed wiring board. Adhesives are needed only if the components cannot be held in place by the tackiness of the solder paste. After

component placement, the populated PWB is transported to the preheater to evaporate solvents in the solder paste. In this stage, the temperature of the board is raised to 100 to 150°C at a rate low enough to prevent solvent boiling and the formation of solder balls; 2°C/second has been accepted as an industry norm.[20]

The next stage in the reflow process is slow heating, which increases the temperature to the solder melting point and activates the flux in the solder paste. The activated flux removes oxides and contaminants from the surfaces of the metals to be joined. The next stage consists of melting the solder. During the melting stage, the temperature of the solder paste is raised to just above its melting point. When the solder melts, it replaces the liquid flux formed in the previous step. The temperature is held above the melting point while the solder coats the surfaces to be joined. The temperature must be regulated to allow the melted solder particles to coalesce and form a fillet around the component lead. Excessive temperatures increase the fluidity of the solder, causing it to move away from the desired joint location, while excessively long periods at the wetting temperature may lead to intermetallic formation and embrittlement. Fillet formation is the most critical part of the reflow process. The last stage involves cooling the solder joint, either by conducting the heat through the board layers or by natural or forced convection to the ambient air.

Reflow methods differ according to the method of heat transfer to the reflow site. Conductive, convective, and radiative heat transfers are variously used for this purpose. Conductive heat-transfer methods include the soldering iron, hot-bar reflow, and conductive-belt reflow. Radiative heat-transfer methodologies include infrared, laser, and optical fiber systems. Convective heat transfer is used in vapor-phase (or condensation) reflow systems. Each method has its own merits and limitations. The most common technique for mass reflow soldering is infrared heating/convection because it allows temperature control throughout the reflow soldering cycle. However, vapor-phase soldering is still considered the best way to limit the maximum temperature reached during reflow, as the temperature in the oven never rises above the temperature of the saturated vapor.

Connectors

Connectors can be classified based on the electronic elements being connected. Board-to-board connectors are used to connect two printed circuit boards; wire-to-board connectors are usually used to connect a source of power to a printed circuit board or to bus signals with high fidelity; and wire-to-wire connectors are generally used externally to the electronic equipment. In contemporary systems, connectors are a vital link in completing an electronic system. Connectors should be chosen carefully, taking into consideration current, voltage, impedance, EMI/EMR shielding, and allowable losses.

A connector is composed of four basic elements:[21] the contact interface, the contact finish, the contact spring, and the connector housing. The contact interfaces are of two kinds: separable and permanent. The separable connector allows for mating and unmating, while the permanent interface does not. The contact finish aids in establishing a contact interface and protecting the contact springs from corrosion. The contact spring is the electrically conducting element between the permanent connection to the subsystems and the separable interface that supplies the force necessary to maintain contact at the separable interface. The connector housing performs both mechanical and electrical functions by electrically insulating the individual contacts and supporting and locating them mechanically.

Board-to-board connectors are an important part of today's packaging. Commercial applications are making high-density connectors inexpensive. Mezzanine assemblies are gaining popularity, as processor modules become more complex in order to accommodate processor upgrades. One example is Intel's Pentium II® Mobile Module, which uses a 400-pin, 1.27-mm pitch connector.[22]

10A.1.4 Electronic Packaging Future

The future of electronic packaging promises an accelerating change. For several years now, industry roadmaps have assessed the capabilities, needs, direction, and future of electronic packaging. Hundreds

of companies, government agencies, consortia, and universities have contributed to roadmaps. There are a number of these long-range plans, including the National Electronics Manufacturing Initiative (NEMI),[14] the Japan Printed Circuit Association's *Report on the Technology Roadmap for Advanced System Integration and Packaging*,[23] and IPC's *The National Technology Roadmap for Electronic Interconnections*.[24] Many large companies and research universities involved in packaging, such as the University of Maryland's CALCE Electronic Products and Systems Center, also have roadmaps that complement and fill research holes.

The most recent findings are that cost is king in nearly all applications. Cost, a major driver in the advancement of technology, is also driving the trend toward electronics manufacturing service (EMS) suppliers or contract manufacturers. Cost and business pressures are fueling the 25% annual increase in EMS providers. The advantages of EMS include optimizing manufacturing cycle times, reducing working capital, and improving quality. The NEMI roadmap suggests that a downside of this EMS trend will be a slowing of research and development, as the EMS providers will push to keep the *status quo* rather than advance to less proven areas. The NEMI roadmap predicts that this will also lead to shorter product cycles and more conservative, evolutionary packaging approaches. If the new technology cannot achieve cost parity with existing technologies, then industry acceptance will be delayed. One example of this is flip chip packaging, competing with those technologies that are now widely used: CSP and BGA.[25]

The NEMI, IPC, and the Semiconductor Industry Association roadmaps all predict a movement from peripheral to array packaging; consequently, peripheral packages will not decrease below a 0.5-mm pitch. The growth of flip chip interconnection was seen as a growing trend in the industry by all roadmaps. The NEMI roadmap sees integral passives (resistors and capacitors) as an important part of the future in portable electronics, whereas the IPC sees the enabling technologies as immature.[26]

Some recent predictions from the NEMI roadmap are summarized in Tables 10A.1.9 and 10A.1.10. The cost of I/O and labor to assemble packages will continue to fall dramatically. Substrate lines, spaces, and device I/O pitch will also shrink to keep pace with silicon integration. Noncontact test infrastructure

TABLE 10A.1.9 NEMI Predictions for the Future[25]

	1999	2003
IC package costs per I/O (¢)	0.7–0.9	0.4–0.6
Board assembly conversion (¢)	0.7–0.9	0.5
Substrate lines and spaces (μm)	70	35
Process test pads (mm)	0.5	Non-contact
Flip chip pad pitch (μm)	180	130

TABLE 10A.1.10 Package Forecast by Lead Count Range

I/O	1998	2003
4–18	28,495	38,768
20–32	16,347	20,489
36–68	8442	18,502
72–100	2043	5160
104–144	1551	3222
148–208	922	2378
212–304	405	1070
308+	511	1710
Total	58,716	91,299

From Berry, S., The future of leadcounts, *HDI*, 2(4), 14–16, 1999. With permission.

will be required to provide efficient and cost effective assurance that assemblies are functional. A shift to higher pin-count packages is also expected, as shown in Table 10A.1.10. While the current percentage of packages with a pin count greater than 304 is less than 1%, the percentage in the year 2003 will nearly double.

Electronic packaging technology will continue to enable many compact and high-speed products of the 21st century. The need for innovations in cost, yields, and productivity will certainly challenge those engineers involved in the many facets of packaging. These improvements will eventually touch and, it is hoped, enhance the quality of life of all the world's people.

References

1. Pecht, M., *Handbook of Electronic Package Design*, Marcel Dekker, New York, 1991.
2. Pecht, M., *Integrated Circuit, Hybrid, and Multichip Module Package Design Guidelines*, John Wiley & Sons, NewYork, 1993.
3. Pecht, M., Evans, J., and Evans, J., *Quality Conformance and Qualification of Microelectronic Packages and Interconnects*, John Wiley & Sons, New York, 1994.
4. Pecht, M., Nguyen, L.T., and Hakim, E.B., *Plastic Encapsulated Microelectronics*, John Wiley & Sons, New York, 1994.
5. Pecht, M.G., Agarwal, R., McCluskey, P., Dishongh, T., Javadpour, S., and Mahajan, R., *Electronic Packaging: Materials and Their Properties*, CRC Press, Boca Raton, FL, 1999.
6. Lau, J.H., Wong, C.P., Prince, J.L., and Nakayama, W., *Electronic Packaging: Design, Materials, Process, and Reliability*, McGraw-Hill, New York, 1998.
7. Hwang, J.S., BGA and CSP solder spheres, *Surface Mount Technology*, April, p. 18, 1999.
8. Baker, J.D., Nation, B.J., Achari, A., and Waite, G.C., On the adhesion of palladium silver conductors under heavy aluminum wire bonds, *International Journal for Hybrid Microelectronics*, 4(2), 155–160, 1981.
9. Houston, T., Inside the newest Sony camcorder, *Portable Design*, May, pp. 28–32, 1997.
10. Lockard, S.C., Hansen, J.M., and Nelson, G.H. Multimetal layer TAB for high performance digital applications, in *Proceedings of the Technical Program, National Electronic Packaging and Production Conference West*, 1990, pp. 1113–1122.
11. Walker, J., *Multichip Module Packaging and Assembly*, Advancement Course 11, Surface Mount Technology Conference, Secaucus, NJ, 1992.
12. Lau, J.H., Erasmus, S.J., and Rice, D.W., Overview of tape automated bonding technology, in *Electronic Materials Handbook*, Vol.1, ASM International, Material Park, OH, 1991, pp. 274–295.
13. 3M, Electronic Products Division, *Specifications and Design Guidelines*, Austin, TX, 1987.
14. Combo memories — stack 'em high (editorial), *Portable Design*, December, 18 pp., 1997.
15. Bosnjak, M. and Schlemmer, S., Mechanical microvia formation, *Printed Circuit Fabrication*, May, pp. 36–40, 1999.
16. IPC, *HDI and Microvia Technology*, Institute for Interconnecting and Packaging Electronic Circuits (IPC), Northbrook, IL, 1999 (www.ipc.org/html/hdi.htm).
17. Virsik, P., Microvias: the hole story, *Printed Circuit Fabrication*, May, pp. 46–48, 1999 (www.pcfab.com/db_area/archive/1999/9905/microvias.html).
18. Singer, A. and Bhatkal, R., A cost analysis of microvia technologies: photo vs. plasma vs. laser, in *Proceedings of the IPC Printed Circuits Expo*, San Jose, CA, March 1997, 802 pp.
19. Rasul, J., Bratschun, W., and McGowen, J., Microvia bare board reliability assessment, in *Proceedings of the IPC Printed Circuits Expo*, San Jose, CA, March 1997, 802 pp.
20. Linman, D.L., Vapor phase for high reliability soldering, in *Proceedings of the Technical Program: National Electronic Packaging and Production Conference West*, Anaheim, CA, February 1991, pp. 278–285.
21. Mroczkowski, R.S., Materials considerations in connector design, in *Proceedings of the First Electronics Materials and Process Conference*, American Society for Materials, Chicago, IL, 1988.

22. Pentium MMX module extends Intel's hegemony (editorial), *Portable Design*, March, 16 pp., 1997.
23. Japan Printed Circuit Association (English version by IPC), *Report on the Technology Roadmap for Advanced System Integration and Packaging,* Institute for Interconnecting and Packaging Electronic Circuits (IPC), Northbrook, IL, 1998.
24. IPC, *The National Technology Roadmap for Electronic Interconnections,* Institute for Interconnecting and Packaging Electronic Circuits (IPC), Northbrook, IL, 1997.
25. Feinstein, L., Highlights of the 1998 NEMI roadmap, *HDI,* 2(4), 36–37, 1999.
26. Munie, G.C., Checking the roadmap, " *Surface Mount Technology*, March, pp. 50–52, 1999.
27. Berry, S. The future of leadcounts, *HDI,* 2(4), 14–16, 1999.

Supplemental References

Bhandarkar, S., Dasgupta, A., Pecht, M., and Barker, D., *Effects of Voids in Solder-Filled Plated-Through Holes,* IPC Technical Paper TP-863, 33rd IPC Conference, 1990, pp. 1–6.

Doane, D.A. and Franzon, P., Electrical design of digital multichip moldules, in *Multichip Modules: Technologies and Alternatives*, Van Nostrand-Reinhold, New York, 1993, pp. 368–393 (wire bonding, pp. 525–568).

Leonard, C. and Pecht, M., How failure prediction methodology affects electronic equipment design, *Quality and Reliability Engineering International,* 6(4), 243–250, 1990.

Manko, H.H., *Solders and Soldering*, McGraw-Hill, New York, 1992.

Markstein, H.W., The interconnection needs of the high-end devices market for ECL, ASICS and gate arrays is being satisfied by tape automated bonding, *Electronic Packaging and Production*, 33–36, 1990.

Pecht, M.G., *Soldering Processes and Equipment*, John Wiley & Sons, New York, 1993.

Woodgate, R.L., *Handbook of Machine Soldering*. John Wiley & Sons, New York, 1988.

Kreith, F; et. al. "Transportation"
Mechanical Engineering Handbook
Ed. Frank Kreith
Boca Raton: CRC Press LLC, 1999

Transportation

Frank Kreith
University of Colorado

Michael D. Meyer
Georgia Institute of Technology

John Leonard II
Georgia Institute of Technology

Paul W. Shuldiner
University of Massachusetts

Kenneth B. Black
University of Massachusetts

Paul Schonfeld
University of Maryland

Paul Norton
National Renewable Energy Laboratory

James B. Reed
National Conference of State Legislatures

Janet B. Goehring
National Conference of State Legislatures

Introduction

Transportation is important for an industrial society because the movement of goods and people has significant impact on a country's economy. In the United States the transportation sector accounts for approximately 20% of the gross domestic product. The transportation sector also accounts for almost two thirds of the total U.S. petroleum consumption, with cars and light duty trucks accounting for more than 20% of the total U.S. energy use. The cost of imported petroleum in 1993 was $56 billion—almost 10% of all imports. The transportation sector is also the nation's largest single source of air pollution, with personal vehicles producing 26% of volatile organic compounds, 32% of nitrous oxides, and 62% of carbon monoxide. Thus, better ways to promote transportation efficiency is one of the most important facets of reducing fossil fuel consumption and improving environmental quality.

Transportation engineering is a highly interdisciplinary field dealing with the planning, design, construction, maintenance, and operation of various transportation modes. This section presents an overview of transportation engineering, emphasizing planning, design, operation, environmental impacts, legal,

system analysis, and emerging issues such as alternative fuels, electric vehicles, and intelligent highway systems. Emphasis is placed on those facets of transportation that impact the mechanical engineering profession, but the area also needs social, political, and management inputs to arrive at appropriate engineering solutions of specific problems.

10.1 Transportation Planning

M. D. Meyer

Transportation planning is the process of producing the information necessary to determine the most cost-effective strategy for improving the performance of the transportation system. Transportation planning can occur on different scales of analysis — at the national, multistate, state, metropolitan, subarea or corridor, and local levels. At each level the planning methodologies used, databases developed, and information produced are often heavily influenced by government regulations. No matter what modeling approaches are used in transportation planning, a fundamental concept that influences the way transportation demand is estimated is called **derived demand**. Derived demand means that a trip is taken to accomplish some activity at the destination, and that the trip itself is simply a means of reaching this activity. Transportation planning must therefore relate trip making to the types of activities that occur in a study area and also to the characteristics of the trip maker that will influence the way these trips are made. This end is reached by combining similar uses of land into a land use category that can then be used in transportation planning to estimate how many trips are attracted to each type of land use (e.g., the number of trips to schools, shopping centers, residential units, office complexes, etc.).

Basic Framework of Transportation Planning

The basic framework for transportation planning at any scale of application is shown in Figure 10.1.1. The steps shown in this framework are discussed in the sections that follow.

Inventory of Facilities

Knowing what a **transportation network** consists of and the condition and performance of these facilities is an important starting point for transportation planning. Transportation investment is usually aimed at upgrading the physical *condition* of a facility (e.g., repaving a road or building a new bridge) or at improving its *performance* (e.g., providing new person-carrying capacity by providing preferential treatment for high-occupancy vehicles or building a new road to serve existing demand).

Collect and Maintain Socioeconomic and Land Use Data

Given the concept of derived demand, a basic starting point for planning is characterizing the variables that will influence this demand. Current land use is readily attained through land use inventories. The methods of estimating future land use range from trends analysis to large-scale land use models that predict household and employment sites decades into the future. Socioeconomic characteristics often include the level of income, number of members in the household, number of autos in the household, number of children, age of the head of household, and highest level of education achieved.

Define Goals and Objectives

Goals are generalized statements that indicate the desired ultimate achievement of a transportation plan. *Objectives* are more specific statements that indicate the means by which these goals will be achieved. The identification of goals and objectives is critical in that they define the evaluation criteria that will be used later in the planning process to assess the relative impacts of alternative projects and strategies. These criteria are often called *measures of effectiveness*.

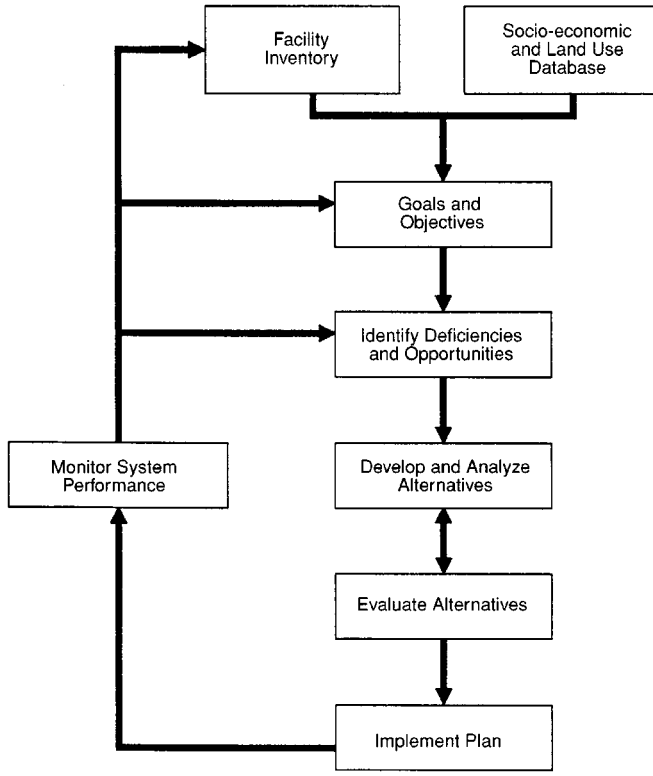


FIGURE 10.1.1 Basic elements of transportation planning.

Identify System Deficiencies or Opportunities

The methods used to identify transportation deficiencies and opportunities can vary widely. In some cases large-scale transportation network models are used to estimate future traffic volumes and then to compare them to the capacity of the road network to handle such volumes. This volume-to-capacity (V/C) ratio has served as the major means of identifying the location of system deficiencies. However, other *performance measures* can also be used in the transportation planning process that are much broader than the traditional V/C ratios, such as level of accessibility to economic activities, average travel times, and so on.

Develop and Analyze Alternatives

Various types of strategies can surface from the planning process: (1) those focused on improvements to highways — for example, adding new lanes, improving traffic control through signals or signing, or improving traffic flow through channelization; and (2) other strategies, such as reducing the demand for transportation through flexible working hours, increasing average vehicle occupancy through such measures as carpools or transit use, or raising the "price" of travel through the use of tolls. More recently, the application of advanced transportation technologies to the operation of a road system, known as *intelligent transportation systems*, has become an important strategy in many cities.

Evaluate Alternatives

Evaluation brings together all of the information gathered on individual alternatives and provides a framework to compare the relative worth of the alternatives. This evaluation process most often relies on the various measures of effectiveness that relate to the goals and objectives defined at the beginning of the planning process.

Implement Plan

The major products of the transportation planning process are the *transportation plan* and, at the state and metropolitan levels, the *transportation improvement program (TIP)*, which lists all of the transportation projects that will be implemented over the next few years in the region.

Monitor System Performance

Transportation planning continually examines the performance and condition of the transportation system to identify where improvements can be made. Some means of monitoring system performance is necessary to systematically identify areas where these improvements might occur. This system monitoring has traditionally been based on traffic volumes obtained from traffic counters placed at strategic locations in the network (e.g., increasing traffic volumes during a specific time period indicate congestion).

Transportation Modeling

The level of transportation analysis can vary according to the level of complexity and scale of application. In most cases, however, the modeling process consists of four major steps—trip generation, trip distribution, mode split, and trip assignment. Each study area (whether a nation, state, metropolitan area, or neighborhood) is divided into zones of homogeneous characteristics (e.g., similar household incomes) that can then be used as the basic foundation for estimating trips from, or attracted to, that zone. In most cases, transportation planners try to create a **zonal system** by defining zones that are related to other data collection activities (e.g., U.S. Census tracts). The transportation system is represented in models as a network of *links* and *nodes*. Links represent line-haul facilities, such as roads or transit lines, and nodes represent points of connection, such as intersections. Given the complex nature of transportation systems, the typical transportation network consists of links representing only highly used facilities. The transportation modeling process can thus be represented as shown in Figure 10.1.2.

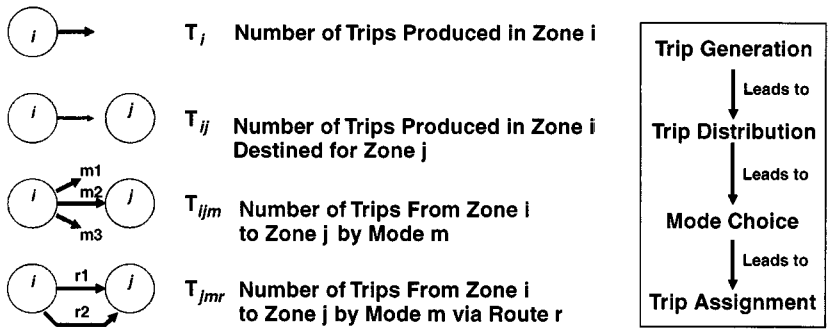


FIGURE 10.1.2 Transportation modeling framework.

Trip generation is the process of analytically deriving the number of trips that will be generated from a location or zone based on socioeconomic information of the travelers or, in the case of freight movement, zonal economic characteristics. The trip generation stage also includes predicting the number of trips that will be attracted to each zone in the study area:

- Number of trips produced in a zone = *f* (Socioeconomic characteristics, land use, transportation mode availability)
- Number of trips attracted to a zone = *f* (Attractiveness of zone)

There are two approaches for estimating trips generated. The first approach uses *trip rate models*. Based on existing data, trip rates can be calculated that relate trip-making to variables known to influence travel behavior. For an example, see Table 10.1.1.

TABLE 10.1.1 Cross-Classification Analysis

	Number of People in Households		
	1	**2**	**3+**
Low income	2.4	3.3	4.5
Medium income	3.5	3.8	4.8
High income	3.9	4.2	5.4

The other approach is to use *regression models*. Based on existing data, multiple regression equations are estimated relating trip generation to variables found to be significant in travel behavior. For example,

$$\text{Zone estimate:} \quad T_{iz} = 184.2 + 120.6\left(X1_i\right) + 34.5\left(X2_i\right)$$

$$\text{Household estimate:} \quad T_{ih} = 0.64 + 2.3\left(X1_{ih}\right) + 1.5\left(X2_{ih}\right)$$

$$\text{Zone attractions:} \quad T_j = 54.2 + 0.23(X3) + 0.43(X4)$$

where

T_{iz} = Total trips generated in zone i
T_{ih} = Total trips generated per household in zone i
T_j = Total trips attracted to zone j
$X1_i$ = Total number of workers in zone i
$X2_i$ = Total number of automobiles in zone i
$X1_{ih}$ = Number of employees per household in zone i
$X2_{ih}$ = Number of automobiles per household in zone i
$X3$ = Total number of office employees in zone j
$X4$ = Total number of retail employees in zone j

Trip distribution is the process of estimating the number of trips that originate in each zone in the study area and their corresponding destinations. The result of the trip distribution process is a matrix called a *trip table*, which shows the number of trips originating in each study zone and their corresponding destinations for the time period being examined. The most commonly used method for distributing trips in a zonal system is the gravity model, which is of the following form:

$$T_{ij} = P_i \frac{A_j \times F_{ij} \times K_{ij}}{\sum_j \left(A_j \times F_{ij} \times K_{ij}\right)} \tag{10.1.1}$$

where

T_{ij} = Number of trips originating in zone i and destined to zone j
P_i = Number of trips produced in zone i
A_j = Level of attractiveness of zone j (e.g., number of employees)
F_{ij} = Friction or impedance factor between zones i and j (a value usually a function of travel time)
K_{ij} = Socioeconomic adjustment factors for trips between zones i and j (a value that represents variables that influence trip making not accounted for by other variables)

Mode choice is the process of estimating the percentage of individuals who will use one mode of transportation over the others available for a given trip. The basic concept in making this estimation is that each mode of transportation has associated with it some empirically known characteristics that, when combined in a mathematical equation, can define that mode's *utility*. Variables such as travel time, travel cost, modal reliability, and so on are often incorporated into a mode's **utility function**, along with socioeconomic characteristics of the traveler. Freight models use a similar concept in estimating

commodity flows by mode. One of the most familiar forms of mode choice models is the logit model, which predicts mode shares based on the following equation:

$$P_{ik} = \frac{e^{U_k}}{\sum_{m=1}^{n} e^{U_m}} \tag{10.1.2}$$

where

P_{ik} = Probability of individual i choosisng mode k
U_k = Utility of mode k
U_m = Utility of mode m
n = Number of modes available for trip

The utility of each mode is often represented as a linear function of those variables found to influence an individual's choice of mode. For example, a utility function for the automobile mode might be of the form

$$U_a = 6.3 - 0.21(X_1) - 0.43(X_2) - 0.005(X_3) \tag{10.1.3}$$

where

U_a = Utility of automobile
X_1 = Access and egress time when automobile is chosen
X_2 = Line-haul time
X_3 = Travel cost

The utility functions of other modes available for a specific trip would be similarly specified. The probabilities would then be multiplied by the total number of trips to obtain the number of trips made by mode.

Trip assignment is the process of estimating the travel flows through a transportation network based on a trip table (which is produced in trip distribution). The basic concept found in all trip assignment models is that travelers choose routes that will minimize travel time—that is, they will choose the shortest path through the network. Link performance functions that relate travel time to number of vehicles on the link (see Figure 10.1.3) are used to iteratively update estimated link travel times so that minimum-path travel times reflect the effect of congestion. Recent research in trip assignment methods has introduced the concept of stochastic assignment, which means that some subset of trip routes will in fact have associated with them some characteristics that attract certain types of travelers, even if the travel time is longer.

Defining Terms

Derived demand: An assumption that travelers make a trip to accomplish some activity at the destination and that the trip itself is simply a means of reaching this activity.

Transportation network: A transportation system is represented in models as a network of *links* and *nodes*. Links represent line-haul facilities, such as roads or transit lines, and nodes represent points of connection, such as intersections. Given the complex nature of transportation systems, the typical transportation network consists of links representing only higher functionally classified facilities or those links that represent known highly traveled paths.

Utility function: A mathematical formulation that assigns a numerical value to the attractiveness of individual modes of transportation based primarily on that mode's characteristics.

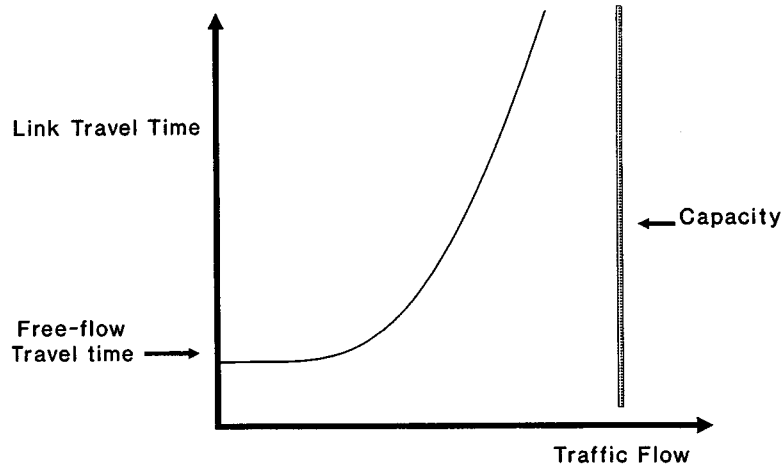


FIGURE 10.1.3 Link performance function.

Zonal system: Each study area (whether a nation, state, metropolitan area, or neighborhood) is divided into zones of homogeneous characteristics (e.g., similar household incomes) that can then be used as the basic foundation for estimating trips from or attracted to, that zone. In most cases transportation planners try to define zones that are related to other data collection activities (e.g., U.S. Census tracts).

References

Institute of Transportation Engineers. 1993. *Trip Generation Handbook*, 5th ed. ITE, Washington, DC.

Johnson, E. 1993. *Avoiding the Collision of Cities and Cars*. Report on a Study Project Sponsored by the American Academy of Arts and Sciences, September 1993.

Mannering, F. and Kilareski, W. 1990. *Principles of Highway Engineering and Traffic Analysis*. John Wiley & Sons, New York.

Meyer, M. and Miller, E. 1984. *Urban Transportation Planning: A Decision-Oriented Approach*. McGraw-Hill, New York.

Newell, G. 1980. *Traffic Flow on Transportation Networks*. MIT Press, Cambridge, MA.

Ogden, K. 1992. *Urban Goods Movement*. Ashgate, Brookfield, VT.

Ortuzar, J. and Willumsen, L. 1994. *Modelling Transport*, 2nd ed. John Wiley & Sons, New York.

Papacostas, C.S. 1992. *Fundamentals of Transportation Engineering*, 2nd ed. Prentice Hall, Englewood Cliffs, NJ.

Stover, V. and Koepke, F. 1988. *Transportation and Land Development*. Institute of Transportation Engineers, Washington, D.C.

Further Information

Transportation Research Board, National Academy of Sciences
2101 Constitution Ave., N.W.
Washington, DC 20418
American Association of State Highway and Transportation Officials
444 N. Capitol St., N.W.
Suite 225
Washington, DC 20001
Institute of Transportation Engineers
525 School St., S.W.
Suite 410
Washington, DC 20024

10.2 Design of Transportation Facilities

John Leonard, II and Michael D. Meyer

The efficient movement of people and goods requires transportation systems and facilities that are designed to provide sufficient capacity for the demands they face in a safe manner. In addition, in most modern societies, the design of transportation facilities must explicitly minimize harm to the natural and human-made environment while providing for mitigation measures that relate to those impacts that are unavoidable. In many ways the critical challenge to today's designers of transportation projects is successfully designing a facility that minimally harms the environment.

The design of a transportation facility almost always takes place within the context of a much broader **project development process**. This process can vary in complexity with the type of project under design and with the scale of implementation. The importance of the project development process to the designer is that it

- Establishes the key characteristics of the project that must be considered in the design
- Indicates the time frame that will be followed for project design
- Establishes which agencies and groups will be involved in the process and when this involvement will likely occur
- Links the specific elements of the project design with other tasks that must be accomplished for the project to be constructed
- Satisfies legal requirements for a design process that is open for review and comment
- Indicates the specific products that must be produced by the designers to complete the project design process

In most cases the project development process consists of a well-defined set of tasks each of which must be accomplished before the next task can occur. These tasks include both technical activities and public involvement activities that are necessary for successful project development.

Components of the Project Development Process

Identify Project Need

A project need can be identified through a formal planning process or from a variety of other sources, including elected officials, agency managers, transportation system users, and citizens. Important in this early portion of project development is an indication of what type of improvement is likely to be initiated. For example, a project could relate to one or more of the following types of improvement strategies:

- *New construction.* A transportation facility constructed at a new location.
- *Major reconstruction.* Addition of new capacity or significant changes to the existing design of a facility, but usually occurring within the area where the current facility is located.
- *Rehabilitation/restoration.* Improvements to a facility usually as it is currently designed and focusing on improving the physical condition of the facility or making minor improvements to enhance safety.
- *Resurfacing.* Providing new pavement to a transportation facility that prolongs its useful life.
- *Spot improvements.* Correction of a problem or hazard at an isolated or specific location.

Establish Project Limits and Context

One of the very first steps in the design process is to define the boundaries or limits of the project. This implies establishing how far the project will extend beyond the area being targeted for improvement and the necessary steps to ensure smooth connections to the existing transportation system.

Establish Environmental Impact Requirements

The design of a project will very much be affected by environmental laws or regulations that require design compliance with environmental mandates. These environmental mandates could relate to such things as wetlands, historic properties, use of public park lands, water quality, navigable waterways, fish and wildlife, air quality, noise, and archaeological resources. One of the first steps in project development is to determine whether the likely project impacts are significant enough to require environmental study.

Develop Strategy for Interagency Coordination and Public Involvement

Depending on the complexity and potential impact of a project, the project designer could spend a great deal of time interacting with agencies having some role in or jurisdictional control over areas directly related to the project. These agencies could have jurisdiction by law (e.g., wetlands) or have special expertise that is important to project design (e.g., historic preservation). In addition to interagency coordination, transportation project development is often subject to requirements or public outreach and/or public hearings.

Initiate Project Design and Preliminary Engineering

Topographic data of the study area and forecasted vehicular volumes expected to use the facility in the design year are used as input into the preliminary design of the horizontal and vertical alignment of the facility, that is, the physical space the facility will occupy once finished. This preliminary engineering step also includes the preparation of initial right-of-way (ROW) plans, which indicate the amount of land that must be available to construct the facility.

Project Engineering

Once preliminary engineering has provided the basic engineering information for the project, the more detailed project design begins. This entails specific layouts of horizontal and vertical geometry, soils/subsurface examination and design, design of utility location, drainage design, more detailed ROW plans, and initial construction drawings. Concurrent with this design process, the environmental process continues with updated information on project changes that might cause additional environmental harm, the initiation of any permitting process that might be needed to construct the project (e.g., environmental agency permission to affect wetlands), and public hearings/meetings to keep the public involved with project development.

Final Engineering

The final engineering step is the culmination of the design process, which basically completes the previous design plans to the greatest level of detail. This step includes finalizing ROW plans, cost estimates, construction plans, utility relocation plans, and any agreements with other agencies or jurisdictions that might be necessary to complete the project. Environmental permits are received and final project review for environmental impacts is completed.

Basic Concepts of Project Design

Human Factors

Human factors have a great deal of influence on the design of transportation facilities in such areas as width of facility, length and location of access/egress points, braking distance, location of information/guidance aids such as signs, and geometric characteristics of the facility's alignment. The driver-vehicle-roadway interface is shown in Figure 10.2.1.

Vehicle or User Performance Factors

The dynamics of vehicle motion plan an important role in determining what is effective and safe design. The key vehicle characteristics that become important in design criteria include:

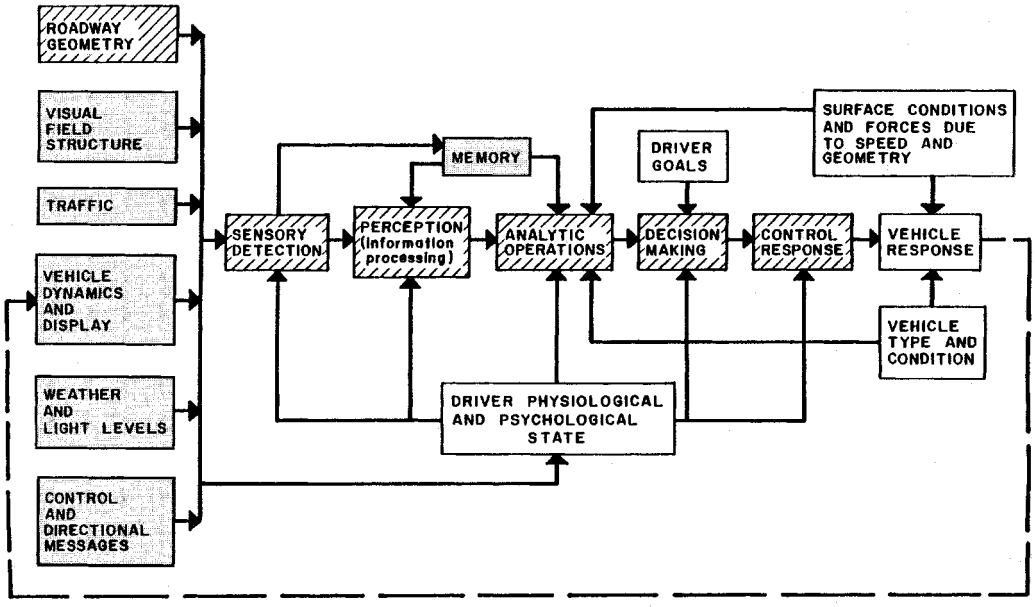


FIGURE 10.2.1 Driver-vehicle-roadway interface. Gray shaded boxes, information inputs; cross-hatched boxes, driver guidance and control process.

- *Vehicle size*. Influences vertical and horizontal clearances, turning radii, alignment width, and width of vehicle storage berths.
- *Vehicle weight*. Influences strength of material needed to support vehicle operations.
- *Vehicle or user performance*. Influences specifications for horizontal and vertical geometry, braking distances, and needed capacity to allow passing and successful maneuvering (e.g., assumed walking speed of pedestrians crossing a road).

Classification Schemes

The transportation system serves many functions, ranging from providing access to specific locations to providing high-speed, high-capacity movement over longer distances. Classification schemes are used to represent these various roles and influence the design criteria that are associated with the facilities in each classification category. A common **functional classification** scheme for highways is shown in Figure 10.2.2.

Capacity and Level of Service

Every design usually begins with some estimation of the demand for the transportation facility that will likely occur if the facility is built. The key design question then becomes, what facility capacity (e.g., number of road lanes, runways, transit lines, or vehicle departures) is necessary if a certain level of service is desired? The definition of **level of service (LOS)** thus becomes a critical element in establishing this important design factor (see Figure 10.2.3).

Design Standards

Design standards dictate minimum or maximum values of project characteristics that are associated with a particular facility type. Design standards usually result from extensive study of the relationship between various design characteristics and vehicle performance and the safe handling of the vehicles by human operators. Design standards often vary by the "design speed" of the facility (and thus the

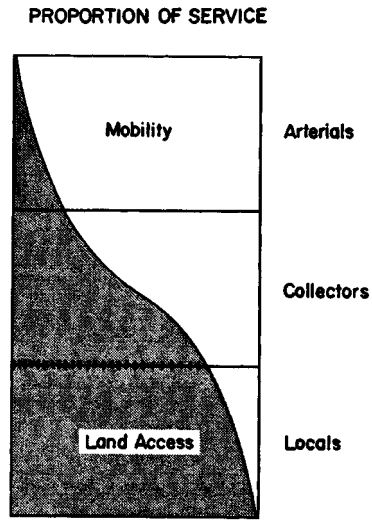


FIGURE 10.2.2 Relationship of functionally classified systems in service traffic mobility and land access. (*Source*: AASHTO, 1990, Figure I-5.)

importance of the-facility classification) and by the "design vehicle." Design standards are often the basis for developing typical cross sections (see Figure 10.2.4 and 10.2.5).

Intermodal Transportation Terminals or Transfer Facilities

Terminals or transfer facilities are locations where users of the transportation system change from one mode of travel to another. The effective design of such facilities is a critical element of successful transportation system performance, given the potential bottlenecks they represent if not designed appropriately. The design of terminals and transfer facilities must pay special attention to the needs of the users of the facility, in that they serve to establish the effective capacity of the facility. For example:

1. Internal pedestrian movement facilities and areas (stairs, ramps, escalators, elevators, corridors, etc.)
2. Line-haul transit access area (entry control, fare collection, loading, and unloading)
3. Components that facilitate movements between access modes and the station (ramps or electric doors)
4. Communications (public address systems and signage)
5. Special provisions for disabled patrons (elevators and ramps)

The criteria that could relate to the design of such a facility include threshold values for pedestrian level of service, delay at access points, connectivity from one area of the facility to another, and low-cost maintenance. For the vehicle side of such terminals, special consideration must be given to the performance of the design vehicle (e.g., turning radii of buses or semitrailer trucks) and the vehicle storage requirements (e.g., the number, size, and orientation of loading/unloading berths).

Advanced Technology Projects

One of the characteristics of transportation system development in recent years has been the increased application of advanced (usually electronic) technologies to improve system performance. The following steps apply to designing advanced technology projects.

FREEWAYS

Level of Service	Maximum Density (pc/mi/ln)	Minimum Speed (mph)	Max Service Flow Rate (pcphpl)	Maximum *v/c* Ratio
		Free-Flow Speed = 70 mph		
A	10.0	70.0	700	0.318/0.304
B	16.0	70.0	1120	0.509/0.487
C	24.0	68.5	1644	0.747/0.715
D	32.0	63.0	2015	0.916/0.876
E	36.7/39.7	60.0/58.0	2200/2300	1.000
F	var	var	var	var
		Free-Flow Speed = 65 mph		
A	10.0	65.0	650	0.295/0.283
B	16.0	65.0	1040	0.473/0.452
C	24.0	64.5	1548	0.704/0.673
D	32.0	61.0	1952	0.887/0.849
E	39.3/43.4	56.0/53.0	2200/2300	1.000
F	var	var	var	var
		Free-Flow Speed = 60 mph		
A	10.0	60.0	600	0.272/0.261
B	16.0	60.0	960	0.436/0.417
C	24.0	60.0	1440	0.655/0.626
D	32.0	57.0	1824	0.829/0.793
E	41.5/46.0	53.0/50.0	2200/2300	1.000
F	var	var	var	var
		Free-Flow Speed = 55 mph		
A	10.0	55.0	550	0.250/0.239
B	16.0	55.0	880	0.400/0.383
C	24.0	55.0	1320	0.600/0.574
D	32.0	54.8	1760	0.800/0.765
E	44.0/47.9	50.0/48.0	2200/2300	1.000
F	var	var	var	var

Note: In table entries with split values, the first value is for four-lane freeways, and the second is for six- and eight-lane freeways.

(A)

PEDESTRIAN WALKWAYS

Level of Service	Space (sq ft/ped)	Expected Flows and Speeds		
		Ave. Speed, *S* (ft/min)	Flow Rate, *v* (ped/min/ft)	Vol/Cap Ratio, *v/c*
A	≥130	≥260	≤ 2	≤0.08
B	≥ 40	≥250	≤ 7	≤0.28
C	≥ 24	≥240	≤10	≤0.40
D	≥ 15	≥225	≤15	≤0.60
E	≥ 6	≥150	≤25	≤1.000
F	< 6	<150	—Variable—	

*Average conditions for 15 min.

(B)

FIGURE 10.2.3 Example level of service characteristics (*Source*: Transportation Research Board. 1994. *Highway Capacity Manual*. Washington, D.C.).

SIGNALIZED INTERSECTIONS

Level of Service	Stopped Delay per Vehicle (sec)
A	≤5.0
B	5.1 to 15.0
C	15.1 to 25.0
D	25.1 to 40.0
E	40.1 to 60.0
F	>60.0

(C)

ARTERIAL ROADS

Arterial Class	I	II	III
Range of Free Flow Speeds (mph)	45 to 35	35 to 30	35 to 25
Typical Free Flow Speed (mph)	40 mph	33 mph	27 mph
Level of Service	Average Travel Speed (mph)		
A	≥35	≥30	≥25
B	≥28	≥24	≥19
C	≥22	≥18	≥13
D	≥17	≥14	≥ 9
E	≥13	≥10	≥ 7
F	<13	<10	< 7

(D)

FIGURE 10.2.3 (continued) Example level of service characteristics.

Defining Terms

Design standard: Physical characteristics of a proposed facility that are professionally accepted and often based on safety considerations.

Functional classification: Classifying a transportation facility based on the function it serves in the transportation system. Such classification becomes important in that design standards are often directly related to the functional classification of a facility.

Level of service: An assessment of the performance of a transportation facility based on measurable physical characteristics (e.g., vehicular speed, average delay, density, flow rate). Level of service is usually subjectively defined as ranging from A (good performance) to F (bad or heavily congested performance).

Project development process: The steps that are followed to take a project from initial concept to final engineering. This process includes not only the detailed engineering associated with a project design but also the interaction with the general public and with agencies having jurisdiction over some aspect of project design.

References

American Association of State Highway and Transportation Officials. 1990. *A Policy on the Geometric Design of Highways and Streets.* AASHTO, Washington, DC.

RECOMMENDED ROADWAY SECTION WIDTHS

Functional Class	U/R	Number of Lanes	Travel Lane	Shoulder	
				Right	Left[1]
Freeway	Urban	4–8	12	10	4[2]
	Rural	4–8	12	10	4[2]
Arterial	Urban	Multilane with median	12	10	4
	Urban	Multilane without median	11–12	8–10[3]	N/A
	Rural	2 lane	12	See Table 5.2	N/A
	Rural	Multilane with median	12	8–10	4

(A)

WIDTH OF USABLE SHOULDER—EACH SIDE OF TRAVEL WAY
RURAL TWO-LANE ARTERIAL

	Design Traffic Volume				
	Current ADT Under 400	Current ADT Over 400	DHV 100–200	DHV 200–400	DHV Over 400
All design speeds	4 ft	6 ft	6 ft	8 ft	10 ft

(B)

RECOMMENDED WIDTH OF TRAVEL WAY AND GRADED SHOULDER
RURAL COLLECTOR

	Design Traffic Volume				
Design Speed (mph)	Current ADT Under 400	Current ADT Over 400	DHV 100–200	DHV 200–400	DHV Over 400
30	20 ft	20 ft	20 ft	22 ft	24 ft
40	20 ft	22 ft	22 ft	22 ft	24 ft
50	20 ft	22 ft	22 ft	24 ft	24 ft
60	22 ft	22 ft	22 ft	24 ft	24 ft
Graded Shoulder (Each Side)*					
All speeds	2 ft	4 ft	6 ft	8 ft	8 ft

(C)

* If right-of-way permits.

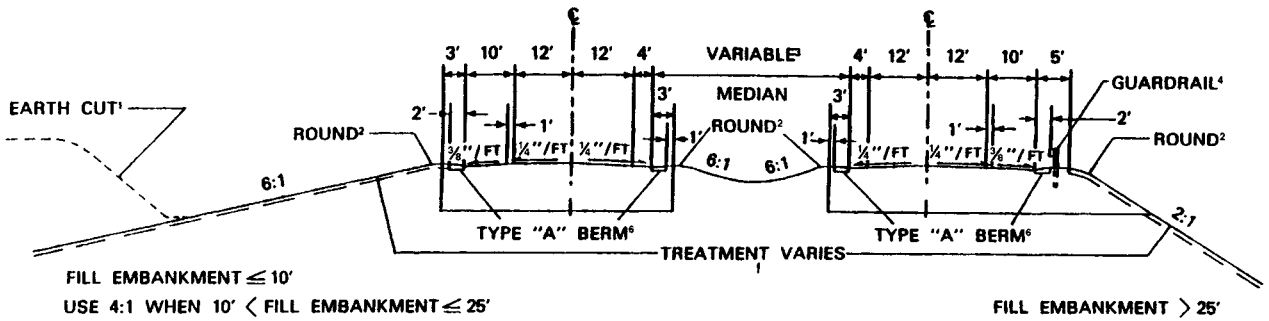
FIGURE 10.2.4 Example design criteria. (*Source*: Massachusetts Department of Public Works. 1988. *Highway Design Manual.* Boston, MA).

Mannering, F. and Kilareski, W. 1990. *Principles of Highway Engineering and Traffic Analysis.* John Wiley & Sons, New York.

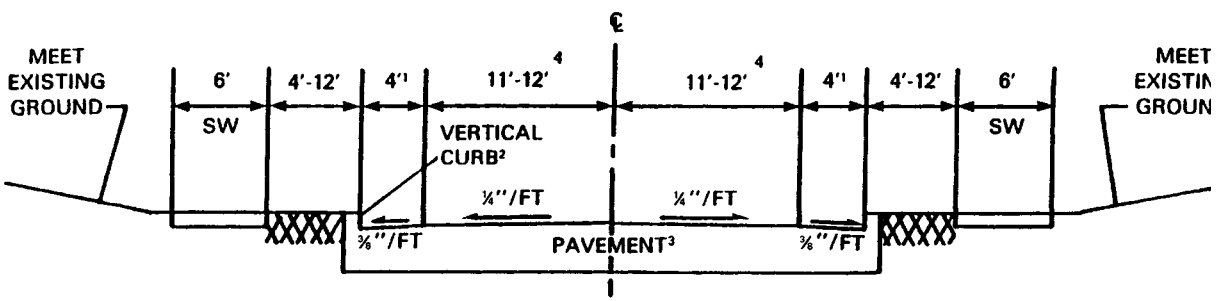
Massachusetts Department of Public Works. 1988. *Highway Design Manual.* MDPW, Boston, MA.

Transportation Research Board. 1994. *Highway Capacity Manual.* Special Report 209, TRB, Washington, DC.

Wright, P. and Ashford, N. 1989. *Transportation Engineering, Planning and Design.* John Wiley & Sons, New York.



Typical Freeway/Expressway Section



Typical Urban Collector and Local Road

FIGURE 10.2.5 Example cross sections.

Further Information

Transportation Research Board, National Academy of Sciences, 2101 Constitution Ave., N.W., Washington, D.C. 20418, http://www.nas.edu/trb/.

American Association of State Highway and Transportation Officials, 444 N. Capitol St., N.W., Suite 225, Washington, D.C. 20001.

Institute of Transportation Engineers, 525 School St., S.W., Suite 410, Washington, D.C. 20024, http://www.ite.org/.

Intelligent Transportation Society of America, 400 Virginia Avenue, S.W., Suite 800, Washington, D.C., 20024-2730, http://www.itsa.org/.

10.3 Operations and Environmental Impacts

Paul W. Shuldiner and Kenneth B. Black

Highways are intended to provide for the safe and efficient movement of vehicles. These objectives are secured through careful geometric design of the layout and physical features of roads and streets and through appropriate operational procedures and devices designed to guide, advise, and regulate vehicle operators in their use of these facilities. Proper design of operational procedures and devices is based on a thorough understanding of the complex interactions among driver, vehicle, and roadway characteristics. Provided here is an overview of the fundamentals upon which such an understanding can be built.

Fundamental Equations

Flow, q, is given by the following formula:

$$q = \frac{n \times 3600}{T} \text{ vehicles per hour} \tag{10.3.1}$$

where n = number of vehicles passing a point in T seconds. **Density**, k, is given by

$$k = \frac{n}{L} \times 5280 \text{ vehicles per mile} \tag{10.3.2}$$

where n = number of vehicles occupying a length of road and L is in feet, or by

$$k = \frac{n}{L} \times 1000 \text{ vehicles per kilometer} \tag{10.3.3}$$

where n = number of vehicles occupying a length of road and L is in meters.

Average **speed**, $\bar{u}$, is given by

$$\bar{u}_t = \frac{1}{n}\sum_{i=1}^{n} u_i \tag{10.3.4}$$

where u_i = speed of the ith vehicle, termed the *time mean speed*, and by

$$\bar{u}_s = ns \Big/ \sum_{i=1}^{n} t_i \tag{10.3.5}$$

where t_i = time for the ith vehicle to traverse distance s, termed the *space mean speed*. The values $\bar{u}_s$ and $\bar{u}_t$ usually differ very slightly from one another unless there are wide variations among the speeds of individual vehicles.

Flow, Speed, and Density Relationships

There exists a dimensional identity between flow, density, and speed such that

$$q = k \times \bar{u}_s \tag{10.3.6}$$

For simplicity, it is often assumed that average speed, u, decreases linearly as density, k, increases. Given this assumption and the dimensional identity expressed in Equation (10.3.6), it follows that flow, density, and speed are related as shown in Figure 10.3.1. It may be seen from Figure 10.3.1 that:

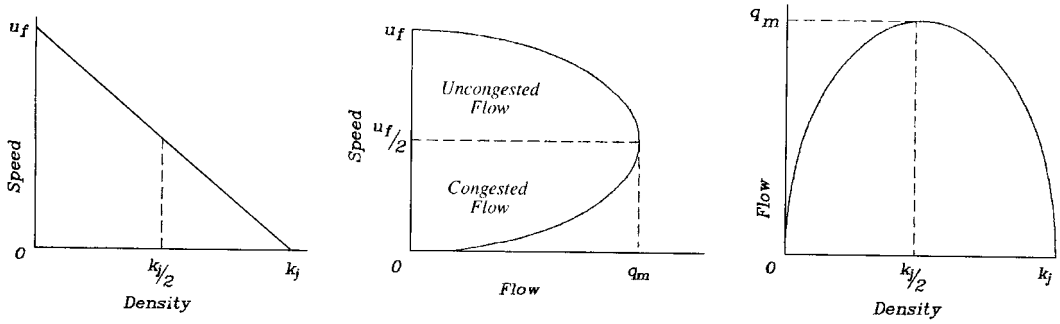


FIGURE 10.3.1 Relationships between flow, density, and speed.

1. When density on the highway is zero, there are no vehicles on the highway and the flow is zero.
2. As density rises to its maximum, flow goes to zero as vehicles line up—essentially bumper to bumper.
3. As density increases, flow also increases—up to the point where $k = kj/2$—and then decreases as density increases.
4. As flow increases from zero to a maximum (the capacity of the roadway), speed also decreases. Once the flow approaches capacity, any impediment to the steady movement of traffic will reduce both flow and speed abruptly, with a concomitant rapid increase in density.

Traffic Measurements

Measurements of traffic operating conditions are useful in the design of streets and highways and of the devices and measures necessary to promote safe and efficient operation of these facilities. Several of the more common types of traffic measurements and their applications are summarized as follows:

Measurement Type	Typical Measurement Device	Common Applications
Volume, roadway	Pneumatic road tubes and counters	Design of new facilities or capacity additions
	Magnetic induction loops	Traffic control
Volumes, intersection movements	Pneumatic tubes and counters	Traffic control
	Manual recording	Intersection design
Speed	Radar	Traffic control
	Pneumatic road tubes	Level of service
	Magnetic induction loops	Congestion management
Travel time and delay	"Floating cars" in traffic stream	Level of service
	License plate matching using laptop computers, pencil and paper, video camcorders	Congestion management
Accident records	Standardized police reports	Traffic control
	Condition and collision diagrams	Geometric improvements

Level of Service (LOS)

The **level of service** is a measure of how freely a given traffic system is operating. It is affected primarily by traffic **volume**, but other factors such as proportion of heavy vehicles, roadway geometrics, and

incidents also affect the level of service. LOS is designated by letters A through F. LOS A is characterized as *free flow* and is applied to conditions of very low traffic volume in which there is little or no restriction to traffic flow. LOS F is characterized as *forced flow* with congested, stop-and-go movement. LOS C involves some restriction to free movement of traffic and is the level of service commonly used as a design criterion.

Highway Capacity

The **capacity** of a highway depends on many geometric and traffic characteristics, the most important of which are number of lanes, width of lanes, lateral clearance to obstructions, percentage of the traffic stream made up of trucks and buses, and percent grade of the roadway. Ideally, lane widths should be at least 12 feet. Obstruction should be a minimum of 6 feet from the edge of pavement. Vertical grades of 3% or more that are one-half mile long or longer will reduce the capacity of a highway, especially where commercial vehicles are present. For freeways, maximum lane capacity under ideal conditions is generally assumed to be 2000 vehicles per hour. Ideal conditions, however, are seldom achieved.

A more useful number is the service flow. This is the ideal flow with several adjustment factors applied:

$$q_s = q^* N f_w f_{hv} \tag{10.3.7}$$

where

q_s = Service flow in vehicles per hour
q^* = Ideal flow in passenger cars per lane per hour
N = Number of freeway lanes
f_w = Adjustment factor for effect of lane widths and close lateral obstructions
f_{hv} = Adjustment factor for effect of heavy vehicles

See TRB (1985) for appropriate values of f_w and f_{hv}.

Intersection Capacity

The 1985 *Highway Capacity Manual* provides a quick method for estimating the capacity of a signalized intersection. This method is based on critical movement analysis and is known as the *planning procedure*. To apply this procedure, the sum of a through lane of traffic is added to the opposing left-turn traffic, and the highest resulting sum for both the north-south and east-west directions is considered the critical volume for an intersection. These volumes are compared with threshold values to establish the intersection capacity. These thresholds are presented in TRB (1985).

The analysis of capacity of unsignalized intersections is done by studying the degree to which motorists will accept gaps in traffic when they pull away from a stop or yield sign. The critical gap is the gap, in seconds, that 50% of the drivers will accept. The potential capacity minus the used capacity of a lane is the reserve capacity. It is this capacity that is the basis of establishing the level of service of the unsignalized intersection.

Traffic Control Devices

Traffic signs, signals, and pavement markings **(traffic control devices)** are used to regulate, warn, and guide traffic in the interest of promoting safe and efficient operation. The effectiveness of such devices is greatly enhanced through the uniformity of their design, including shape, color, and message, and through their use and placement on the roadway. Uniformity is ensured by adherence to the standards promulgated in the *Manual of Uniform Traffic Control Devices* (U.S. DOT and FHA, 1988) or the equivalent manual adopted by each state highway agency.

Stop Sign Warrants

Since stop signs are an inconvenience to motorists, they should only be used where one or more of the following conditions warrants.

1. At an intersection of a less important road with a main road where application of the normal right-of-way rule is unduly hazardous
2. Where a street enters a through highway or street
3. At an unsignalized intersection in a signalized area
4. At other intersections where a combination of high speed, restricted view, and serious accident record indicates a need for control by a stop sign

Consideration should be given to other less restrictive measures before these warrants are applied.

Traffic Signal Warrants

Traffic control signals should not be installed unless at least one of several signal warrants is met. Eleven signal warrants are used to account for such factors as traffic volume, accidents, and pedestrian volumes. The description of these warrants is presented in U.S. DOT and FHA (1988). In addition to a review of these warrants, an engineering study should be conducted to show that the installation of such a signal would improve the overall safety and operation of the intersection.

Air Quality Effects

Air quality effects of vehicles are a function of the number, type, and condition of vehicles on the highway and of the speed and efficiency with which these vehicles operate. Carbon monoxide, hydrocarbons, and oxides of nitrogen are the primary pollutants emitted by vehicular traffic. The level of emission of each pollutant is complicated and difficult to predict. In general, the emission of carbon monoxide and hydrocarbons is minimized at speeds between 35 and 45 miles per hour, whereas the emission of oxides of nitrogen increases over this speed range. Therefore, there is no ideal speed at which the level of emission of all three pollutants will be minimized. Nevertheless, smoothly flowing traffic will tend to produce less pollution, and therefore, whatever can be done to reduce delay and congestion—especially when combined with a reduction in total vehicle miles traveled—will result in an improvement in air quality.

Defining Terms

Capacity: A measure of a facility's ability to accommodate the movement of vehicles.

Density (k): Number of vehicles occupying a given length of roadway at an instant in time; usually expressed as veh/mi or veh/km.

Level of service (LOS): The quality of movement experienced by motorists in a traffic stream; usually expressed in terms of the freedom of movement of motorists, ranging from little or no restriction (LOS A) to essentially bumper-to-bumper congestion (LOS F).

Speed (u): Average speed of vehicles in a traffic stream; usually expressed as miles per hour (mph) or kilometers per hour (kph).

Traffic control devices: Signs, signals, and pavement markings placed on streets or highways to guide, inform, or regulate the movement of vehicles on those facilities.

Volume (q): Time rate of flow of vehicles; usually expressed as veh/h.

References

Baerwald, J.E. (Ed.) 1990. *Transportation and Traffic Engineering Handbook*. Prentice Hall, Englewood Cliffs, NJ.

Carter, E.C. and Homburger, W.S. 1978. *Introduction to Transportation Engineering*. Reston, VA.

Garber, N.J. and Hoel, L.A. 1988. *Traffic and Highway Engineering*, West, St. Paul, MN.

Mannering, F.L. and Kilareski, W.P. 1990. *Principles of Highway Engineering and Traffic Analysis*. John Wiley & Sons, New York.

McShane, W.R. and Ross, R.P. 1990. *Traffic Engineering*. Prentice Hall, Englewood Cliffs, NJ.

Pline, J.L. (Ed.) 1992. *Traffic Engineering Handbook*, 4th ed. Prentice Hall, Englewood Cliffs, NJ.

TRB. 1985. *Highway Capacity Manual*. Special Report 209, Transportation Research Board, Washington, DC.

U.S. DOT and FHA. 1988. *Manual of Uniform Traffic Control Devices*. U.S. Department of Transportation, Federal Highway Administration, Washington, DC.

Further Information

Transportation Research Board
National Research Council
2101 Constitution Ave. NW
Washington, DC 20418
202-334-2934
Fax 202-334-2003

American Society of Civil Engineers
United Engineering Center
345 East 47th Street
New York, NY 10017-2398
212-705-7496
Fax 212-980-4681

Institute of Transportation Engineers
525 School Street, SW
Suite 410
Washington, DC 20014-2797
202-544-8050
Fax 202-863-5486

10.4 Transportation Systems

Paul Schonfeld

The various forms of transportation that have been developed over time are called **modes**. The classification of modes may be very broad (e.g., highway transportation or air transportation) or more restrictive (e.g., chartered helicopter service). The major distinctions among transportation modes that help to classify them include:

1. Medium (e.g., air, space, surface, underground, water, underwater)
2. Users (e.g., passengers vs. cargo, general-purpose vs. special trips or commodities, common vs. private carrier)
3. Service type (scheduled vs. **demand responsive**, fixed vs. variable route, nonstop vs. express or local, mass vs. personal)
4. Right-of-way type (exclusive, semi-exclusive, shared)
5. Technology:
   a. Propulsion (e.g., electric motors, diesel engines, gas turbines, linear induction motors, powered cables)
   b. Energy sources (e.g., petroleum fuels, natural gas, electric batteries, electric power from conducting cables)
   c. Support (e.g., aerodynamic lift, flotation on water, steel wheels on two steel rails, monorails, air cushions, magnetic levitation, suspension from cables)
   d. Local control (e.g., lateral control by steering wheels, wheel flanges on railroad vehicles, rudders, longitudinal control by humans or automatic devices)
   e. Network guidance and control systems (with various degrees of automation and optimization)

A mode may be defined by its combination of such features. The number of conceivable combinations greatly exceeds the number of modes that have been actually tried, which, in turn, exceeds the number of successful modes. Success may be limited to relatively narrow markets and applications (e.g., for helicopters or aerial cablecars) or may be quite general. Thus, automobiles are successful in a very broad range of applications and have become the basis for distinct transportation modes such as taxis, carpools, or ambulances.

The relative success of various transportation modes depends on available technology and socioeconomic conditions at any particular time, as well as on geographic factors. As technology or socioeconomic conditions change, new transportation modes appear, develop, and may later decline as more effective competitors appear. For many centuries water transportation was considerably cheaper than overland transportation. Access to waterways was quite influential in the location of economic activities and cities. Access to good transportation is still very important to industries and communities. Technological developments have so drastically improved the relative effectiveness of air transportation that within a short period (approximately 1950 to 1965) aircraft almost totally replaced ships for transporting passengers across oceans. It is also notable that as economic prosperity grows, personal transportation tends to shift from the walking mode to bicycles, motorcycles, and then automobiles. Geography can significantly affect the relative attractiveness of transportation modes. Thus, natural waterways are highly valuable where they exist. Hilly terrain decreases the economic competitiveness of artificial waterways or conventional railroads while favoring highway modes. In very mountainous terrain even highways may become uncompetitive compared to alternatives such as helicopters, pipelines, and aerial cablecars.

The relative shares of U.S. intercity passenger and freight traffic are shown in Table 10.4.1. The table shows the relative growth since 1929 of airlines, private automobiles, and trucks and the relative decline of railroad traffic.

TABLE 10.4.1 Volume of U.S. Intercity Freight and Passenger Traffic

Millions of Revenue Freight Ton-Miles and Percentage of Total

Year	Railroads	%	Trucks	%	Great Lakes	%	Rivers and Canals	%	Oil Pipelines	%	Air	%	Total
1929	454,800	74.9	19,689	3.2	97,322	16.0	8,661	1.4	26,900	4.4	3	0.0	607,375
1939	338,850	62.3	52,821	9.7	76,312	14.0	19,937	3.7	55,602	10.2	12	0.0	543,534
1944	746,912	68.6	58,264	5.4	118,769	10.9	31,386	2.9	132,864	12.2	71	0.0	1,088,266
1950	596,940	56.2	172,860	16.3	111,687	10.5	51,657	4.9	129,175	12.2	318	0.0	1,062,637
1960	579,130	44.1	285,483	21.7	99,468	7.6	120,785	9.2	228,626	17.4	0.1	0.0	1,314,270
1970	771,168	39.8	412,000	21.3	114,475	5.9	204,085	10.5	431,000	22.3	3,295	0.2	1,936,023
1980	932,000	37.5	555,000	22.3	96,000	3.9	311,000	12.5	588,000	23.6	4,840	0.2	2,486,840
1986	889,000	35.5	634,000	25.3	68,000	2.7	325,000	13.0	578,000	23.1	7,340	0.3	2,501,340
1987	968,000	36.3	668,000	25.1	78,000	2.9	358,000	13.4	585,000	22.0	8,720	0.3	2,665,720

Millions of Revenue Passenger-Miles and Percentage of Total (Except Private)

Year	Railroads	%	Buses	%	Air Carriers	%	Inland Waterways	%	Total (except private)	Private Automobiles	Private Airplanes	Total (including private)
1929	33,965	77.1	6,800	15.4	—	—	3,300	7.5	44,065	175,000	—	219,065
1939	23,669	67.7	9,100	26.0	683	2.0	1,486	4.3	34,938	275,000	—	309,938
1944	97,705	75.7	26,920	20.9	2,177	1.7	2,187	1.78	128,989	181,000	1	309,990
1950	32,481	47.2	26,436	38.4	8,773	12.7	1,190	1.7	68,880	438,293	1,299	508,472
1960	21,574	28.6	19,327	25.7	31,730	42.1	2,688	3.6	75,319	706,079	2,228	783,626
1970	10,903	7.3	25,300	16.9	109,499	73.1	4,000	2.7	149,702	1,026,000	9,101	1,184,803
1980	11,000	4.5	27,400	11.3	204,400	84.2	NA	—	242,800	1,300,400	14,700	1,557,900
1986	11,800	3.4	23,700	6.9	307,900	89.7	NA	—	343,400	1,450,100	12,400	1,805,900
1987	12,300	3.4	22,800	6.2	329,100	90.4	NA	—	364,200	1,494,900	12,400	1,871,500

Note: Railroads includes all classes, including electric railways, Amtrak and Auto-Train.
Source: Transportation Policy Associates.
Source: Railroad facts, 1988 Edition, Association of American Railroads.

Transportation System Components

The major components of transportation systems are:

1. Links
2. Terminals
3. Vehicles
4. Control systems

Certain "continuous-flow" transportation systems such as pipelines, conveyor belts, and escalators have no discrete vehicles and, in effect, combine the vehicles with the link.

Transportation systems may be developed into extensive networks. The networks may have a hierarchical structure. Thus, highway networks may include freeways, arterials, collector streets, local streets, and driveways. Links and networks may be shared by several transportation modes (e.g., cars, buses, trucks, taxis, bicycles, and pedestrians on local streets). Exclusive lanes may be provided for particular modes (e.g., pedestrian or bicycles) or groups of modes (e.g., buses and carpools).

Transportation terminals provide interfaces among modes or among vehicles of the same mode. They may range from marked bus stops or truck loading zones on local streets to huge airports or ports.

Evaluation Measures

Transportation systems are evaluated in terms of their effects on their suppliers, users, and environment. Both their costs and benefits may be classified into supplier, user, and external components. Private transportation companies normally seek to maximize their profits (i.e., total revenues minus total supplier costs). Publicly owned transportation agencies should normally maximize net benefits to their jurisdictions, possibly subject to financial constraints.

From the supplier's perspective, the major indicators of performance include measures of **capacity** (maximum throughput), **speed, utilization rate** (i.e., fraction of time in use), **load factor** (i.e., fraction of maximum payload actually used), energy efficiency (e.g., Btu per ton-mile or per passenger mile), and labor productivity (e.g., worker hours per passenger mile or per ton-mile). Measures of environmental impact (e.g., noise decibels or parts of pollutant per million) are also increasingly relevant. To users, price and service quality measure, including travel time, wait time, access time, reliability, safety, comfort (ride quality, roominess), simplicity of use, and privacy are relevant in selecting modes, routes, travel times, and suppliers.

Air Transportation

Air transportation is relatively recent, having become practical for transporting mail and passengers in the early 1920s. Until the 1970s its growth was paced primarily by technological developments in propulsion, aerodynamics, materials, structures, and control systems. These developments have improved its speed, load capacity, energy efficiency, labor productivity, reliability, and safety, to the point where it now dominates long-distance mass transportation of passengers overland and practically monopolizes it over oceans. Airliners have put ocean passenger liners out of business because they are much faster and also, remarkably, more fuel and labor efficient. However, despite this fast growth, the cargo share of air transportation is still small.

For cargoes that are perishable, high in value, or urgently needed, air transportation is preferred over long distances. For other cargo types, ships, trucks, and railroads provide more economical alternatives. In the 1990s, the nearest competitors to air cargo are containerships over oceans and trucks over land. For passengers, air transportation competes with private cars, trains, and intercity buses over land, with practically no competitors over oceans. The growth of air transportation has been restricted to some extent by the availability of adequate airports, by environmental concerns (especially noise), and by the fear of flying of some passengers.

There are approximately 10,000 commercial jet airliners in the world, of which the largest (as of 1997) are Boeing B-747 types, of approximately 800,000 lb gross takeoff weight, with a capacity of 550 passengers. The economical cruising speed of these and smaller "conventional" (i.e., **subsonic**) airliners has stayed at around 560 mph since the late 1950s. A few supersonic transports (SSTs) capable of cruising at approximately 1300 mph were built in the 1970s (the Anglo-French Concorde and the Soviet Tu-144) but, due to high capital and fuel costs, were not economically successful. About eight Concorde SSTs are still operating, with government subsidies.

The distance that an aircraft can fly depends on its payload, according to the following equation:

$$R = \frac{V}{c'}\left(\frac{L}{D}\right)\ln\left(W_{TO}/W_L\right) \tag{10.4.1}$$

where

R = range (mi)
c' = specific fuel consumption (lb fuel/lb thrust × h)
(L/D) = lift-to-drag ratio (dimensionless)
W_{TO} = aircraft takeoff weight (lb) = $W_L + W_F$
W_L = aircraft landing weight (lb) = $W_E + W_R + P$
W_E = aircraft empty weight (lb)
W_R = reserve fuel weight (lb)
W_F = consumed fuel weight (lb)
P = payload (lb)

This equation assumes that the difference between the takeoff weight and landing weight is the fuel consumed. For example, suppose that for a Boeing B-747 the maximum payload carried (based on internal fuselage volume and structural limits) is 260,000 lb, maximum W_{TO} is 800,000 lb, W_R = 15,000 lb, W_E = 370,000, L/D = 17, V = 580 mph, and c' = 0.65 lb/lb thrust × h. The resulting weight ratio [W_{TO}/W_L = 800/(370 + 15 + 260)] is 1.24 and the range R is 3267 mi. Payloads below 260,000 lb allow greater ranges.

Most airline companies fly scheduled routes, although charter services are common. U.S. airlines are largely free to fly whatever routes (i.e., origin–destination pairs) they prefer in the United States. In most of the rest of the world, authority to serve particular routes is regulated to negotiated by international agreements. The major components of airline costs are direct operating costs (e.g., aircraft depreciation or rentals, aircrews, fuel, and aircraft maintenance) and indirect operating costs (e.g., reservations, advertising and other marketing costs, in-flight service, ground processing of passengers and bags, and administration).

The efficiency and competitiveness of airline service is heavily dependent on efficient operational planning. Airline scheduling is a complex problem in which demand at various times and places, route authority, aircraft availability and maintenance schedules, crew availability and flying restrictions, availability of airport gates and other facilities, and various other factors must all be considered. Airline management problems are discussed in (Wells, 1984).

Airports range from small unmarked grass strips to major facilities requiring many thousands of acres and billions of dollars. Strictly speaking, an airport consists of an airfield (or "airside") and terminal (or "landside"). Airports are designed to accommodate specified traffic loads carried by aircraft up to a "design aircraft," which is the most demanding aircraft to be accommodated. The design aircraft might determine such features as runway lengths, pavement strengths, or terminal gate dimensions at an airport. Detailed guidelines for most aspects of airport design (e.g., runway lengths and other airfield dimensions, pavement characteristics, drainage requirements, allowable noise and other environmental impacts, allowable obstruction heights, lighting, markings, and signing) are specified by the U.S. Federal Aviation Administration (FAA) in a series of circulars.

Airport master plans are prepared to guide airport growth, usually in stages, toward ultimate development. These master plans:

1. Specify the airport's requirements.
2. Indicate a site if a new airport is considered.
3. Provide detailed plans for airport layout, land use around the airport, terminal areas, and access facilities.
4. Provide financial plans, including economic and financial feasibility analysis.

Major new airports tend to be very expensive and very difficult to locate. Desirable airport sites must be reasonably close to the urban areas they serve yet far enough away to ensure affordable land and acceptable noise impacts. Many other factors—including airspace interference with other airports, obstructions (e.g., hills, buildings), topography, soil, winds, visibility, and utilities—must be reconciled. Hence, few major new airports are being built, and most airport engineering and planning work in the United States is devoted to improving existing airports. Governments sometimes develop multi-airport system plans for entire regions or countries.

National agencies (such as the FAA in the United States) are responsible for traffic control and airspace management. Experienced traffic controllers, computers, and specialized sensors and communication systems are required for this function. Increasingly sophisticated equipment has been developed to maintain safe operations even for crowded airspace and poor visibility conditions. For the future we can expect increasing automation in air traffic control, relying on precise aircraft location with global positioning satellite (GPS) systems and fully automated landings. Improvements in the precision and reliability of control systems are increasing (slowly) the capacity of individual runways as well as the required separation among parallel runways, allowing capacity increases in restricted airport sites.

Railroad Tranportation

The main advantages of railroad technology are low frictional resistance and automatic lateral guidance. The low friction reduces energy and power requirements but limits braking and hill-climbing abilities. The lateral guidance provided by wheel flanges allows railroad vehicles to be grouped into very long trains, yielding economies of scale and, with adequate control systems, high capacities per track. The potential energy efficiency and labor productivity of railroads is considerably higher than for highway modes, but is not necessarily realized, due to regulations, managerial decisions, demand characteristics, or terrain.

The main competition of railroads include automobiles, aircraft, and buses for passenger transportation, and trucks, ships, and pipelines for freight transportation. To take advantage of their scale economies, railroad operators usually seek to combine many shipments into large trains. Service frequency is thus necessarily reduced. Moreover, to concentrate many shipments, rail cars are frequently re-sorted into different trains, rather than moving directly from origin to destination, which results in long periods spent waiting in classification yards, long delivery times, and poor vehicle utilization. An alternative operational concept relying on direct nonstop "unit trains" is feasible only when demand is sufficiently large between an origin–destination pair.

Substantial traffic is required to cover the relatively high fixed costs of railroad track. Moreover, U.S. railroads, which are privately owned, must pay property taxes on their tracks, unlike their highway competitors. By 1920 highway developments had rendered low-traffic railroad branch lines noncompetitive in the United States. Abandonment of such lines has greatly reduced the U.S. railroad network, even though the process was retarded by political regulation.

The alignment of railroad track is based on a compromise between initial costs and operating costs. The latter are reduced by a more straight and level alignment, which requires more expensive earthwork, bridges, or tunnels. Hay (1982) provides design guidelines for railroads.

In general, trains are especially sensitive to gradients. Thus, compared to highways, railroad tracks are more likely to go around rather than over terrain obstacles, which increases the **circuity factors** for railroad transportation.

The resistance for railroad vehicles may be computed using the Davis equation (Hay, 1982):

$$r = 1.3 + 29/w + bV + CAV^2/wn + 20G + 0.8D \tag{10.4.2}$$

where

- G = gradient (%)
- D = degree of curvature
- r = unit resistance (lb of force per ton of vehicle weight)
- w = weight (tons per axle of car or locomotive)
- n = number of axles
- b = coefficient of flange friction, swaying, and concussion (0.045 for freight cars and motor cars in trains, 0.03 for locomotives and passenger cars, and 0.09 for single-rail cars)
- C = drag coefficient of air [0.0025 for locomotives (0.0017 for streamlined locomotives) and single- or head-end-rail cars, 0.0005 for freight cars, and 0.00034 for trailing passenger cars, including rapid transit]
- A = cross-sectional area of locomotives and cars (usually 105 to 120 ft^2 for locomotives, 85 to 90 ft^2 for freight cars, 110-120 ft^2 for multiple-unit and passenger cars, and 70 to 110 ft^2 for single- or head-end-rail cars)
- V = speed (mph)

The coefficients shown for this equation reflect relatively old railroad technology and can be significantly reduced for modern equipment (Hay, 1982). The equation provides the unit resistance in pounds of force per ton of vehicle weight. The total resistance of a railroad vehicle (in lb) is

$$R_v = rwn \tag{10.4.3}$$

The total resistance of a train R is the sum of resistances for individual cars and locomotives. The rated horsepower (hp) required for a train is:

$$\text{hp} = \frac{RV}{375\eta} \tag{10.4.4}$$

where η = transmission efficiency (typically about 0.83 for a diesel electric locomotive).

The hourly fuel consumption for a train may be computed by multiplying hp by a specific fuel consumption rate (approximately 0.32 lb/hp × h for a diesel electric locomotive).

Diesel electric locomotives with powers up to 5000 hp haul most trains in the United States. Electric locomotion is widespread in other countries, especially those with low petroleum reserves. It is especially competitive on high-traffic routes (needed to amortize electrification costs) and for high-speed passenger trains. Steam engines have almost disappeared.

The main types of freight rail cars are box cars, flat cars (often used to carry truck trailers or intermodal containers), open-top gondola cars, and tank cars. Passenger trains may include restaurant cars and sleeping cars. Rail cars have tended toward increasing specialization for different commodities carried, a trend that reduces opportunities for back hauls. Recently, many “double-stack” container cars have been built to carry two tiers of containers. Such cars require a vertical clearance of nearly 20 ft, as well as reduced superelevation (banking) on horizontal curves. In the United States standard freight rail cars with gross weights up to 315,000 lb are used.

High-speed passenger trains have been developed intensively in Japan, France, Great Britain, Italy, Germany, and Sweden. The most advanced (in 1995) appear to be the latest French TGV versions, with cruising speeds of 186 mph and double-deck cars. At such high speeds, trains can climb long, steep grades (e.g., 3.5%) without slowing down much. Construction costs in hilly terrain can thus be significantly reduced. Even higher speeds are being tested in experimental railroad and magnetic levitation (MAGLEV) trains.

Highway Transportation

Highways provide very flexible and ubiquitous transportation for people and freight. A great variety of transportation modes, including automobiles, buses, trucks, motorcycles, bicycles, pedestrians, animal-drawn vehicles, taxis, and carpools, can share the same roads. From unpaved roads to multilane freeways, roads can vary enormously in their cost and performance. Some highway vehicles may even travel off the roads in some circumstances. The vehicles also range widely in cost and performance, and at their lower range (e.g., bicycles) are affordable for private use even in poor societies.

Flexibility, ubiquity, and affordability account for the great success of highway modes. Personal vehicles from bicycles to private automobiles offer their users great freedom and access to many economic and social opportunities. Trucks increase the freedom and opportunities available to farmers and small businesses. Motor vehicles are so desirable and affordable that in the United States the number of registered cars and trucks approximates the number of people of driving age. Other developed countries are approaching the same state despite strenuous efforts to discourage motor vehicle use.

The use of motor vehicles brings significant problems and costs. These include:

1. Road capacity and congestion. Motor vehicles require considerable road space, which is scarce in urban areas and costly elsewhere. Shortage of road capacity results in severe congestion and delays.
2. Parking availability and cost.
3. Fuel consumption. Motor vehicles consume vast amounts of petroleum fuels. Most countries have to import such fuels and are vulnerable to price increases and supply interruptions.
4. Safety. The numbers of people killed and injured and the property damages in motor vehicle accidents are very significant.
5. Air quality. Motor vehicles are major contributors to air pollution.
6. Regional development patterns. Many planners consider the low-density “sprawl” resulting from motor vehicle dominance to be inefficient and inferior to the more concentrated development produced by mass transportation and railroads.

In the United States trucks have steadily increased their share of the freight transportation market, mostly at the expense of railroads, as shown in Table 10.4.1. They can usually provide more flexible, direct, and responsive service than railroads, but at higher unit cost. They are intermediate between rail and air transportation in both cost and service quality. With one driver required per truck, the labor productivity is much lower than for railroads, and there are strong economic incentives to maximize the load capacity for each driver. Hence, the tendency has been to increase the number, dimensions, and weights allowed for trailers in truck-trailer combinations, which requires increased vertical clearances (e.g., bridge overpasses), geometric standards for roads, and pavement costs.

The main reference for highway design is the AASHTO manual [AASHTO, 1990]. For capacity, the main reference is the Transportation Research Board *Highway Capacity Manual* (TRB, 1985). Extensive software packages have been developed for planning, capacity analysis, geometric design, and traffic control.

Water Transportation

Water transportation may be classified into (1) marine transportation across seas and (2) inland waterway transportation; their characteristics differ very significantly. Inland waterways consist mostly of rivers, which may be substantially altered to aid transportation. Lakes and artificial canals may also be part of inland waterways. Rivers in their natural states are often too shallow, too fast, or too variable in their flows. All these problems may be alleviated by impounding water behind dams at various intervals. (This also helps generate electric power.) Boats can climb or descend across dams by using **locks** or other elevating systems (Hochstein, 1981). In the U.S. inland waterway network there are well over 100 major lock structures, with chambers up to 1200 ft long and 110 ft wide. Such chambers allow up to 18 large barges (35×195 ft) to be raised or lowered simultaneously.

In typical inland waterway operations, large diesel-powered "towboats" (which actually push barges) handle a rigidly tied group of barges (a "tow"). Tows with up to 48 barges (35×195 ft, or about 1300 tons per barge) are operated on the lower Mississippi, where there are no locks or dams. On other rivers, where locks are encountered at frequent intervals, tow sizes are adjusted to fit through locks. The locks constitute significant bottlenecks in the network, restricting capacity and causing significant delays.

Table 10.4.1 indicates that the waterway share of U.S. freight transportation has increased substantially in recent years. This is largely attributable to extensive improvements to the inland waterway system undertaken by the responsible agency, the U.S. Army Corps of Engineers.

The main advantage of both inland waterway and marine transportation is low cost. The main disadvantage is relatively low speed. Provided that sufficiently deep water is available, ships and barges can be built in much larger sizes than ground vehicles. Ship costs increase less than proportionally with ship size, for ship construction, crew, and fuel. Energy efficiency is very good at low speeds [e.g., 10–20 knots (nautical mi/h)]. However, at higher speeds the wave resistance of a conventional ship increases with the fourth power of speed (V^4). Hence, the fuel consumption increases with V^4 and the power required increases with V^5. Therefore, conventional-displacement ships almost never exceed 30 knots in commercial operation. Higher practical speed may be obtained by lifting ships out of the water on hyrdrofoils or air cushions. However, such unconventional marine vehicles have relatively high costs and limited markets at this time, ships have increased in size and specialization. Crude oil tankers of up to 550,000 tons (of payload) have been built. Tankers carrying fluids are less restricted in size than other ships because they can pump their cargo from deep water offshore without entering harbors. Bulk carriers (e.g., for coal, minerals, or grains) have also been built in sizes exceeding 300,000 tons. They may also be loaded through conveyor belts built over long pier structures to reach deep water. General cargo ships and containerships are practically always handled at shoreline berths and require much storage space nearby.

The use of intermodal containers has revolutionized the transportation of many cargoes. Such containers greatly reduce the time and cost required to load and unload ships. Up to 4500 standard 20-ft containers ($20 \times 8 \times 8$ ft) can be carried at a speed of about 24 knots on recently built containerships.

Port facilities for ships should provide shelter from waves and sufficiently deep water, including approach channels to the ports. In addition, ports should provide adequate terminal facilities, including loading and unloading equipment, storage capacity, and suitable connections to other transportation networks. Ports often compete strenuously with other ports and strive to have facilities that are at least equal to those of competitors. Since ports generate substantial employment and economic activities, they often receive financial and other support from governments.

Geography limits the availability of inland waterways and the directness of ship paths across oceans. Major expensive canals (e.g., Suez, Panama, Kiel) have been built to provide shortcuts in shipping routes. These canals may be so valuable that ship dimensions are sometimes compromised (i.e., reduced) to fit through these canals. In some parts of the world (e.g., Baltic Sea, North Sea, most U.S. coasts) the waters are too shallow for the largest ships in existence. Less efficient, smaller ships must be used there.

The dredging of deeper access channels and ports can increase the allowable ship size, if the costs and environmental impacts are considered acceptable.

Public Transportation

Public transportation is the term for ground passenger transportation modes available to the general public. It connotes public availability rather than ownership. "Conventional" public transportation modes have fixed routes *and* fixed schedules and include most bus and rail transit services. "Unconventional" modes (also labeled "paratransit") include taxis, carpools and van pools, rented cars, dial-a-ride services, and subscription services.

The main purposes of public transportation services, especially conventional mass transportation services in developed countries, are to provide mobility for persons without automobiles (e.g., children, poor, nondrivers); to improve the efficiency of transportation in urban areas; to reduce congestion effects, pollution, accidents, and other negative impacts of automobiles; and to foster preferred urban development patterns (e.g., strong downtowns and concentrated rather than sprawled development).

Conventional services (i.e., bus and rail transit networks) are quite sensitive to demand density. Higher densities support higher service frequencies and higher network densities, which decrease user wait times and access times, respectively. Compared to automobile users, bus or rail transit users must spend extra time in access to and from stations and in waiting at stations (including transfer stations). Direct routes are much less likely to be available, and one or more transfers (with possible reliability problems) may be required. Thus, mass transit services tend to be slower than automobiles unless exclusive rights-of-way (e.g., bus lanes, rail tunnels) can favor them. Such exclusive rights-of-way can be quite expensive if placed on elevated structures or in tunnels. Even when unhindered by traffic, average speeds may be limited by frequent stops and allowable acceleration limits for standing passengers. Prices usually favor mass transit, especially if parking for automobiles is scarce and expensive.

The capacity of a transit route can be expressed as:

$$C = FLP \tag{10.4.5}$$

where

C = one-way capacity (passengers/hour) past a certain point
F = service frequently (e.g., trains/hour)
L = train length (cars/train)
P = passenger capacity of cars (spaces/car)

For rail transit lines where high capacity is needed in peak periods, C can reach 100,000 passengers/hour (i.e., 40 trains/hour × 10 cars/train × 250 passenger spaces/car). There are few places in the world where such capacities are required. For a bus line the train length L would usually be 1.0. If no on-line stops are allowed, an exclusive bus lane also has a large capacity (e.g., 1000 buses/hour × 90 passenger spaces/bus), but such demand levels for bus lanes have not been observed.

The average wait time of passengers on a rail or bus line depends on the headways, which is the interval between successive buses or trains. This can be approximated by:

$$\overline{W} = \overline{H}/2 + \text{var}(H)/2\overline{H} \tag{10.4.6}$$

where

$\overline{W}$ = average wait time (e.g., minutes)
$\overline{H}$ = average headway (e.g., minutes)
var(H) = variance of headway (e.g., minutes^2)

It should be noted that the headway is the inverse of the service frequency.

The number of vehicles N required to serve a route is:

$$N = RFL \tag{10.4.7}$$

where R = vehicle round trip time on route (e.g., hours).

The effectiveness of a public transportation system depends on many factors, including demand distribution and density, network configuration, routing and scheduling of vehicles, fleet management, personnel management, pricing policies, and service reliability. Demand and economic viability of services also depend on how good and uncongested the road system is for automobile users.

Engineers can choose from a great variety of options for propulsion, support, guidance and control, vehicle configurations, facility designs, construction methods, and operating concepts. New information and control technology can significantly improve public transportation systems. It will probably foster increased automation and a trend toward more personalized (i.e., taxilike) service rather than mass transportation.

Defining Terms

Capacity: The maximum flow rate that can be expected on a transportation facility. "Practical" capacity is sometimes limited by "acceptable:" delay levels, utilization rates, and load factors.

Circuity factor: Ratio of actual distance on network to shortest airline distance.

Delay: Increase in service time due to congestion or service interruptions.

Demand-responsive: A mode whose schedule or route is adjusted in the short term as demand varies, such as taxis, charter airlines, and "TRAMP" ships.

Load factor: Fraction of available space or weight-carrying capability that is used.

Lock: A structure with gates at both ends which is used to lift or lower ships or other vessels.

Modes: Distinct forms of transportation.

Subsonic: Flying below the speed of sound (Mach 1), which is approximately 700 mph at cruising altitudes of approximately 33,000 ft.

Utilization rate: Fraction of time that a vehicle, facility, or equipment unit is in productive use.

References

AASHTO (American Society of State Highway and Transportation Officials). 1990. *A Policy on Geometric Design of Highways and Streets.* Washington, DC.

Brun, E. 1981. *Port Engineering.* Gulf Publishing, Houston.

Hay, W.W. 1982. *Railroad Engineering*. John Wiley & Sons, New York.

Hochstein, A. 1981. *Waterways Science and Technology.* Final Report DACW 72-79-C-003. U.S. Army Corps of Engineers, August.

Homburger, W.S. 1982. *Transportation and Traffic Engineering Handbook.* Prentice-Hall, Englewood Cliffs, NJ.

Horonjeff, R. and McKelvey, F. 1994. *Planning and Design of Airports*. McGraw-Hill, New York.

Morlok, E.K. 1976. *Introduction to Transportation Engineering and Planning.* McGraw-Hill, New York.

TRB (Transportation Research Board). 1985. *Highway Capacity Manual.* Special Report 209. TRB, Washington, DC.

Vuchic, V. 1981. *Urban Public Transportation.* Prentice-Hall, Englewood Cliffs, NJ.

Wells, A.T. 1984. *Air Transportation.* Wadsworth Publishing Co., Belmont, CA.

Wright, P.H. and Paquette, R.J. 1987. *Highway Engineering.* John Wiley & Sons, New York.

Further Information

The ITE Handbook (Homburger, 1992) and Morlok (1978) cover most transportation modes. Horonjeff and McKelvey (1994), Hay (1982). Wright and Paquette (1987), Brun (1981), and Vuchic (1981) are more specialized textbooks covering airports, railroads, highways, ports, and urban public transportation systems, respectively. Periodicals such as *Aviation Week & Space Technology, Railway Age, Motor Ship, and Mass Transit* cover recent developments in their subject areas.

10.5 Alternative Fuels for Motor Vehicles

Paul Norton

Overview

Motor vehicle fuels that provide an alternative to conventional gasoline or diesel include ethanol, methanol, natural gas, propane, electricity, and biodiesel. Legislation such as the Clean Air Act Amendments of 1990 (CAAA) and the Energy Policy Act of 1992 (EPACT) has increased interest in alternative fuels. This legislation promotes alternative fuels for their potential to decrease emissions from cars and trucks and lessen the dependence of the United States on imported petroleum.

Modifications must be made to conventional vehicles to use alternative fuels. Electric vehicles (EVs) are powered by one or more electric motors that are energized by batteries. This requires a complete replacement of the conventional engine, fuel system, and drive train. The changes required for the other alternative fuels are less dramatic. All the remaining alternative fuels are burned in internal combustion engines similar to those used for conventional gasoline or diesel fuel. However, the engine and fuel system must be modified to account for the properties of the alternative fuel. The components that are typically added or changed to transform a gasoline vehicle into a compressed natural gas vehicle are shown in Figure 10.5.1. A major challenge is finding space on the vehicles for enough gas cylinders to give the vehicle an acceptable range between refuelings. These cylinders store the natural gas at pressures of 3000 to 3600 psi and must be protected from road debris and corrosive substances such as battery acid.

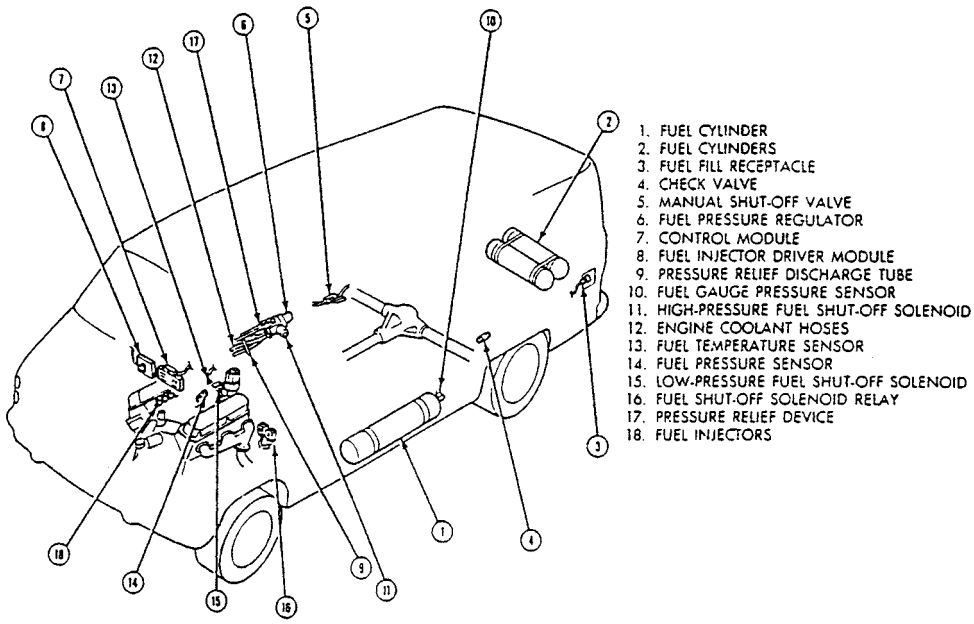


FIGURE 10.5.1 Typical Components of a compressed natural gas vehicle.

In light-duty "flexible-fueled" or "variable-fueled" vehicles, ethanol and methanol is blended with gasoline in the vehicle's fuel tank. The car can be fueled on any mix of the specified alcohol and gasoline up to 85% alcohol. Therefore an ethanol flexible-fueled vehicle (FFV) could be fueled on 85% ethanol (E85) at one refueling, and pure gasoline at the next. A sensor in the fuel line constantly measures the percentage of ethanol in the fuel and the engine control system adjusts the air/fuel ratio accordingly. The typical unique components in an FFV are shown in Figure 10.5.2. All components in the fuel system

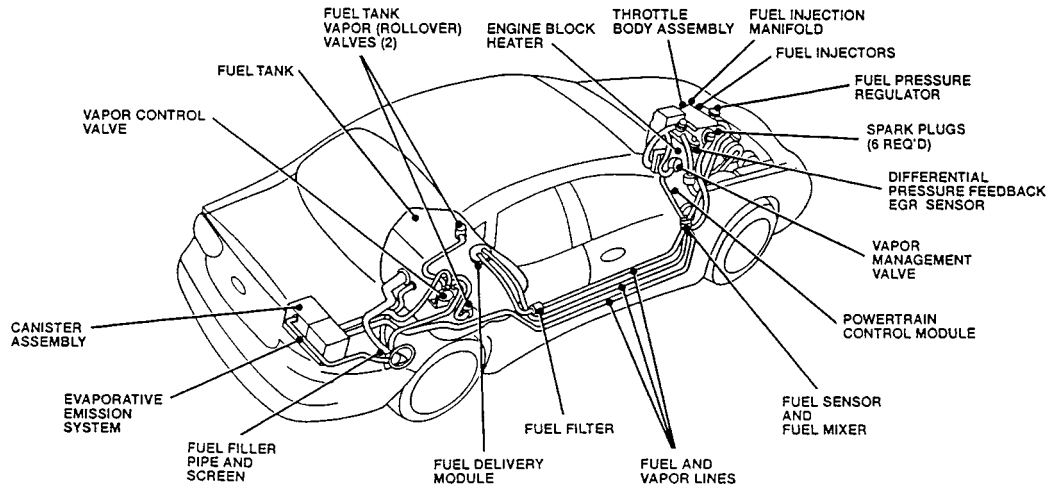


FIGURE 10.5.2 Components of a flexible fuel vehicle.

must be compatible with ethanol or methanol. The fuel injectors must have a higher flow capacity to account for the lower heating value of ethanol and methanol compared to gasoline.

Vehicles can be modified to run on an alternative fuel by the original manufacturer (OEM) or they can be modified by an aftermarket conversion shop. In general, alternative fuel vehicles produced by an OEM go through a much more rigorous development and testing process than an aftermarket conversion. The alternative fuel system for the OEM vehicle is designed for one specific model of vehicle or engine, whereas the aftermarket conversion is designed to fit on a wide variety of vehicles. Therefore the OEM system can be optimized for the one vehicle it was designed for, but the aftermarket conversion kit must be a compromise that will work with many different vehicles. Today's cars and trucks are highly developed and difficult to improve upon. Recent studies have found that compressed natural gas and propane aftermarket conversions had higher emissions in some regulated properties than they did on gasoline before the conversions (Motta et al., 1996). The same study points out that tests on OEM compressed natural gas cars showed substantially lower levels of all four regulated emissions than their gasoline powered counterparts.

Advantages and Disadvantages of Alternative Fuels

Each alternative fuel has unique physical properties that lead to advantages and disadvantages compared to conventional fuels and to each other. All the alternative fuels contain less energy per unit volume than gasoline or diesel fuel. This leads to the need to store more fuel on the vehicle to achieve the same operating range between refuelings as a gasoline car or diesel truck. In many cases, this disadvantage is balanced by lower fuel cost and/or lower emissions than a gasoline or diesel vehicle. A summary of some of the properties of the most popular alternative fuels is shown in Table 10.5.1. Some of the advantages and disadvantages of using these fuels are outlined in Table 10.5.2.

TABLE 10.5.1 Properties of Conventional and Alternative Fuels

Fuel	Cetane Number	Research Octane Number	Motor Octane Number	Density (lb/gal)	LHV[a] (Btu/gal)	LHV[a] Btu/lb)	DEE[b] Gallons	DEE[b] Pounds
Liquids at atm. Pressure								
100% Ethanol		109	90	6.6	75600	11500	0.59	0.66
100% Methanol	approx. 5	109	89	6.7	56200	8400	0.44	0.48
100% Soy Methyl-Ester (Biodiesel)	49			7.3	120200	16500	0.93	0.95
#2 Diesel	45			7.4	129000	17400	1.00	1.00
#1 Diesel	45			7.6	125800	16600	0.98	0.95
Gasoline		90–100	80–90	6.0	115400	19200	0.89	1.10
Gases at atm. Pressure								
Compressed natural gas (CNG)[c]		>127	122	n/a	n/a	20400	n/a	1.17
Liquified natural gas (LNG)[d]		>127	122	3.5	78000	22300	0.60	1.28
Propane (LPG)		109	96	4.2	83600	19900	0.65	1.14
Dimethyl ether (DME)	55–60			5.6	74800	13600	0.58	0.78

[a] Lower heating value.

[b] Diesel energy equivalent = the LHV of the alternative fuel divided by the LHV of #2 Diesel = the number of gallons (or pounds) of diesel fuel that has the same energy content as one gallon (or pound) of the alternative fuel.

[c] Average values for natural gas in the United States.

[d] 100% Methane at –263°F.

References

Motta, R. et al. April 1996. *Compressed Natural Gas and Liquefied Petroleum Gas Conversions: The National Renewable Energy Laboratory's Experience*, National Renewable Energy Laboratory. NREL/SP-425-20514.

Owen, K. and Coley, T. 1995. *Automotive Fuels Reference Book*, 2nd ed. Society of Automotive Engineers.

Reda Moh. Beta, ed. 1995. *Alternative Fuels, A Decade of Success and Promise*, Society of Automotive Engineers, SAE No. PT-48.

Whalen, P. et al. *19th Alternative Fuel Light-Duty Vehicles: Summary of Results from the National Renewable Energy Laboratory's Vehicle Evaluation Data Collection Efforts*, NREL/SP-425-20821, May.

Federal Alternative Motor Fuels Programs; Fourth Annual Report to Congress, United States Department of Energy, DOE/GO-10095-150, July, 1995.

TABLE 10.5.2 Advantages and Disadvantages of Alternative Fuels

Fuel	Advantages	Disadvantages
Liquefied Petroleum Gas (LPG) Liquefied petroleum gas is the most widely used alternative fuel. In the U.S., LPG is mostly propane with small amounts of butane and ethane. In 1993, an estimated 350,000 vehicles powered by propane (mostly conversions) were plying the nation's highways.	• Most widely available alternative fuel with an estimated 11,000 refueling sites nationwide • LPG is nontoxic • Liquefies at relatively low pressures • One gallon of liquefied propane has about 70% energy content of gasoline • Can reduce carbon monoxide and ozone forming hydrocarbon emissions • Refueling stations are relatively inexpensive	• 45% of LPG in the U.S. is derived from oil • Proper detection and ventilation are required for indoor storage and maintenance of vehicles • Starting problems may occur at very low temperatures due to low vapor pressure • Perception of a safety problem
Compressed Natural Gas (CNG) Natural gas is composed primarily of methane. It is stored on the vehicle under 3000 to 4000 psi of pressure. There are currently an estimated 50,000 vehicles operating on compressed natural gas in U.S. fleets.	• High North American resource base • Natural gas reserves are more globally distributed than oil reserves • Can reduce all regulated emissions • Is especially effective in reducing particulate emissions from large trucks • Less expensive than gasoline • Eliminates evaporative emissions because the fuel system is sealed • Can be produced from renewable resources	• Requires bulky pressure cylinders and high pressure fuel lines • Fuel system costs as much as $5,000 more than gasoline for cars • Refueling infrastructure is expensive • Energy storage density is low, making it difficult to achieve adequate range • Composition of natural gas varies regionally
Methanol Methanol is an alcohol which can be produced from coal, wood, methane, or natural gas. It is often blended with gasoline; e.g., "M85" contains 85 percent methanol and 15 percent gasoline to overcome cold start and poor flame visibility problems.	• Can be produced domestically but near-term sources are likely to be imports • Can be produced from renewable resources, but is currently produced predominantly from natural gas • Can reduce ozone-forming hydrocarbon emissions and particulate emissions • Eliminates toxic benzene emissions • A liquid fuel at ambient temperatures	• May increase formaldehyde emissions • Corrosive and toxic • Requires the use of special methanol-compatible materials in fuel system • Poor flame visibility • Has a lower energy content than gasoline, therefore requires larger fuel tanks • Currently significantly more expensive than gasoline • Lack of refueling infrastructure • Difficult to start in cold temperatures
Ethanol Ethanol is an alcohol derived from biomass-corn and other agricultural products such as sugar cane. Ethanol is often blended with gasoline to avoid taxation as an alcoholic beverage.	• Is produced domestically from renewable resources • Can reduce carbon dioxide and ozone-forming hydrocarbon emissions • Eliminates toxic benzene emissions • A liquid fuel at ambient temperatures • May increase acetaldehyde emissions • Has higher energy content than methanol	• Currently significantly more expensive than gasoline • Has a lower energy content than gasoline; therefore requires larger fuel tanks • Lack of refueling infrastructure • Requires the use of special ethanol-compatible materials in fuel system • Ability to meet large-scale demand is questionable • May compete for land with food crops

TABLE 10.5.2 Advantages and Disadvantages of Alternative Fuels (continued)

Fuel	Advantages	Disadvantages
Reformulated Gasoline* (RFG) Reformulated gasoline is made by changing the composition of conventional gasoline to reduce emissions. It often contains an oxygenate such as methyl tertiary butyl ether (MTBE).	• Reduces air pollution • Can be used in all existing cars • Can use existing refueling infrastructure	• Slightly more expensive • Does not reduce oil imports significantly • May increase engine wear and maintenance costs
Liquefied Natural Gas (LNG) Natural gas can be stored on the vehicle as a cryogenic liquid at about –260F. LNG is used primarily for heavy-duty vehicles.	Same advantages as CNG plus: • Has higher storage density than CNG so a higher range can be achieved • Usually higher methane content than CNG • Does not require high pressure cylinders for storage	• Is a cryogenic liquid with special handling requirements • Requires vacuum-insulated storage tank • Refueling infrastructure is expensive • Lack of refueling infrastructure • Low cetane number for diesel engines
Electricity Electricity is stored in a battery bank in the vehicle. The vehicle is powered by one or more electric motors.	• No tailpipe emissions • Reduces noise pollution • Can use domestic fuel sources	• Not cost competitive • Limited range per charge • Limited recharging infrastructure • May increase sulfur emissions if power is generated by coal • Potential battery disposal problem
Biodiesel Biodiesel is a methyl ester that can be made from a variety of products. Most biodiesel used in the U.S. is made from soybeans. Biodiesel is usually blended with diesel. The most common blend is 20% biodiesel, 80% conventional diesel.	• Is made from a renewable resource • Is produced domestically • Can reduce carbon monoxide, hydrocarbon, and particulate matter emissions • Is used in existing diesel engines without modifications • Is nontoxic • Has energy content comparable to diesel	• Is several times more expensive than conventional diesel • May reduce reliability and durability of the engine • Can increase emissions of oxides of nitrogen • Needs to be blended with conventional diesel to reduce gelling effect at low temperatures • May not blend easily with conventional diesel under some conditions
Dimethyl Ether (DME) DME can be made from natural gas and is a liquid at relatively low pressures (similar to LPG). It appears to be a good candidate for heavy-duty trucks. Prototype engines are currently being produced.	• Can reduce emissions significantly • Has a high cetane number so it can be used in a diesel engine • Can be produced from domestic renewable sources	• No commercially available engines at this time • Limited fuel supply • No refueling infrastructure • Uncertain production cost

* Reformulated gasoline is considered a “clean fuel” in the Clean Air Act Amendments, but is not considered an alternative fuel in the Energy Policy Act of 1992.

10.6 Electric Vehicles

Frank Kreith

Electric vehicles have recently received nationwide attention because the California Air Resources Board (CARB) has developed a plan that will require 10% of all vehicles sold in California by the year 2003 to be zero emission vehicles (ZEV) and other states are considering similar mandates. The only type of vehicle currently available that can meet the ZEV regulation is an electric vehicle (EV) powered by a bank of batteries. Figure 10.6.1 is a schematic diagram of the major components of a typical battery powered electric vehicle. Such vehicles can be built from stock automobiles with relatively minor body modification, and several types are commercially available. At this time, lead-acid batteries are the only kind ready for deployment on a large scale, but some car manufacturers are promising nickel–metal hydride batteries for their 1998 electric cars.

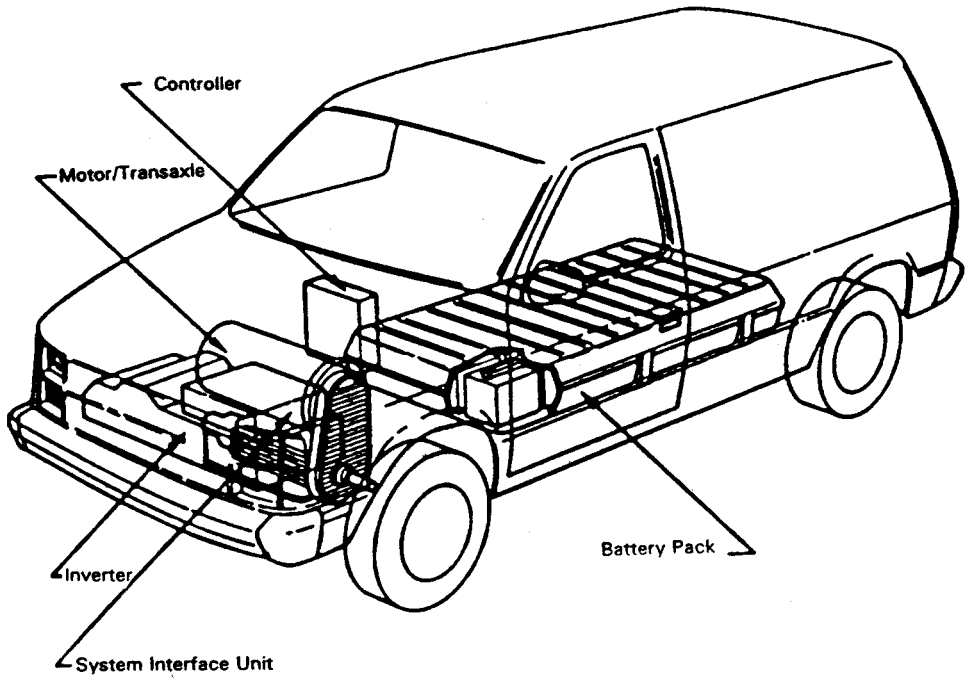


FIGURE 10.6.1 Major components of a battery powered electric vehicle. Source: Brown, P.J. 1990. Overview of the U.S. government's electric vehicle program, in *Electric Vehicles, A Decade of Transition* (Society of Automotive Engineers PT-40), p. 5.

The total cost of operating an electric vehicle consists of the purchase price and operating expenses. Electric vehicles are not yet in mass production and the cost of vehicles available in 1997 was more than twice that of equivalent gasoline powered vehicles. In addition to the cost of electric power required to charge the batteries, operating costs also include battery replacement. A typical battery pack for an electric vehicle costs between $2000 and $8000 and must be replaced after approximately 500 charge and discharge cycles for lead acid batteries and 1000 to 2000 cycles for nickel based batteries. Battery development is focusing on increasing charge cycles, reducing recharge time, and eliminating maintenance by using sealed units. Table 10.6.1 gives the performance characteristics of some batteries currently available. The main problem with deploying batteries in vehicles is their low power-to-weight ratio, called specific power, which is between 100 and 200 watts per kilogram with currently available technology. The driving range of currently available electric vehicles is between 50 and 100 miles

TABLE 10.6.1 Near-Term Battery System Performance Characteristics

Battery Type	1990 Energy Storage Density (Wh/kg)	1990 Specific Power (W/kg)	1990 Cycle Life	1995 Energy Storage Density (Wh/kg)	1995 Specific Power (W/kg)	1995 Cycle Life
Lead-Acid						
Flooded - Flat	37	150	300	40	150	300-600
Flooded - Tubular	34	80	500	35	100	600-1000
Sealed	30	160	200	33	200	300-700
Lead-fiberglass mesh	N/A	N/A	N/A	45	450-500	1000
Alkaline						
Ni-Cd	47	180	300	50	200	2000

Source: Burke, A.F. 1991. Battery availability for near-term (1998) electric-vehicles, *Electric Vehicle R&D* (Society of Automotive Engineers, SP-880, September), 17, with additions from Taylor Moore. 1994. Producing the near-term EV battery, *EPRI J.*, April/May. 6-13 and Prof. A. Frank, U. of California.

between recharging, but that range is reduced by factors such as increased vehicle payload, high speeds, steep hills, cold temperatures, and the use of auxiliary systems such as air-conditioning.

From an efficiency perspective, electric vehicle power systems are approximately as efficient as internal combustion engines according to EPA. Overall vehicle energy conversion efficiency as shown in Table 10.6.2 is between 15 and 18% with current battery technology, compared to gasoline engine powered vehicles, which have an overall efficiency of about 16%. Although electric vehicles produce no tail pipe emission, air pollution is generated by the power plants that provide the electricity to charge the batteries, and environmental/health problems may also be associated with the manufacture and disposal of batteries. Table 10.6.3 compares emission from electric vehicles with those from fossil fueled vehicles. It can be seen that electric vehicles would reduce emissions of hydrocarbons and carbon monoxide, but would produce more emissions in other categories of pollutants such as carbon dioxide, nitrous oxides, and sulfur oxides. Actual vehicle emissions would vary from region to region due to variations in the local electric generation fuel mix.

TABLE 10.6.2 Vehicle Power System Efficiencies

Primary Energy Source	Power System Component Efficiency						
	Gasoline Engine	Electric Vehicle					
	Gasoline	Coal Plant (Battery type)			Natural Gas Co-Gen Plant (Battery type)		
		Tubular Lead-Acid	Sealed Lead-Acid	Nickel-Cadmium	Tubular Lead-Acid	Sealed Lead-Acid	Nickel-Cadmium
System Component							
Energy conversion to electricity*		36%	36%	36%	43%	43%	43%
Internal combustion engine	20%						
Drive train efficiency	80%	90%	90%	90%	90%	90%	90%
Electric motor and controller		85%	85%	85%	85%	85%	85%
Battery pack		65%	81%	68%	65%	81%	68%
Battery charger		85%	85%	85%	85%	85%	85%
Overall Energy Conversion	16%	15%	19%	16%	18%	23%	19%

* Energy conversion to electricity = Electric output/Primary fuel input.

Source: McCoy, G.A. and Lyons, J.K. 1993. *Electric Vehicles: An Alternative Fuels Vehicle, Emissions, and Refueling Infrastructure Technology Assessment.* Washington State Energy Office, p. 19.

TABLE 10.6.3 Comparison of Emissions from Near-Term Alternative Fuels to Gasoline and Diesel Fuel

Vehicle Type	Fuel Source	Increase or Decrease in Emissions on a Per-Mile Basis					Sources
		CO_2	NO_x	NMHC	CO	SO_x	
	Crude oil	Baseline	Baseline	Baseline	Baseline	Baseline	
Reformulated gasoline	Crude oil	Slightly more	Slightly less	Slightly less	Slightly less	Much less	1
Compressed natural gas	Natural gas	19% less	Slightly more	Less	Much less	Less	1
Electric vehicles	Coal	20% more	More	Much less	Much less	Much more	1[a]
Electric vehicles	Natural gas	30% to 50% less	30% to 50% less	—	—	—	2
Electric vehicles	Current fuel mix	11% more	26% more	99% less	—	—	3[b]
Electric vehicles	Projected year 2000 fuel mix	—	3% more	99% less	98% less	2100% more	4[c]
Electric van	Current fuel mix	54% less	33% less	99% less	99% less	—	5
	Post-1995 mix	54% less	83% less	99% less	99% less	—	
Electric automobile	1995 mix	—	20% less to 30% more	99% less	99% less	60% to 170% more	6
Electric vehicles	Projected Southern California Utility mix for 1998 to 2010	71% less	79% less	99% less	99% less	17% more	7
Electric vehicles	California mix	—	35% less	99% less	99% less	2000% more	8[d]

Key: CO_2, Carbon dioxide; NO_x, Nitrogen oxides; NMHC, Reactive hydrocarbons; CO, Carbon monoxide; SO_x, Sulfur oxides.

[a] Includes corrections conforming to the data of Wang, Q., LeLuchi, M., and Sperling, D. 1990. Emission Impacts of Electric Vehicles, *J. Air Waste Manage. Assoc.* 40(9), 1275–1284.

[b] EVs with lead-acid batteries were compared to TIER I, Enh I/M gasoline vehicles, The current fuel mix is 54% coal, 33% nuclear and hydroelectric, 4% oil, and 11% natural gas (EPRI).

[c] Assumes a fuel mix of 54% coal, 10% natural gas, and the remainder from non-fossil sources.

[d] The projected 1994 California electric generation fuel mix is 16% coal; 33% natural gas, 82% hydroelectric, solar, wind and geothermal; 15% nuclear; 2% biomass; and 2% "uncommitted." (2, p. 44).

Sources

1. Gordon, D. 1991. *Steering a New Course: Transportation, Energy and the Environment.* Union of Concerned Scientists, Washington, D.C., pp. 75-76.
2. Plotkin, S.E. 1990. Presentation on electric vehicles before the Environment and Energy Study Institute. Washington, D.C.
3. U.S. Environmental Protection Agency (EPA). *Preliminary Electric Vehicle Emissions Assessment.* (Draft document released for review November 3, 1993.)
4. Sperling, D. and DeLuchi, M.A., *Transportation Energy Futures, Annual Review of Energy.* University of California Transportation Center, University of California, Berkeley, March 1989.
5. Moore, T. "They're New, They're Clean, They're Electric," *EPRI J.*, April/May 1991, p. 6.
6. Wang, Q., DeLuchi, M., and Sperling, D., Emission Impacts of Electric Vehicles, *Journal of Air and Waste Management Association*, 40(9) 1990, 1275–1284.
7. Hwang, R. et al. 1994. *Driving Out Pollution: The Benefits of Electric Vehicles*, The Union of Concerned Scientists, Berkeley, p. 8.
8. Technical Support Document, *Zero-Emissions Vehicle Update*, (California Air Resource Board, April 1994).

Hybrid and Fuel Cell Vehicles

In view of the shortcomings of battery powered electric vehicles, efforts are currently underway to develop automotive hybrid and fuel cell powered vehicles.

A hybrid vehicle employs a small engine and a generator to supplement battery power to the electric motor drive, thereby increasing the vehicle's range and reducing the size of the engine and the battery pack. Although hybrid vehicles are technically not "zero emission," they do reduce emission appreciably and overcome some of the shortcomings of battery powered pure electric cars.

Hybrid vehicles may be either charge sustaining or charge depleting. A charge sustaining hybrid is one that maintains the electrical energy stored on board by charging the storage device with liquid fuel from the tank. Auxiliary off-board electrical energy from a stationary power plant may or may not be required. A charge depleting hybrid is one which does not charge the electrical energy storage device directly with on-board fuel. The electrical storage device is charged by electricity from a stationary power plant only. For both, supplementary charge can be obtained by converting the kinetic energy of the vehicle to electric power during braking and deceleration.

A hybrid vehicle configuration that drives the wheels of a vehicle only electrically is called a series arrangement. When the wheels can be powered both electrically and mechanically, it is called a parallel arrangement. The two arrangements are shown schematically in Figure 10.6.2. Both parallel and series configurations may be used for city and highway driving. The parallel system may require more on-off cycles of the gasoline engine for city driving but will be more efficient in general. Both series and parallel vehicle configurations can operate in a zero tailpipe emission mode when the batteries are charged. When the batteries are drained to a certain level, the engine can turn on to recharge them.

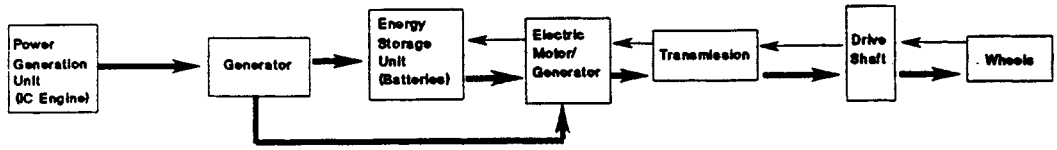


FIGURE 10.6.2A Series Hybrid Electric Vehicle.

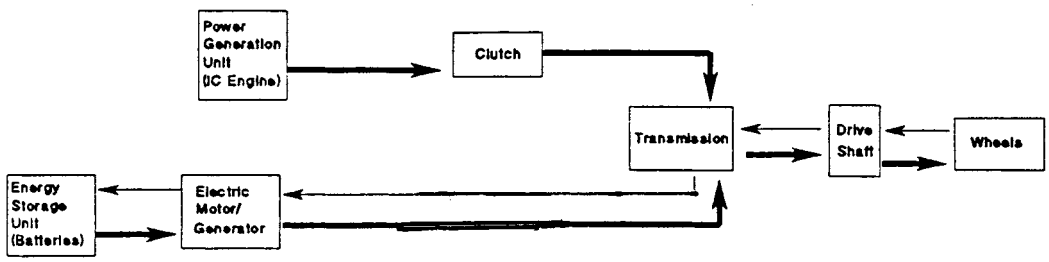


FIGURE 10.6.2B Parallel Hybrid Electric Vehicle.

Some advantages of the parallel configuration include:

- More power for a given motor size because both the engine and the motor can supply power simultaneously;
- Greater continuous operation efficiency because the engine is directly coupled to the drive shaft; and,
- Fewer components in the drive line and all of the components are available.

The benefits of a series configuration include:

- Since the engine drives a generator, it never idles and can operate at a speed that gives optimum performance;
- A smaller engine can be used, which reduces pollution and improves mileage; and,
- The configuration provides for a variety of options to arrange the engine and vehicle components.

The parallel configuration is best suited for longer trips and highway driving, while the series configuration is best suited for start-and-stop city driving.

A fuel cell converts the chemical energy of hydrogen and oxygen directly into electric power. In a fuel cell powered vehicle, hydrogen, stored under pressure in the vehicle or produced on board through a chemical process, is mixed with oxygen from the air to produce electricity in a fuel cell. The problem is creating hydrogen for a fuel cell. If it is done on-board by a reformer, the overall efficiency is very poor. If it is done off-board by a hydrogen generation plant, storing is a problem as is overall energy efficiency. However, if stored as a compressed gas, enough hydrogen to provide a 250 mile range can be stored in a 35 to 40 gallon canister. Other storage methods may be even more compact. A fuel cell electric vehicle would meet the zero emission requirements and a fuel cell powered 20 passenger bus is currently under test in Canada. Chapter 8 provides more technical background on fuel cells.

References

Kreith, F., Norton, P., and Potestio, D. 1995. Electric vehicles: promise and reality. *Transportation Q.* 49(2), 5–21.

Goehring, J., Kreith, F., Reed, J.B., and Sikes, K.P. 1996. Hybrid electric vehicles: options for state policymakers. NCSL Transportation Series, National Conference of State Legislatures, Denver, CO.

Burke, A.F. 1996. Battery availability for near-term (1998) electric vehicles, *Electric Vehicle R&D*, Society of Automotive Engineers, SP-880, September, 17.

Electric Vehicles and Clean Air. 1991. Electric Power Research Institute, Technical Brief, April.

Wouk, V. 1995. Hybrids: then and now. *IEEE Spectrum*, July, 16–21.

Further Information

National Conference of State Legislatures, 1560 Broadway, Suite 700, Denver, CO 80202; telephone: (303) 830-2200, fax: (303) 863-8003.

U.S. Department of Energy, Office of Transportation Technologies, Washington, D.C.

U.S. Department of Energy, National Renewable Energy Laboratory, Hybrid Electrical Vehicle Web Site available from http:/www.hev.doe.gov.

10.7 Intelligent Transportation Systems

James B. Reed and Janet B. Goehring

Introduction

Computers, communications systems, and automotive technologies provide the basis for intelligent transportation systems (ITS). These systems apply new technologies to address the many challenges of surface transportation—safety, productivity, environmental concerns, and mobility in the face of increasing congestion. ITS applications include, among others, on-board computer mapping devices to help drivers find unknown locations, robotic systems that automatically control braking and acceleration, "Mayday" beacons for stranded vehicles, sophisticated traffic control centers, and electronic toll collectors that do not require stopping. Most ITS initiatives emphasize highway and automobile improvements, though some focus on public transportation.

The goals of ITS are to improve transportation safety, reduce congestion, and develop new technologies. ITS can potentially reduced the national highway death toll of about 40,000 annually. According to the U.S. Department of Transportation (DOT) each year traffic accidents alone cause $70 billion in costs, not counting the loss of life or the consequences of long-term injuries. In addition, traffic congestion delays cause staggering losses in productivity of approximately $100 billion per year. An estimated 2 billion hours and 1 billion gallons of gasoline are wasted in America's collective traffic jams annually. ITS may cut travel times and fuel consumption by 10% or more, without constructing new roads. A Federal Highway Administration estimate predicts a yearly reduction in fatalities by at least 8 percent and 400,000 fewer injuries by 2011. Fewer accidents, in turn, decrease congestion since 60% of highway jamming is caused by accidents.

The **Intelligent Vehicle-Highway Systems Act of 1991**, part of the landmark Intermodal Surface Transportation Efficiency Act of 1991 (ISTEA), authorized a program of research, development, and operational testing to establish ITS as a component of the nation's surface transportation system. The program is intended to enhance the capacity, efficiency, and safety of the federally supported highway system and the states' efforts to attain mandated air quality goals. Its goal is reducing societal, economic, and environmental costs associated with traffic congestion. In addition, the legislation promotes an ITS industry in the United States, as well as the development of a technological base that would improve industrial and economic competitiveness, facilitate technology transfer, and establish a U.S. presence in this emerging field of technology.

The Intelligent Transportation Society of America (ITS America), a partnership of constituencies formed in 1990 and later designated a federal advisory committee, assists in the coordination and acceleration of the development and deployment of advanced transportation technologies in the United States. ITS America, a nonprofit educational and scientific association, includes over 750 members from federal, state, and local government, private industry, universities, research institutions and automobile organizations. ITS joined with the U.S. Department of Transportation in working to develop the comprehensive National ITS Program Plan (March 1995).

Major Elements of ITS

ITS is categorized into seven basic groupings or bundles of services.

- Travel and transportation management
- Travel demand management
- Public transportation operations
- Electronic payment
- Advanced vehicle control and safety systems

- Emergency management
- Commercial vehicle operations

Table 10.7.1 summarized from the *1995 ITS National Program Plan*, lists the "bundles" and the associated user services.

TABLE 10.7.1 ITS Bundles and User Services

Bundle	User Services
Travel and transportation management	En route driver information, route guidance, traveler services information, traffic control, incident management, emissions testing and mitigation
Travel demand management	Demand management and operations, pre-trip travel information, ride matching
Public transportation operations	Public transportation management, en-route transit information, personalized public transit, public travel security
Electronic payment	Electronic payment services, "smart cards"
Commercial vehicle operations	Electronic clearance, automated roadside safety inspection, on-board safety monitoring, administrative processes, freight mobility, hazardous material incident response
Emergency management	Emergency notification, personal security, emergency vehicle management
Advanced vehicle control and safety systems	Collision avoidance, vision enhancement, safety readiness, pre-crash restraint deployment, automated highway systems

Intelligent Transportation Infrastructure

In order to assist the linking of ITS systems and technologies, the DOT developed with ITS America a national ITS framework referred to as "system architecture" (Figure 10.7.1).

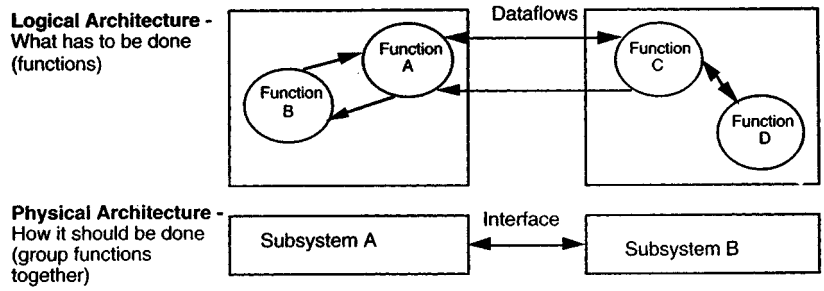


FIGURE 10.7.1 Schematic Diagram of ITS Architecture.

The architecture defines a series of *functions* that an ITS system would carry out, e.g., traffic surveillance or operating transit vehicles. These high-level functions are broken down into subfunctions. Higher level functions are depicted graphically as data flow diagrams or *DFDs*. Each of the lower level functions is given a precise definition, known as a process specification or P-spec for short.

One of the primary roles of the architecture is to assign functions to various components, or *subsystems*, of the overall ITS system. Traffic management or transit management are example of subsystems. However, the various subsystems cannot act independently of one another. Information must be exchanged between these subsystems if the overall ITS system is to act in a coordinated, integrated fashion. Therefore, the architecture also defines the connections or *interfaces* between these subsystems in terms of *data flows* as shown in Figure 10.7.1.

To enhance the deployment efforts of ITS projects and test the architecture, the DOT initiated the Intelligent Transportation Infrastructure (ITI) in 1996. Also called "Operation TimeSaver," the ITI

promises to reduce the travel time of Americans by 15% once the infrastructure is fully in place. The ITI seeks to link individual ITS components already in existence in major metropolitan areas and to put the national architecture into action. DOT established nine integrated components for the ITI focus: traffic signal control, freeway management, incident management, electronic fare payment, electronic toll collection, railroad grade crossings, emergency management services, and regional multimodal traveler information.

Traffic Signal Control

Traffic signal control systems react to changing traffic conditions to improve transportation system efficiency. State-of-the-art traffic controls, often using fuzzy logic control, have the capability to modify signal timings in response to traffic demand. The signal systems employ a variety of mechanisms to collect and process real time traffic conditions. When integrated into an ITI system, the control devices can give priority to emergency response vehicles or transit to facilitate the efficient flow of traffic.

To provide metropolitan-wide signal coordination, the systems in one region should be able to share traffic flow information electronically with the systems of neighboring jurisdictions. This information sharing ensures smooth traffic flow along major corridors.

Freeway Management

Real-time information about traffic and road conditions aids in the efficient management of the freeway system. Methods for gathering such information include video monitors, microwave radar, ultrasonic monitors, and citizen and police reports. The information is provided to travelers through infrastructure-based devices such as variable message signs and highway advisory radio broadcasts. The relevant information can also be used to support incident management activities and can help determine the severity and type of incident to direct the appropriate response team. The freeway management system can also be used to streamline the traffic logistics of a large special event, such as, for example, the fuzzy system control in Munich.

Transit Management

Transit management aids in the on-time performance of transit systems. Transit fleet management includes real-time transit information to travelers, monitoring the location of buses and other equipment, flexible rerouting of vehicles and automated maintenance monitoring. Other enhancements include displays of routes and schedules for passengers on the vehicle.

Incident Management

Rapid and efficient response to accidents and other incidents is a key factor in saving lives and easing travel delays. Incident management systems help fire and police to quickly identify and respond to incidents. Objectives of these systems include coordination of services across regional boundaries, improvement of response times, the use of in-vehicle navigation equipment to assist emergency vehicles in locating the scene of the accident, and the reduction of traffic congestion from the incidents.

Electronic Fare Payment

Electronic fare payment involves the use of a one payment system such as a “smart card” for transit fares, parking fees, and other travel related fees. The cards reduce the time needed for travelers to search for exact change or figure out fare amounts. The cards can work as credit/debit cards or stored value cards.

Electronic Toll Collection

Electronic toll collection reduces delays at toll plazas and the operating costs of toll agencies. The systems include payment cards or tags on the vehicle and a communications link between the vehicle and the toll crossing. As the vehicle passes the toll station the payment is automatically processed. The use of common toll readers and tags that promote interoperability further decreases the time and cost to travelers.

Railroad Grade Crossings

Intersections of a highway and railroad tracks pose a unique safety hazard, because trains that travel at high speeds may take more than one mile to stop. The ITI deployment objectives include automated warnings at rail crossings, advanced warning systems to travelers, and the coordination of rail movements with traffic signal control systems.

Emergency Management Services

Emergency management services assist emergency vehicles in responding in a more efficient manner to incidents. One ITI application includes an on-board vehicle mapping system. Also included are automated vehicle location devices to inform dispatchers of the exact location of the emergency vehicle. The timely and accurate information made available by these systems will also improve the response to hazardous materials incidents.

Regional Multimodal Traveler Information

Traveler information systems provide integrated road and transit information to travelers to assist them in making travel choices in a metropolitan area. This system promotes regional coordination in the collecting, processing, and presentation of traveler information. One method of making the information available is through a multimedia kiosk placed in transit stations, tourist centers, or other convenient locations. The information may also be provided to the public via a range of communications systems such as broadcast radio, television, or the Internet.

The ITS national architecture addresses the systems that make up ITI and arranges data to be shared. The architecture supports the deployment process and helps to create an integrated system infrastructure that promises to ease surface transportation.

Defining Terms

Intelligent transportation infrastructure: The systems related to traffic detection, monitoring, communications, and control that are required to support ITS products and services in metropolitan and rural areas.

Intermodal: Possessing the ability to transfer people and goods between transportation modes such as rail, air, water, and highway.

ITS system architecture: The written framework that describes how system components interact and work together to achieve system goals.

Fuzzy logic (see Section 19.15).

Real-time traffic information: Instantaneous video or other data on traffic status, i.e., congestion and collisions, that is conveyed from a central location to travelers and others who require such data.

References

Costantino, J. 1994. "Overcoming Institutional Barriers to IVHS." *Transportation Q.*, Winter 3–13.

Euler, G.W. and Robertson, H.D., eds. 1995. *National ITS Program Plan*, ITS America, Washington, D.C.

Intelligent Transportation Society of America. 1994. *ITS Information Sourcebook: A Media Professional's Guide to Key Contacts for Intelligent Transportation Systems Information*, ITS America, Washington, D.C.

ITS America. 1994. *ITS Architecture Development Program Phase I.* Washington, D.C.

U.S. Department of Transportation. 1996. *Intelligent Transportation Systems (ITS) Projects.* Washington, D.C.

U.S. General Accounting Office. 1994. *Smart Highways: An Assessment of Their Potential to Improve Travel.* GAO/PEMD-91-18, Washington, D.C.

Further Information

National Conference of State Legislatures
1560 Broadway, Suite 700
Denver, CO 80202
(303) 830-2200

U.S. Department of Transportation
Joint Program Office for Intelligent Transportation Systems
400 7th Street SW HVH-1
Washington, D.C. 20590
(202) 366-2196

ITS America
400 Virginia Ave., SW, Suite 800
Washington, D.C. 20024
(202) 484-4847

National Governor's Association
444 Capitol Street, Suite 267
Washington, D.C. 20001
(202) 624-7740

Albano, L.D.; et. al. “Engineering Design”
Mechanical Engineering Handbook
Ed. Frank Kreith
Boca Raton: CRC Press LLC, 1999

Engineering Design

Leonard D. Albano
Worcester Polytechnic Institute

Nam P. Suh
Massachusetts Institute of Technology

Michael Pecht
University of Maryland

Alexander Slocum
Massachusetts Institute of Technology

Mark Jakiela
Massachusetts Institute of Technology

Kemper Lewis
Georgia Institute of Technology

Farrokh Mistree
Georgia Institute of Technology

J.R. Jagannatha Rao
University of Houston

11.1 Introduction

Nam P. Suh

Traditionally, the design field has been identified with particular end products, e.g., mechanical design, electrical design, ship design. In these fields, design work is largely based on specific techniques to foster certain product characteristics and principles. Examples include the principles of constant wall thickness, lightweight construction, and shortest load path. Also, the design field has been subdivided by an increasing reliance on specialized knowledge and the division of labor. Precision engineering and robotics are examples of subfields that are distinguished by the accuracy and reliability that the product must have. In the field of precision engineering, for instance, the dimensions of interest are nanometers, which are often encountered in the semiconductor industry. Each one of these fields also has its specific know-how and paradigms to support effective design have also sub-divided the design field.

There are three branches of design. The traditional school, which still dominates, believes that design requires experience and cannot be taught. The second group deals with optimization as a subset of design, using computer-based tools such as genetic algorithms, fuzzy logic, and the like. The third school of thought believes that there are axioms that govern good design decisions. A good designer needs to use all three methodologies.

11.2 Elements of the Design Process

Nam P. Suh

All design activities must do the following:

1. *Know* the "*customers' needs*."
2. *Define the essential problems* that must be solved to satisfy the needs.
3. *Conceptualize the solution through synthesis*, which involves the task of satisfying several different functional requirements using a set of inputs such as product design parameters within given constraints.
4. *Analyze* the proposed solution to establish its optimum conditions and parameter settings.
5. *Check the resulting design solution* to see if it meets the original customer needs.

Design proceeds from abstract and qualitative ideas to quantitative descriptions. It is an iterative process by nature: new information is generated with each step, and it is necessary to evaluate the results in terms of the preceding step. Thus, design involves a continuous interplay between *the requirements the designer wants to achieve* and *how the designer wants to achieve these requirements*.

Designers often find that a clear description of the design requirements is a difficult task. Therefore, some designers deliberately leave them implicit rather than explicit. Then they spend a great deal of time trying to improve and iterate the design, which is time consuming at best. To be efficient and generate the design that meets the perceived needs, the designer must specifically state the users' requirements before the synthesis of solution concepts can begin.

Solution alternatives are generated after the requirements are established. Many problems in mechanical engineering can be solved by applying practical knowledge of engineering, manufacturing, and economics. Other problems require far more imaginative ideas and inventions for their solution. The word "creativity" has been used to describe the human activity that results in ingenious or unpredictable or unforeseen results (e.g., new products, processes, and systems). In this context, creative solutions are discovered or derived by inspiration and/or perspiration, without ever defining specifically what one sets out to create. This creative "spark" or "revelation" may occur, since our brain is a huge information storage and processing device that can store data and synthesize solutions through the use of associative memory, pattern recognition, digestion and recombination of diverse facts, and permutations of events. Design will always benefit when "inspiration" or "creativity," and/or "imagination" plays a role, but this process must be augmented by amplifying human capability systematically through fundamental understanding of cognitive behavior and by the development of scientific foundations for design methods.

11.3 Concept of Domains

Nam P. Suh

Design is made up of four *domains*: the customer *domain*, the *functional domain*, the *physical domain*, and the *process domain* (see Figure 11.3.1). The domain on the left relative to the domain on the right represents "what the designer wants to achieve," whereas the domain on the right represents the design solution, or "how the designer proposes to satisfy the problem." Therefore, the design process can be defined as mapping from the "what" domain to the "how" domain. During product design, the mapping is from the functional domain to the physical domain. In manufacturing process design, the designer maps from the physical domain to the process domain.

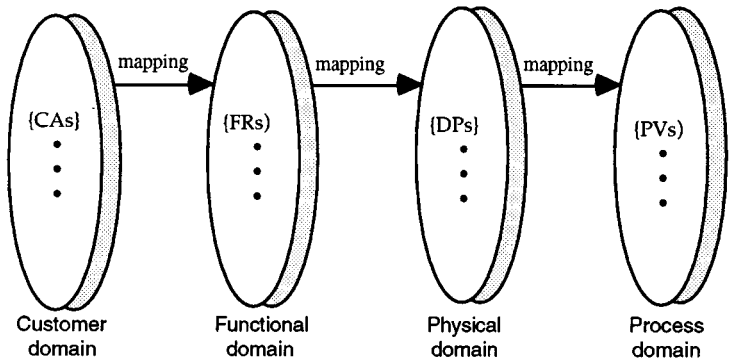


FIGURE 11.3.1 Four domains of the design world. {x} are characteristic vectors of each domain.

The customer domain is characterized by customer needs or the attributes the customer is looking for in a product, process, system, or material. In the functional domain, the designer formally specifies customer needs in terms of functional requirements (FRs). In order to satisfy these FRs, design parameters (DPs) are conceived in the physical domain. Finally, a means to produce the product specified in terms of DPs is developed in the process domain, which is characterized by process variables (PVs).

In mechanical engineering, design typically refers to product design and, often, hardware design. However, mechanical engineers also deal with other equally important designs such as software design, design of manufacturing processes and systems, and organizations. All designers go through the same thought process, although some believe that their design is unique and different from those of everyone else. In materials science, the design goal is to develop materials with certain properties (i.e., FRs). This is done through the design of microstructures (i.e., DPs) to satisfy these FRs, and through the development of material processing methods (i.e., PVs) to create the desired microstructures. To establish a business, the goals are {FRs}, and they are satisfied by structuring the organization in terms of its departments {DPs} and finding the human and financial resources {PVs} necessary to staff and operate the enterprise. Similarly, universities must define the mission of their institutions (i.e., FRs), design their organizations effectively to have an efficient educational and research enterprise (i.e., DPs), and must deal with human and financial resource issues (i.e., PVs). In the case of the U.S. government, the President of the United States must define the right set of {FRs}, design the appropriate government organization and programs {DPs}, and secure the resources necessary to get the job done {PVs}, subject to the constraints imposed by the Constitution and Congress. In all organizational designs the process domain represents the resources: human and financial.

Table 11.3 shows how seemingly different design tasks in many different fields can be described in terms of the four design domains. In the case of the product design, the customer domain consists of the customer requirements or attributes the customer is looking for in a product; the functional domain consists of functional requirements, often defined as engineering specifications and constraints; the

TABLE 11.3 Characteristics of the Four Domains of the Design World for Various Designs: Manufacturing, Materials, Software, Organizations, and Systems

Domains Character Vectors	Customer Domain {CAs}	Functional Domain {FRs}	Physical Domain {DPs}	Process Domain {PVs}
Manufacturing	Attributes which consumers desire	Functional requirements specified for the product	Physical variables which can satisfy the functional requirements	Process variables that can control design parameters (DPs)
Materials	Desired performance	Required properties	Microstructure	Processes
Software	Attributes desired in the software	Output	Input variables and algorithms	Subroutines
Organization	Customer satisfaction	Functions of the organization	Programs or offices	People and other resources that can support the programs
Systems	Attributes desired of the overall system	Functional requirements of the system	Machines or components, subcomponents	Resources (human, financial, materials, etc.)

physical domain is the domain in which the key design parameters {DPs} are chosen to satisfy the {FRs}; and the process domain specifies the manufacturing methods that can produce the {DPs}. It is indeed fortunate that all designs fit into these four domains, since in a given design task, mechanical design, software design, manufacturing issues, and organizational issues must often be considered simultaneously. Because of this logical structure of the design world, generalized design principles can be applied to all design applications, and the issues that arise in the four domains can be considered systematically and concurrently.

Customer needs are often difficult to define. Nevertheless the designer must make every effort to understand customer needs by working with customers to appreciate and establish their needs. Then these needs (or the attributes the customer is looking for in a product) must be translated into functional requirements (FRs). This must be done in a "solution neutral environment." This means that the FRs must be defined without bias to any existing or preconceived solutions. If the FRs are defined based on an existing design, then the designer will simply be specifying the FRs of that product and creative thinking cannot be done.

To aid the process of defining FRs, QFD (quality function deployment) has been used. In QFD customer needs and the possible functional requirements are correlated and important FRs are determined. Experience plays an important role in defining FRs, since qualitative judgment is often necessary for assessing customer needs and identifying the essential problems that must be solved.

11.4 The Axiomatic Approach to Design

Nam P. Suh

The creative process of mapping the FRs in the functional domain to DPs in the physical domain is not unique; the solution varies with a designer's knowledge base and creative capacity. As a consequence, solution alternatives may vary in their effectiveness to meet the customer's needs. The axiomatic approach to design is based on the premise that there are generalizable principles that form the basis for distinguishing between good and bad designs.

Suh (1990) identified two design axioms by *abstracting* common elements from a body of good designs, including products, processes, and systems. The first axiom is called the *Independence Axiom*. It states that the independence of *functional requirements* (FRs) must be always maintained, where FRs are *defined as the minimum set of independent functional requirements* that characterize the design goals. The second axiom is called the *Information Axiom*, which states that among those designs that satisfy the Independence Axiom the design that has the highest probability of success is the best design. During the mapping process (for example, mapping from FRs in the functional domain to DPs in the physical domain), the designer should make correct design decisions using the Independence Axiom. When several designs that satisfy the Independence Axiom are available, the Information Axiom can be used to select the best design.

Axioms are general principles or self-evident truths that cannot be derived or proven to be true; however they can be refuted by counterexamples or exceptions. Through axioms such as Newton's laws and the laws of thermodynamics, the concepts of force, energy, and entropy have been defined. One of the main reasons for pursuing an axiomatic approach to design is the generalizability of axioms, which leads to the derivation of corollaries and theorems. These theorems and corollaries can be used as design rules that precisely prescribe the bounds of their validity because they are based on axioms. The following corollaries are presented in Suh (1990).

Corollary 1: (Decoupling of Coupled Designs)
Decouple or separate parts or aspects of a solution if FRs are coupled or become interdependent in the designs proposed.

Corollary 2: (Minimization of (FRs)
Minimize the number of FRs and constraints.

Corollary 3: (Integration of Physical Parts)
Integrate design features in a single physical part if FRs can be independently satisfied in the proposed solution.

Corollary 4: (Use of Standardization)
Use standardized or interchangeable parts if the use of these parts is consistent with FRs and constraints.

Corollary 5: (Use of Symmetry)
Use symmetrical shapes and/or components if they are consistent with the FRs and constraints.

Corollary 6: (Largest Tolerance)
Specify the largest allowable tolerance in stating FRs.

Corollary 7: (Uncoupled Design with Less Information)
Seek an uncoupled design that requires less information than coupled designs in satisfying a set of FRs.

The ultimate goal of axiomatic design is to establish a science base for design and improve design activities by providing the designer with a theoretical foundation based on logical and rational thought processes and tools.

The First Axiom: The Independence Axiom

The Independence Axiom may be formally stated as:

Axiom 1: The Independence Axiom
Maintain the independence of the functional requirements.

As stated earlier, functional requirements, FRs, are defined as the minimum set of independent requirements that the design must satisfy. A set of functional requirements is the description of design goals. The Independence Axiom states that when there are two or more functional requirements, the design solution must be such that each of the functional requirements can be satisfied without affecting any of the other requirements. This means that the designer must choose the correct set of DPs so that functional dependence or coupling is not introduced. When there is only one FR, the Independence Axiom is always satisfied. In this case, the given design alternatives should be optimized and the second axiom, the Information Axiom, is used to select the best design.

To apply the Independence Axiom, the mapping process from the design goals to the design solutions can be expressed mathematically. The set of functional requirements that define the specific design goals constitutes a vector {FRs} in the functional domain. Similarly, the set of design parameters in the physical domain that describe the design solution also constitutes a vector {DPs}. The relationship between the two vectors can be written as

$$\{\text{FRs}\} = [A]\{\text{DPs}\} \tag{11.4.1}$$

where [A] is the design matrix that characterizes the nature of the mapping. Equation (11.4.1) may be written in terms of its elements as $FR_i = A_{ij}DP_j$. Equation (11.4.1) is a design equation that may be used for the design of a product or the microstructure of a material. For the design of processes, the design equation may be written as

$$\{\text{DPs}\} = [B]\{\text{PVs}\} \tag{11.4.2}$$

where [B] is the design matrix that characterizes the process design.

Designs that satisfy the Independence Axiom must have either a diagonal or triangular design matrix (see Figure 11.4.1). When the design matrix [A] is diagonal, each of the FRs can be satisfied independently by adjusting one DP. Such a design is called an *uncoupled design*. When the matrix is triangular, the independence of FRs can be guaranteed if and only if the DPs are changed in a proper sequence. Such a design is called a *decoupled* or *quasi-coupled design*. Although the design matrix is a second-order tensor (note: stress, strain, and moment of inertia are also second-order tensors), the usual coordinate transformation technique cannot be applied to Equations (11.4.1) or (11.4.2) to find the invariants such as the diagonal matrix, since [A] and [B] typically involve physical phenomena and geometric relationships that are not amenable to coordinate transformation.

In addition to the Independence Axiom, the mapping of the design goals (FRs or DPs) to design solutions (DPs or PVs, respectively) is often subject to constraints, Cs. Constraints establish the bounds on the acceptable design solutions and differ from FRs in that they do not have to be independent. Cost, for example, is often considered a constraint since it is affected by all design decisions, but the design is acceptable as long as the cost does not exceed a set limit.

Example 1: Shaping of Hydraulic Tubes

In many applications (e.g., aircraft industry), steel tubes must be bent to complex shapes without changing the circular cross-sectional shape of the tube. To design a machine and a process that can achieve the task, the functional requirements can be formally stated as

FR1 = Bend the steel tube to prescribed curvatures.
FR2 = Maintain the circular cross section of the bent tube.

$$\begin{Bmatrix} FR_1 \\ FR_2 \\ FR_3 \end{Bmatrix} = \begin{bmatrix} X & 0 & 0 \\ 0 & X & 0 \\ 0 & 0 & X \end{bmatrix} \begin{Bmatrix} DP_1 \\ DP_2 \\ DP_3 \end{Bmatrix}$$

a) Uncoupled design

$$\begin{Bmatrix} FR1 \\ FR2 \\ FR3 \end{Bmatrix} = \begin{bmatrix} X & 0 & 0 \\ X & X & 0 \\ X & X & X \end{bmatrix} \begin{Bmatrix} DP1 \\ DP2 \\ DP3 \end{Bmatrix}$$

b) Decoupled design

$$\begin{Bmatrix} FR1 \\ FR2 \\ FR3 \end{Bmatrix} = \begin{bmatrix} X & 0 & X \\ X & X & 0 \\ X & X & X \end{bmatrix} \begin{Bmatrix} DP1 \\ DP2 \\ DP3 \end{Bmatrix}$$

c) Coupled design

FIGURE 11.4.1 Examples of uncoupled, decoupled, and coupled designs.

One mechanical concept that can do the job is shown schematically in Figure 11.4.2 for a two-dimensional bending case. It consists of a set of matching rollers with semicircular grooves on their periphery. These "bending" rollers can counterrotate at different speeds and move relative to each other to control the bending process. The centers of these two bending rollers are fixed with respect to each other, and the contact point of the bending rollers can rotate about a fixed point. A second set of "feed" rollers, which counterrotate at the same speed, feed the straight tube feedstock into the bending rollers. As the tubes are bent around the rollers, the cross-sectional shape will tend to change to a noncircular shape. The deformation of the cross section is prevented by the semicircular cam profile machined on the periphery of the bending rollers. The DPs for this design are

DP1 = Differential rotation of the bending rollers to bend the tube
DP2 = The profile of the grooves on the periphery of the bending rollers

The kinematics of the roller motion needs to be determined. To bend the tube, one of the bending rollers must rotate faster than the other. In this case, the tube will be bent around the slower roller. The forward speed of the tube is determined by the average speed of the two bending rollers. The motion of these rollers can be controlled digitally using stepping motors.

The design shown in Figure 11.4.2 is an uncoupled design, since each of the proposed DPs only affects one FR. Is this the best design? The only way this question can be answered is to develop alternative designs that satisfy the FRs, constraints (Cs), and the Independence Axiom. Then the information content for each of the proposed designs must be computed so as to use the Information Axiom to select the best among the proposed designs.

Decomposition, Zigzagging, and Hierarchy

In the preceding example the design was completed when the two FRs in the functional domain were mapped to two DPs in the physical domain. However, in many designs both the FRs and DPs must be decomposed into hierarchies because the high level conceptual design ideas need further design details before they can be implemented. To create the hierarchies for the FRs and DPs, the designer must return to the functional domain from the physical domain. As shown in Figure 11.4.3, the design process requires the designer to alternate or zigzag between the functional domain and the physical domain.

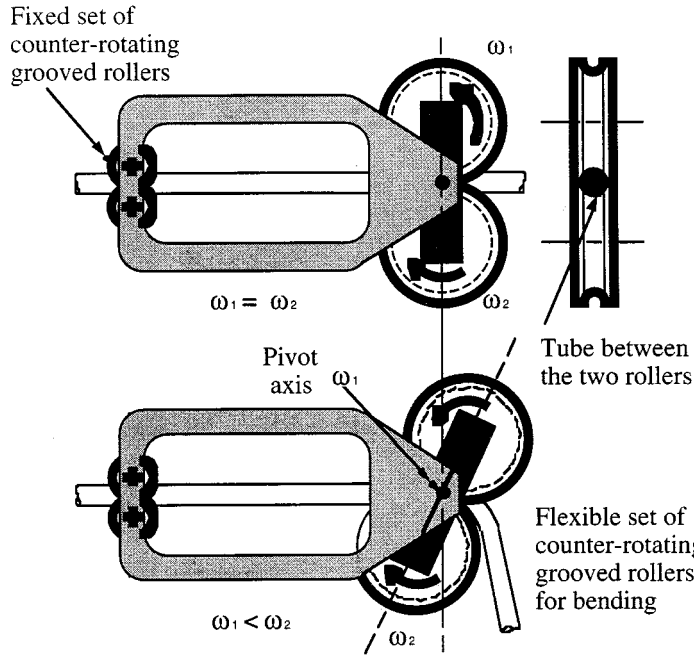


FIGURE 11.4.2 Concept for tube bending.

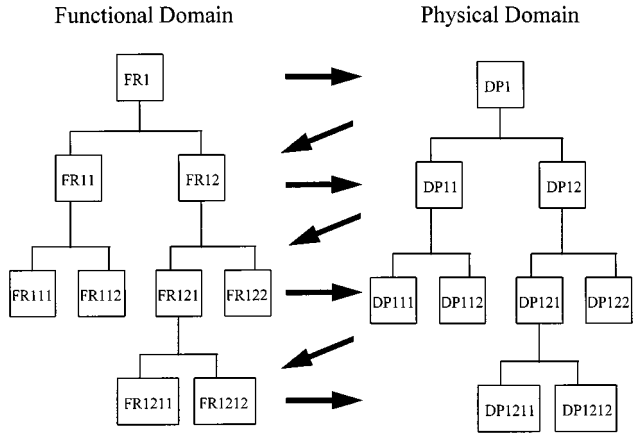


FIGURE 11.4.3 Hierarchical decomposition of FRs and DPs and zigzag mapping.

In many organizations, functional requirements or design specifications are decomposed without zigzagging by remaining only in the functional domain. This means that the designers are not working in a *solution-neutral environment*; they are specifying the requirements for an existing design. For example, assume that the design objective is to develop a vehicle that goes forward, stops, and turns. This vehicle has to satisfy these three FRs. These FRs cannot be decomposed until they are mapped to a set of DPs in the physical domain. This high-level mapping is often referred to as the *conceptual phase* of design. If, for instance, an electric motor is adopted as a means to move forward, then the FR "go forward" would be further decomposed in terms of this physical concept, and the evolving functional hierarchy will be different from that associated with the decision to use gas turbines. Therefore, to define the FRs in a solution-neutral environment, the designer needs to "zig" to the physical domain, and after proper DPs are chosen, the designer must then "zag" to the functional domain for further decomposition.

This process of mapping and zigzagging must continue until the design is completed, resulting in the creation of hierarachical trees for both FRs and DPs. This will be illustrated in the following example.

Example 2: Refrigerator Design

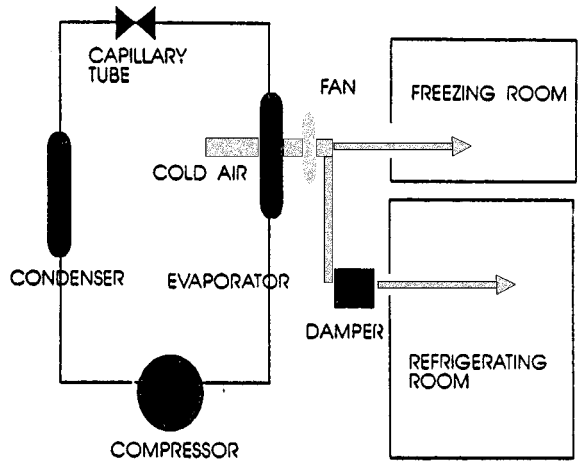
Historically, humankind has had the need to preserve food. Now consumers want an electrical appliance that can preserve food for an extended time. The typical solution is to freeze food for long-term preservation or to keep some food at a cold temperature without freezing for short-term preservation. These needs can be formally stated in terms of two functional requirements:

FR1 = Freeze food for long-term preservation.
FR2 = Maintain food at cold temperature for short-term preservation.

To satisfy these two FRs, a refrigerator with two compartments is designed. The two DPs for this refrigerator may be stated as

DP1 = Freezer section
DP2 = Chiller (i.e., refrigerator) section

To satisfy FR1 and FR2, the freezer section should only affect the food to be frozen and the chiller (i.e., refrigerator) section should only affect the food to be chilled without freezing. In this case, the design matrix will be diagonal. However, the conventional freezer/refrigerator design uses one compressor and one fan which turns on when the freezer section temperature is higher than the set temperature, and the chiller section is cooled by controlling the opening of the vent (see Figure 11.4.4). Therefore, the temperature in the chiller section cannot be controlled independently from that of the freezer section.



ONE COOLING FAN TYPE

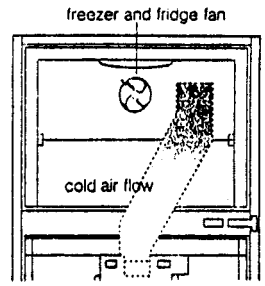
Dependent Control

Fan ----------> Freezing Room
Fan + Damper -----> Refrigerating Room

Thermo Sensor in F-Room Only

The Damper produces large pressure loss

a



b

FIGURE 11.4.4 Conventional refrigerator/freezer cooling system.

Having chosen DP1, FR1 can now be decomposed as

FR11 = Control temperature of the freezer section in the range of –18°C.
FR12 = Maintain the uniform temperature throughout the freezer section at the preset temperature.
FR13 = Control humidity to relative humidity of 50%.

FR2 may be decomposed in the context of DP2 as

FR21 = Control the temperature of the chilled section in the range of 2 to 3°C.
FR22 = Maintain a uniform temperature throughout the chilled section at a preset temperature to within 1°C.

To satisfy the second level FRs, i.e., FR11, FR12, FR21, etc., the designer has to conceive a design and identify the DPs for this level of decomposition. Just as FR1 and FR2 were independent from each other through the choice of proper DP1 and DP2, the FRs at this second level must also be independent from one another.

Suppose that the requirements of the freezer section will be satisfied by pumping chilled air into the freezer section, circulating the chilled air uniformly throughout the freezer section, and monitoring the returning air for temperature and moisture in such a way that the temperature is controlled independently from the moisture content of the air. Then the second level DPs may be chosen as

DP11 = Turn the compressor on or off when the air temperature is higher or lower than the set temperature, respectively.
DP12 = Blow the air into the freezer section and circulate it uniformly throughout the freezer section at all times.
DP13 = Condense the moisture in the returned air when its dew point is exceeded.

The design equation is written as

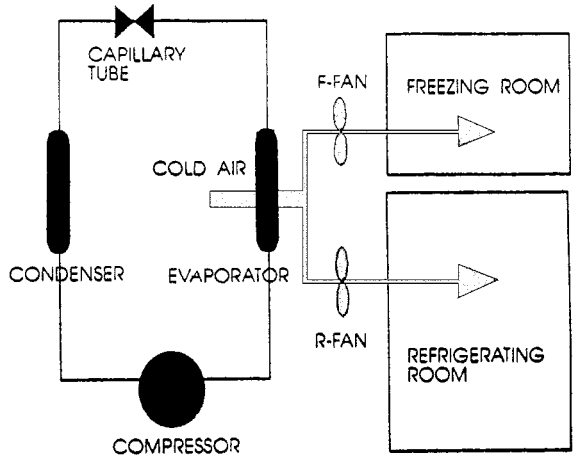
$$\begin{vmatrix} FR12 \\ FR11 \\ FR13 \end{vmatrix} = \begin{vmatrix} x & 0 & 0 \\ x & x & 0 \\ x & 0 & x \end{vmatrix} \begin{vmatrix} DP12 \\ DP11 \\ DP13 \end{vmatrix} \tag{11.4.3}$$

Equation (11.4.3) indicates that the design is a decoupled design.

The chilled section, where the food has to be kept in the range of 2 to 3°C, can now be designed. Here, again, a compressor may be used to control the air temperature within a preset range, and chilled air may be circulated to maintain a uniform temperature throughout the chilled section. This solution would result in a decoupled design as well. One of the design questions to be answered here is whether one compressor and one fan can be used to satisfy both {FR11, FR12, FR13} and {FR21, FR22}. Such a solution would minimize the information content without compromising functional independence (see Corollary 3). Most commercial refrigerators use only one compressor and one fan to achieve these goals (see Figure 11.4.4); however, many of these are coupled designs.

One can propose various specific design alternatives and evaluate the options in terms of the Independence Axiom. If a design allows the satisfaction of these FRs independently, then the design is acceptable for the set of specified FRs. If a solution that satisfies the Independence Axiom cannot be devised, then the designer must compromise the FRs by either eliminating some of the FRs or increasing the tolerance for temperature control, moisture control, etc.

One company has improved the preservation of food in their chilled section by providing one additional fan so as to be able to control the temperature of the chilled section more effectively (see Figure 11.4.5). This can be done since the evaporator was sufficiently cold to cool the air being pumped into the chiller section, even during periods when the compressor was not running. To maintain a uniform temperature distribution, extra vents were provided to insure good circulation of air. In this design, DP21 refers to



TWO COOLING FAN TYPE

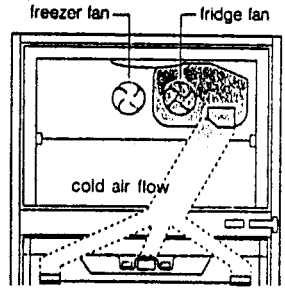
Independent Control

F-Fan ----------> Freezing Room
R-Fan ---------> Refrigerating Room

Thermo Sensor in F-Room and R-Room

No Damper is Required

a



b

FIGURE 11.4.5 New refrigerator/freezer cooling system.

the fan for the chiller section, and DP22 represents the vents. The design matrix for the {FR21, FR22}–{DP21, DP22} relationship is diagonal as shown in the design equation:

$$\begin{vmatrix} FR21 \\ FR22 \end{vmatrix} = \begin{vmatrix} X & 0 \\ 0 & X \end{vmatrix} \begin{vmatrix} DP21 \\ DP22 \end{vmatrix} \tag{11.4.4}$$

The company has compared the performance of the uncoupled design with that of competing coupled designs. For instance, the new design provides much more uniform temperature in the chiller section (Figure 11.4.6a) and much less temporal fluctuation than the existing coupled designs (Figure 11.4.6b). They also found that the uncoupled design saves electricity because air can be defrosted due to the air flow into the chiller section when the compressor is not operating. They also found that this new design enables the use of a quick refrigeration mode in the chilled section by turning on the fan of the chiller section as soon as food is put into the chiller section. To cool 100 g of water from 25 to 10°C it took only 37 vs. 58 min in a conventional refrigerator (Figure 11.4.7).

According to Corollary 3 (Integration of Physical Parts), the innovative idea of using two fans and uniformly positioned ducts may or may not be the best solution if the FRs can be satisfied independently using only one fan. If there is an alternative design that can satisfy the Independence Axiom, the Information Axiom and the detailed calculation of the information content must be used to determine the better of the two designs. As discussed in "The Second Axiom: The Information Axiom" (below), the best design is the one that has the minimum information content since it has the highest probability of success.

In the preceding example, the design matrices were formulated in terms of Xs and Os to model the nature of the relationship between each FR and each DP. In some cases, this simple, conceptual notation is sufficient to complete and implement the design. However, in many cases further steps should be taken to optimize the design. This means that the FR-DP relationships must be modeled more precisely after the conceptual design is represented in terms of Xs and Os.

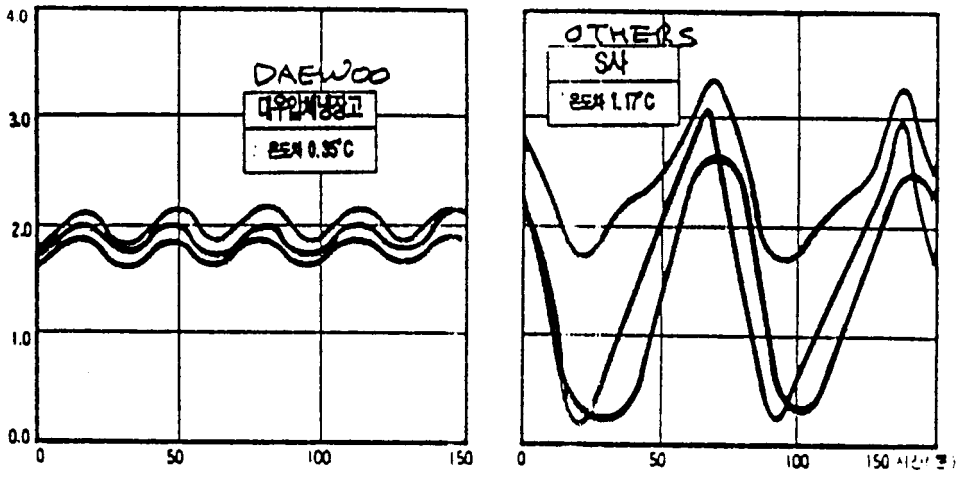
Concurrent Design and Manufacturing Considerations

The previous section, Decomposition, Zigzagging, and Hierarchy, demonstrates use of the Independence Axiom during mapping from the physical domain to the process domain, i.e., product design. In order to implement the chosen DPs, the designer has to map from the physical domain to the process domain (i.e., process design) by choosing the process variables, PVs. This process design mapping must also satisfy the Independence Axiom, although sometimes the solution may simply involve the use of existing processes. When existing processes must be used to minimize capital investment in new equipment, the corresponding PVs must also be used to complete the mapping from the physical domain to the process domain. Therefore, the PVs act as constraints in choosing DPs. Since early design decisions may determine 70 to 80% of manufacturing productivity in many enterprises, the product design and process design (or selection) should be performed at the same time in order to develop designs that can be manufactured without incurring cost overruns and schedule slippage. This is sometimes called *concurrent design*.

The Second Axiom: The Information Axiom

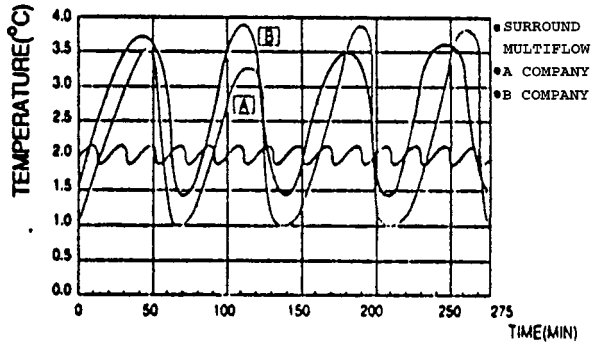
When there is only one FR, the Independence Axiom is always satisfied, and the only task left is to optimize the given design. Various optimization techniques have been advanced to deal with optimization problems involving one objective function. However, when there are two or more FRs, some of these optimization techniques do not work. In these cases, we must first develop a design that is either uncoupled or decoupled. If the design is uncoupled, it can be seen that each FR can be satisfied and the optimum points can be found. If the design is decoupled, the optimization technique must follow a set sequence.

For a given set of FRs, it is most likely that every designer will come up with different designs, all of which are acceptable in terms of the Independence Axiom. However, one of these designs is likely to be the superior alternative. The Information Axiom provides a quantitative means for establishing the merits of a given design, and this value is used to select the best solution. Specifically, the Information Axiom may be stated as



a

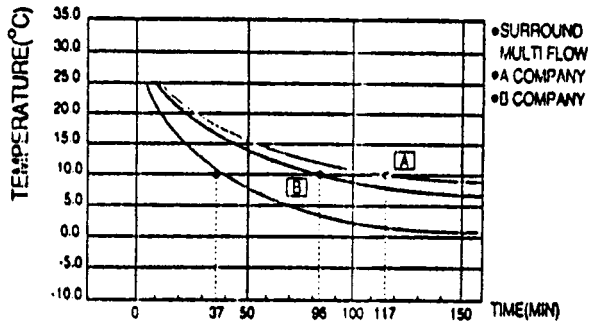
■ +2°C fixed temperature refrigeration
(No variation of temperature)



DIVISION	SURROUND MULTI AIR FLOW	CONVENTIONAL MULTI AIR FLOW	
	FR-820NT	A COMPANY	B COMPANY
TEMPERATURE VARIATION DEGREE	0.6°C	1.5°C	1.9°C

b

FIGURE 11.4.6 Comparison of temperature control.



DIVISION	SURROUND MULTI AIR FLOW	CONVENTIONAL MULTI AIR FLOW	
	FR-820NT	A COMPANY	B COMPANY
ΔT 15°C REFRIGERATE TIME	37 MINUTE	117 MIN	96 MIN

FIGURE 11.4.7 Quick refrigeration performance.

Axiom 2: The Information Axiom
Minimize the information content.

Information is defined in terms of the information content **I** that is related, in its simplest form, to the probability of satisfying a given set of FRs. If the probability of success is p, the information content **I** associated with the probability is defined as

$$\mathbf{I} = -\log_2 p \tag{11.4.5}$$

Equation (11.4.5) defines information content in the units of binary digits or bits. In the general case of an uncoupled design with *n* FRs, **I** may be expressed as

$$I = -\sum_{i=1}^{n} \log_2 p_i = \sum_{i=1}^{n} I_i \tag{11.4.6}$$

where p_i is the probability of DP_i satisfying FR_i. Since there are *n* FRs, the total information content is the sum of all the individual measures. When all probabilities p_i are equal to one, the information content is zero. Conversely, the information content is infinite when one or more probabilities are equal to zero.

A design is called *complex* when its probability of success is low. The quantitative measure for complexity is the information content: complex systems require more information to make the system function (see Equation 11.4.5). Thus, a large system that is comprised of many subsystems and components is not necessarily complex. Even a small system can be complex if its probability of success is low.

Example 3: Cutting a Rod to a Length

Suppose we need to cut rod A to the length 1 +/– 0.000001 m and rod B to the length 1 +/– 0.1 m. Which task is more complicated?

The answer depends on the cutting equipment available for the job! However, most engineers with some practical experience would say that the one that has to be cut within 1 μm would be more difficult, because the probability of success associated with the smaller tolerance is lower than that associated

with the larger tolerance. Therefore, the job with the lower probability of success is more *complex* than the one with higher probability. The Information Axiom links the notion of complexity with the specified tolerances for the FRs: the tighter the tolerance, the more difficult it may be to choose a design solution or a system that can satisfy the FRs.

In practice, the probability of success is a function of the intersection of the tolerances defined by the designer to satisfy the FRs and the ability of the chosen system to produce the part within the specified tolerances. This value can be computed by specifying the *design range* (*dr*) for an FR and by determining the *system range* (*sr*) that the proposed design can provide to satisfy the FR. Figure 11.4.8 illustrates these two ranges graphically. The vertical axis (the ordinate) is for the probability density function and the horizontal axis (the abscissa) is for either FR or DP, depending on the mapping domains involved. When the mapping is between the functional and the physical domains (as in product design), the abscissa is for FR, whereas for mapping between the physical and the process domains (e.g., process design), the abscissa is for DP.

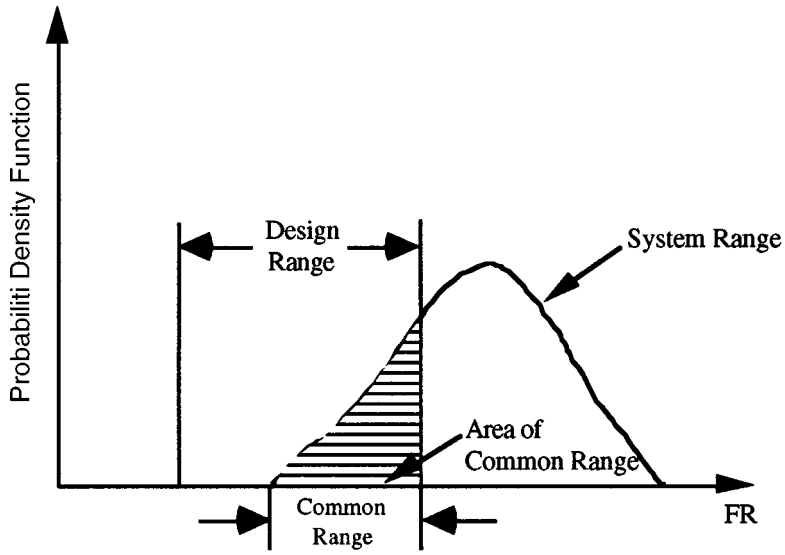


FIGURE 11.4.8 Design range, system range, and common range in a plot of the probability density function (pdf) of a functional requirement.

In Figure 11.4.8 the system range is plotted as a probability density function vs. the specified FR. The intersection of the design range and the system range is called the *common range* (*cr*), and this is the only region where the design requirements are satisfied. Consequently, the area under the common range divided by the area under the system range is equal to the design's probability of success. Then the information content may be expressed as

$$\mathbf{I} = \log_2\left(A_{sr}/A_{cr}\right) \tag{11.4.7}$$

where A_{sr} denotes the area under the system range and A_{cr} is the area within the common range. Furthermore, since A_{sr} =1.0 and there are *n* FRs to satisfy in most cases, the information content may be expressed as

$$I = -\sum_{i=1}^{n} \log_2\left(1/A_{cr}\right)_i \tag{11.4.8}$$

References

Suh, N.P. 1990. *The Principles of Design.* Oxford University Press, New York.

Suh, N.P. June 1995. Axiomatic design of mechanical systems. *Special 50th Anniversary Combined Issue of the Journal of Mechanical Design and the Journal of Vibration and Acoustics. Trans. ASME.* 117:1–10.

11.5 Algorithmic Approaches to Design

Leonard D. Albano

In addition to the axiomatic approach to design (see Section 11.4), there are many methods that are based on an algorithmic approach to design. In algorithmic design, the design process is identified or prescribed so that it leads the designer to a solution that satisfies the design goals. Algorithmic methods can be divided into several categories: pattern recognition, associative memory, analogy, experientially based prescription, extrapolation, interpolation, selection based on probability, etc. This section briefly discusses three common algorithmic methods for design: systematic design, the Taguchi method, and design for assembly (DFA).

Systematic Design

Systematic design methods prescribe how design should be done and provide standardized solutions for fulfilling certain functional requirements. Examples of systematic design methods include Pahl and Beitz (1988), Hubka and Eder (1988), and the German Standards Institute (VDI). This section describes the approach advanced by Pahl and Beitz.

The method of Pahl and Beitz (1988) divides the design process into a number of steps and prescribes methods for dealing with each step. Design starts with an appreciation of customers needs, which are then clarified in terms of an overall function. By definition, the overall function of an engineering system is to convert matter (e.g., energy, materials, signals). Energy, for example, can be converted from chemical energy to mechanical and thermal energy; materials can be shaped, finished, and coated to provide particular geometries and surface finishes; and signal conversion is often used to control the conversion of energy and materials. The overall function is an abstract formulation of customers needs and is defined in terms of inputs and outputs to a system.

Next, the overall function is decomposed into a hierarchy of generic subfunctions (i.e., energy, material, and signal conversions) that the product must perform in order to meet the overall function. The combination of subfunctions is termed a *function structure*. Use of the function structure facilitates the design task by breaking down the overall, complex function into smaller problems that can be divided among a number of experts. Construction of the function structure starts with the logical analysis of the functions that must appear in the solution if the overall function is to be satisfied. For example, a device for producing small, thermoplastic parts should include the following subfunctions: feed plastic pellets, melt plastic pellets, shape molten plastic into desired part geometry, cool part, dispense part, and control material flow. This function structure may be expanded during later steps of the design process as the designer gains additional insight.

Once a simple function structure is established, the designer searches for suitable solution principles to satisfy each of the subfunctions. Solution principles consist of some underlying physical principle or principles to effect the required conversion and some geometric form. Shear and torsion, for example, are physical principles for transferring mechanical energy, and the screw drive is a geometric form that embodies these effects. Published design catalogs, such as the VDI guidelines (VDI), provide a data base of standardized solution principles which facilitates the solution process. A design concept is obtained by integrating the solution principles for each subfunction.

The quality of the proposed design concept increases with the number of solution principles that are considered for each subfunction. Therefore, multiple solution principles should be identified for each subfunction, and these solution principles can be combined in various permutations to produce a rich solution field. Formulating the required subfunctions and the proposed principles in matrices provides a convenient scheme for organizing the results (see Figure 11.5.1). Each row of the matrix refers to a subfunction and the columns of the matrix contain combinations of solution principles.

The reduction of the solution field to a few promising proposals for further development involves a number of ad hoc strategies. One strategy relies on the experience of the designer to identify those

Sub-functions		Solution Concepts					
		1	2	...	j	...	m
1	F_1	S_{11}	S_{12}		S_{1j}		S_{1m}
2	F_2	S_{21}	S_{22}		S_{2j}		S_{2m}
:	:	:	:		:		:
i	F_i	S_{i1}	S_{i2}		S_{ij}		S_{im}
:	:	:	:		:		:
n	F_n	S_{n1}	S_{n2}		S_{nj}		S_{nm}

FIGURE 11.5.1 Matrix approach for developing solution concepts based on mapping different principles S_{xy} to each subfunction F_x.

combinations of solution principles that are not compatible with customers needs or with one another. A second strategy is to use value analysis to evaluate and compare the solution variants generated by the matrix combination. The general procedure for value analysis involves identifying evaluation criteria and assigning relative weights to each criterion. Each solution variant is then assigned a value by combining its weighted value for each criterion. The solution variants are then compared on the basis of their total weighted value, and the best solution is selected.

The selected solution concept is then developed in terms of its layout and detail specifications. Much of the design development relies on design rules and knowledge readily available in handbooks and design guidelines. In addition to the development of production drawings and specifications, detail design often involves optimization of the solution principles and their geometric forms. The reader is referred to Pahl and Beitz (1988) for additional information.

The Taguchi Method

The Taguchi method (1987, 1993) provides a mathematical basis for analyzing product robustness and is intended as a guide for improving design quality. According to Taguchi, higher quality products satisfy customers needs with less variation and are manufactured at lower cost. This definition of quality is based on a cost model that emphasizes the costs associated with product variability. Costs due to variation include those incurred during manufacturing (such as the cost of materials, machine adjustments, and scraps) and those assumed by the customer (e.g., the cost to repair and/or replace the deficient product). The significance of the Taguchi cost model is the fact that it forces the designer to focus on *offline quality control* as a means to improve product quality.

Offline quality control (QC) is a strategy for reducing variability and improving quality during product design and process planning. It is an attempt to take advantage of the fact that most of the cost is committed during the early stages of product development, while only a very small percentage of the cost is actually expended. In contrast, the concept of *online quality control* has been used historically to advance quality products during manufacturing operations. However, during this stage of the product life cycle, much of the cost has been expended and committed. Statistical process control is an example of online quality control.

The practice of offline QC divides the design process into three stages: system design, parameter design, and tolerance design. System design refers to the conceptual phase of design when customers needs are formulated into a design problem and solutions are generated and evaluated. Once a solution is established and defined in terms of its characteristic attributes or parameters, parameter design is used to establish appropriate parameter settings so as to reduce the design's sensitivity to uncontrollable sources of variation. Taguchi refers to these conditions as *noise factors*. Environmental factors (e.g., people and ambient temperature) and time-dependent phenomena (such as tool wear and material

shrinkage) are examples of noise factors. As a last step, tolerance design is used to enhance or fine-tune the quality improvements realized during parameter design. However, tolerance design often involves a trade-off between the improved quality and the increased production costs incurred by tightening tolerances.

Parameter and tolerance design rely on experimental and mathematical techniques to determine the best design, subject to various noise factors. Various strategies may be used to conduct the experiments. Consider a system that involves five design parameters and three levels of settings — high, medium, and low — for each parameter. The simplest strategy is to investigate all possible combinations of parameters and level settings, which would require $3^5 = 243$ sets of experiments. Instead, orthogonal arrays are an efficient and economical alternative to complete enumeration.

Figure 11.5.2 shows a portion of the L18 orthogonal array that enables a reduction in the work scope from 243 to 18 sets of experiments. The column headings A, B, C. D, and E correspond to each of the five design parameters or controllable factors, and the column entries refer to the three level settings 1, 2, and 3. Each row of the array defines an experiment as a unique combination of parameter settings. Any two columns have nine combinations of level settings — (1,1), (1,2), (1,3), (2,1), (2,2), (2,3), (3,1), (3,2), and (3,3) — and each combination appears with the same frequency. This balance is indicative of the array's orthogonality in a statistical sense because the influence of each control factor can be evaluated independently. The basic array can be extended to include the influence of noise factors and their level settings.

Experiment	Parameter Level Setting				
	A	B	C	D	E
1	1	1	1	1	1
2	2	2	2	2	2
3	3	3	3	3	3
4	1	1	2	2	3
5	2	2	3	3	1
6	3	3	1	1	2
7	1	2	1	3	2
8	2	3	2	1	3
9	3	1	3	2	1
10	1	3	3	2	2
11	2	1	1	3	3
12	3	2	2	1	1
13	1	2	3	1	3
14	2	3	1	2	1
15	3	1	2	3	2
16	1	3	2	3	1
17	2	1	3	1	2
18	3	2	1	2	3

FIGURE 11.5.2 Portion of L18 orthogonal array to investigate all combinations of five parameters with three level settings.

Analysis of variation techniques provide a mathematical basis for organizing and interpreting the experimental results. For each performance requirement, signal-to-noise (S/N) ratio is used to express how sensitive each design parameter is to uncontrollable noise (which is indicative of *design robustness*). A robust design has a high S/N ratio and performs well despite the influence of noise. Many equations are available for calculating S/N ratios, and the selection of the appropriate equation is a function of the type of cause–effect relationship under study. Examples include *higher-is-better, lower-is-better,* and *nominal-is-best*. The reader is referred to Taguchi (1987, 1993) for further information.

Design for Assembly

Design for assembly (DFA) supports the analysis and design of products for ease of assembly. Based on a collection of empirical time-study data, if–then rules, and design checklists, DFA methods and tools help the designer to focus on the relationship between the geometric features of a design and its

components and the effort and resources necessary to assemble these components into the desired product. Philips International, Hitachi, Xerox, Ford, and General Electric have been industry leaders in the adoption and dissemination of DFA. In this section the most widely known method, advanced by Boothroyd and Dewhurst (1985), is considered. The method gives the total assembly time and design efficiency, and these measures are then used to guide redesign.

The method of Boothroyd and Dewhurst (1985) starts with the description of a product's assembly sequence in terms of components, component features, and the basic assembly tasks of handling and insertion (which includes fastening). For each component involved in the assembly sequence, a catalog of generic part features is used to classify the component with respect to its difficulty for handling and insertion. These classifications are then used to determine the time required for each assembly task, and the task times are added to give the total operation time for the component. Therefore, the total assembly time (TM) for the design is the sum of the operation times for each component.

Design efficiency is the ratio of an estimated, ideal assembly time to TM. The ideal assembly time is based on a theoretical minimum number of parts (NM). The estimation of NM relies on a checklist for identifying redundant components or components that may be combined to reduce parts count. For manual assembly, the ideal assembly time is specified as 3 sec per part, and design efficiency EM is given by the equation

$$\mathrm{EM} = 3 \times \mathrm{NM}/\mathrm{TM} \tag{11.5.1}$$

The resulting values for assembly time and design efficiency provide the designer with a basis for redesign. Suggested strategies for redesign include reduction in parts count, use of symmetry to facilitate handling, and use of alternative fastening mechanisms to facilitate insertion (e.g., snap-fit elements are faster to insert than parts requiring screw tightening). The reader is referred to Boothroyd and Dewhurst (1985) for further information.

References

Boothroyd, G. and Dewhurst, P. 1985. *Design for Assembly Handbook.* Boothroyd and Dewhurst Associates, Kingston, RI.

Hubka, V. and Eder, W.E. 1988. *Theory of Technical Systems.* Springer-Verlag, New York.

Pahl, G. and Beitz, W. 1988. *Engineering Design.* The Design Council, London.

Taguchi, G. 1987. *System of Experimental Design*, translated by Louise Watanabe Tung. Kraus International Publications, White Plains, NY.

Taguchi, G. 1993. *Taguchi on Robust Technology Development*. translated by S-C Tsai. ASME Press, New York.

VDI. 1987. *VDI Design Handbook 2221: Systematic Approach to the Design of Technical Systems and Products.* German Standards Institute, Dusseldorf.

11.6 Strategies for Product Design

Michael Pecht

Requirements

The design team must have a clear appreciation of user requirements and constraints before a design can begin. Requirements are established within the context of product effectiveness attributes and include limits on parameters such as speed, miles per gallon, computations per second, and accuracy, as well as constraints on size, weight, reliability, and cost. Often, contractual requirements and company policy can dictate a product type, such as the use of an in-house technology or product, or the exclusion of a technology or product.

Requirements must be addressed holistically in the design of a product and must not be limited to those that affect the product's immediate performance. For example, the supportability of the product is a critical product requirement for many products in terms of the ease of maintenance and accessibility to the products, spares, support equipment, the time and required personnel to repair the product, and the ease of discovering or detecting where a problem occurs. A poor definition of requirements could lead to a design where, for example, the air conditioning compressor has to be removed to replace a spark plug, or where special tools are required to replace the oil in a car. Note also that the design team should be aware of the impact of administrative policies on the reliability of products as reflected in the product's scenario for use. For example, outdated standards can contribute to failures that are not the result of either faulty design or human error.

The Life Cycle Usage Environment

The life cycle usage environment or scenario for use of a product goes hand-in-hand with the product requirements. The life cycle usage information describes the storage, handling, and operating stress profiles and thus contains the necessary "load" input information for effective assessment and development of design guidelines, screens, and tests. The stress profile of a product is based on the application profile and the internal stress conditions of the product. Because the performance of a product over time is often highly dependent on the magnitude of the stress cycle, the rate of change of the stress, and even the time and spatial variation of stress, the interaction between the application profile and internal conditions must be specified. Specific information about the product environment includes absolute temperature, temperature ranges, temperature cycles, temperature gradients, vibrational loads and transfer functions, chemically aggressive or inert environments, and electromagnetic conditions. The life cycle usage environment can be divided into three parts: the application and life profile conditions, the external conditions in which the product must be stored, handled, and operated, and the internal product-generated stress conditions.

The application and life profile conditions include the application length, the number of applications in the expected life of the product, the product life expectancy, the product utilization or nonutilization (storage, testing, transportation) profile, the deployment operations, and the maintenance concept or plan. This information is used to group usage platforms (i.e., whether the product will be installed in a car, boat, satellite, underground), develop duty cycles (i.e., on–off cycles, storage cycles, transportation cycles, modes of operation, and repair cycles), determine design criteria, develop screens and test guidelines, and develop support requirements to sustain attainment of reliability and maintainability objectives. The external operational conditions include the anticipated environment(s) and the associated stresses that the product will be required to survive. The stresses include temperature, vibrations, shock loads, humidity or moisture, chemical contamination, sand, dust and mold, electromagnetic disturbances, radiation, etc.

The internal operational conditions are associated with product-generated stresses, such as power consumption and dissipation, internal radiation, and release or outgassing of potential contaminants. If

the product is connected to other products or subsystems in a system, stresses associated with the interfaces (i.e., external power consumption, voltage transients, electronic noise, or dissipation) must also be included.

Characterization of Materials, Products, and the Manufacturing Processes

Design is intrinsically linked to the materials, products, interfaces, and manufacturing processes used to establish and maintain functional and structural integrity. It is unrealistic and potentially dangerous to assume defect-free and perfect-tolerance materials, products, and structures. Materials often have naturally occurring defects and manufacturing processes can induce additional defects to the materials, products, and structures. The design team must also recognize that the production lots or vendor sources for products that comprise the design are subject to change. Even greater variability in products characteristics is likely to occur during the fielded life of a product as compared to its design or production life cycle phases.

Design decisions involve the selection of products, materials, and controllable process techniques, using tooling and processes appropriate to the scheduled production quantity. Often, the goal is to maximize product and configuration standardization; increase package modularity for ease in fabrication, assembly, and modifications; increase flexibility of design adaptation to alternate uses; and utilize alternate fabrication processes.

Design Guidelines and Techniques

Generally, new products replace existing products. The replaced product can be used for comparisons (i.e., a baseline comparison product). Lessons learned from the baseline product can be used to establish new product parameters, to identify areas of focus in the new product design, and to avoid the mistakes of the past.

Once the products, materials, processes, and stress conditions have been characterized, design begins. In using design guidelines, there may not be a unique path to follow. Multiple branches may exist depending upon the input design constraints. The design team should explore enough of the branches to gain confidence that the final design is the best for the prescribed input information. The design team should also use guidelines for the complete design and not those limited to specific aspects of an existing design. This statement does not imply that guidelines cannot be used to address only a specific aspect of an existing design, but the design team may have to trace through the implications that a given guideline suggests.

Design guidelines that are based on physics of failure models can also be used to develop tests, screens, and derating factors. Tests can be designed from the physics of failure models to measure specific quantities and to detect the presence of unexpected flaws, manufacturing, or maintenance problems. Screens can be designed to precipitate failures in the weak population while not cutting into the design life of the normal population. Derating or safety factors can be determined to lower the stresses for the dominant failure mechanisms.

Preferred Products

In many cases, a product or a structure much like the required one has already been designed and tested. This is called a “preferred product or structure” in the sense that variabilities in manufacturing, assembly, and field operation that can cause problems have been identified and corrected. Many design teams maintain a list of preferred products and structures with acceptable performance, cost, availability, and reliability.

Redundancy

Redundancy permits a product to operate even though certain components and interconnections have failed. Redundant configurations can be classified as either active or standby. Elements in active redundancy operate simultaneously in performing the same function. Elements in standby redundancy are

designed so that an inactive one will be switched into service when an active element fails. The reliability of the associated function is increased with the number of standby elements (optimistically assuming that the sensing and switching products of the redundant configuration are working perfectly, and failed redundant components are replaced before their companion component fails).

A design team may often find that redundancy is a way to improve product reliability if time is of prime importance, if the products requiring redundancy are already designed and the products are known to be of poor reliability.

On the other hand, in weighing its advantages, the design team may find that redundancy will

- Prove too expensive, if the products and redundant sensors and switching products are costly
- Exceed the limitations on size and weight, particularly in avionics, missiles, and satellites
- Exceed the power limitations, particularly in active redundancy
- Attenuate the input signal, requiring additional amplifiers (which increase complexity)
- Require sensing and switching circuitry so complex as to offset the reliability advantage of redundancy

Protective Architectures

It is generally desirable to include means in a design for preventing a product, structure, or interconnection from failing catastrophically, and instead allow it to fail safely.

Fuses and circuit breakers are examples used in electronic products to sense excessive current drain and disconnect power from a failed circuit. Fuses within circuits safeguard products against voltage transients or excessive power dissipation and protect power supplies from shorted parts. Thermostats may be used to sense critical temperature, limiting conditions and shutting down the product or a component of the system until the temperature returns to normal. In some products, self-checking circuitry can also be incorporated to sense abnormal conditions and operate adjusting means to restore normal conditions or activate switching means to compensate for the malfunction.

Protective architectures can be used to sense failure and protect against possible secondary effects. In other cases, means can be provided for preventing a failed product or structure from completely disabling the product. For example, a fuse or circuit breaker can disconnect a failed product from a product in such a way that it is possible to permit partial operation of the product after the failure, in preference to total product failure. By the same reasoning, degraded performance of a product after failure of a product is often preferable to complete stoppage. An example is the shutting down of a failed circuit whose design function is to provide precise trimming adjustment within a deadband of another control product. Acceptable performance may thus be permitted, perhaps under emergency conditions, with the deadband control product alone.

Sometimes the physical removal of a part from a product can harm or cause failure of another part of the product by affecting load, drive, bias, or control. In some cases, self-healing techniques can be employed to self-check and self-adjust to effect changes automatically to permit continued operation after a failure.

The ultimate design, in addition to its ability to act after a failure, would be capable of sensing and adjusting for drifts to avert failures.

In the use of protective techniques, the basic procedure is to take some form of action after an initial failure or malfunction, to provide perhaps reduced performance, and to prevent additional or secondary failures. Such techniques can be considered as enhancing product reliability, although they also affect availability and product effectiveness. No less a consideration is the impact of maintenance, repair, and product replacement. If a fuse protecting a circuit is replaced, what is the impact when the product is reenergized? What protective architectures are appropriate for postrepair operations? What maintenance guidance must be documented and followed when fail-safe protective architectures have or have not been included?

Stress Margins

Products and structures should be designed to operate satisfactorily at the extremes of the parameter ranges, and allowable ranges must be included in the procurement or reprocurement specifications. To guard against out-of-tolerance failures, the design team must consider the combined effects of tolerances on products, subsequent changes due to the range of expected environmental conditions, drifts due to aging over the period of time specified in the reliability requirement, and tolerances in products used in future repair or maintenance functions.

Methods of dealing with product and structural parameter variations include statistical analysis and worst-case analysis. In statistical design analysis, a functional relationship is established between the output characteristics of the structure and the parameters of one or more of its products. In worst-case analysis, the effect that a product has on product output is evaluated on the basis of end-of-life performance values or out-of-specification replacement products.

Derating

The principle of derating is that there are no distinct stress boundaries for voltage, current, temperature, power dissipation, etc. above which immediate failure can occur and below which the product will operate indefinitely. Instead, the life of some products increases in a continuous manner as the stress level is decreased below the rated value. Practically, however, there are minimum stress levels below which increased derating will lower reliability, and the complexity required to step up performance will offset any gain in reliability.

Size and Weight Control

Methods to reduce product volume include:

- Eliminating unused space
- Selecting an exterior shape to require least possible volume when integrated with the intended installation structure
- Eliminating separable interconnections, using optimally shaped components and subassemblies
- Minimizing interconnection requirements
- Eliminating support components such as heat sinks, blowers, and special coolant flow passages
- Eliminating redundant components
- Using components and package elements to perform multiple functions

Methods to lower product weight include using high-strength lightweight materials, eliminating duplicate structures, and providing only that degree of strength required to reach a desired safety factor.

Potential Failure Sites and Failure Mechanisms

The design team must evaluate the potential failure mechanisms, failure stresses, failure sites, and failure modes, given a product architecture, the comprising products and materials, and the manufacturing processes. One approach to evaluation consists of the assessment of a list of candidate or "potential" failure mechanisms. The load conditions that cause the failure mechanisms to occur are then determined in light of the life cycle usage environment identified earlier. Table 11.6.1 presents various failure mechanisms and the associated load conditions.

Once the failure mechanisms and associated stresses have been identified, then failure sites are specified and a detailed reliability assessment is conducted for each location suspected to be a potential area of failure. Table 11.6.2 presents an example case of various failure sites, mechanisms, and load structures for a microelectronic product.

Failures can be categorized by their impact on end product performance. A critical failure is an event that reduces the performance of the end product to unacceptable levels. A noncritical failure, also called a fault, does not reduce performance to unacceptable levels. Failures can also be classified by the time

TABLE 11.6.1 The Load Conditions which Cause Failure Mechanisms

Failure Mechanism	Containment	Δ T (Temperature Cycle Magnitude)	RH (Relative Humidity)	T (Steady-State Temperature)	Vibration/Shock	Maintenance and Handling	Voltage
Brittle fracture		X		X	X	X	
Ductile fracture				X	X		
Yield		X		X	X		
Buckling		X		X	X	X	
Large elastic deformation		X			X		
Interfacial deadhesion	X	X	X	X	X		
Fatigue crack initiation	X	X	X	X	X		
Wear					X		
Creep		X		X	X		
Corrosion	X		X	X		X	
Dendritic growth	X		X			X	X
Fatigue crack propagation		X			X	X	
Diffusion				X			X

TABLE 11.6.2 Example: Failure Sites, Operational Loads, and Failure Mechanisms for a Microelectronic Product

Site	Containment	Δ T (Temperature Cycle Magnitude)	RH (Relative Humidity)	T (Steady-State Temperature)	Vibration/Shock	Voltage	Maintenance and Handling
Die		A, L, G	I		A, L, G		
Die attach		A, F, L, G			A, L, G, C		
Flip-chip solder joints	I, K	L, G	I	B, C, I, N	L, G	M	
Tape automated bonds	I, J, K	L, G	I, J	A, B, I, J, M	L, G	M, J	
Wire	I	G, L	I		G, L, E		
Leads	I, J	L	I, J	I	D, L, H	L, J	D, G
Substrate		L, G			A, B, L, G		
Substrate attach		F, L, G			C, L, G		

Note: A — brittle fracture; B — ductile fracture; C — yield; D — buckling; E — large elastic deformation; F — interfacial deadhesion; G — fatigue crack initiation; H — wear; I — corrosion; J — dendritic growth; K — interdiffusion; L — fatigue crack propagation; M — diffusion.

frame and manner in which the event occurs. The categories are catastrophic, intermittent, out-of-tolerance, and maladjustment.

1. **Catastrophic failures.** A catastrophic failure occurs when a product becomes completely inoperative or exhibits a gross change in characteristics. Examples are an open or shorted resistor or capacitor, a leaky valve, a stuck relay, or a broken switch.
2. **Intermittent failures.** Intermittent failures are nonperiodic failures that can occur within products, interconnects, or product interfaces including software. Examples include switch bouncing and poor transmission of signals. The design team must not only safeguard against the effects of intermittent failures, but also avoid creating possible modes of such failures (e.g., conditions permitting product-to-product or structural breakdowns).
3. **Out-of-tolerance failures.** Out-of-tolerance failures result from degradation, deterioration, drift, and wearout. Examples are the drifting of resistor and capacitor values and the wearing out of relays and solenoid valves and precision gears. The changes can arise as a result of time, temperature or temperature cycles, humidity, altitude, etc. When these changes, considered collectively, reach the point where product performance is below acceptable limits, we say that the product has failed.
4. **Maladjustment failures.** Maladjustment failures are often due to human error. Failures arise from improper adjustment of products, as well as abuse of adjusting products due to lack of understanding of the adjustments and the capabilities of the products. Such failures are hard to evaluate and difficult to avoid but must be considered if the reliability of the product is to be sustained.

Designing for the Application Environment

The design of modern products requires that the team be acquainted with the environmental factors affecting product performance over time. The environment is seldom forgiving and when a weak point exists, performance suffers. Design teams must understand the environment and its potential effects and then must select designs or materials that counteract these effects or provide methods to alter or control the environment within acceptable limits.

In addition, a product can create or perturb an environment. For example, many epoxies out-gas during cure, releasing corrosive or degrading volatiles or particulates into the product. Teflon may release fluorine and polyvinylchloride (PVC) may release chlorine. Certain solid rocket fuels are degraded into a jelly-like mass when exposed to aldehydes or ammonia, either of which can come from a phenolic nozzle cone. Common environmental considerations follow.

Temperature Control

Temperature is a powerful agent for electrical, chemical, and physical deterioration for two basic reasons:

- The physical properties of almost all known materials are modified by changes in temperature, temperature gradients, and temperature extremes.
- The rate of most chemical reactions is influenced by the temperature of the reactants.

Nevertheless, the effects of steady-state temperature and spatial and temporal temperature gradients must be clearly understood before a thermal control method is determined. A good practice is to develop a table of potential failure sites and failure mechanisms and the effects of temperature in its many forms.

Poor heat removal of dissipated I^2R losses, hysteresis, or eddy current losses can result in physical damage or accelerated chemical reaction rates. The latter occurrence, affecting certain types of materials, can cause general degradation and catastrophic or intermittent failures. But the fault is not always with the design; air intakes or outlets could be blocked because of faulty installation, or maintenance personnel might fail to clear or replace air filters.

Thermal management schemes for electronic equipment are selected, once the heat flux and the maximum temperature difference available for heat transfer have been identified. A complete thermal

management strategy involves selecting an appropriate heat removal scheme for each level of packaging. For low heat fluxes, passive thermal management techniques can be used. Such techniques do not require the expenditure of external energy for the heat removal. Interest in such techniques is currently very strong, due to their design simplicity, low cost, and high reliability.

Examples of passive cooling techniques include conduction cooling using high thermal conductivity substrates and/or heat sinks and natural convection air cooling. Conduction cooling can be employed when the heat source is not directly exposed to the coolant. This method is desirable in dense electronic packages, where radiation and convection are not effective cooling means. Natural convection cooling relies on the buoyancy-induced flows generated due to heating. For a maximum temperature rise of 85°C above the environment, heat fluxes of up to about 0.1 W/cm^2 can be removed by natural convection air cooling under nominal sea-level environmental conditions.

For higher heat fluxes, a combination of passive and active techniques can sometimes be used. This may include, for example, a high thermal conductivity substrate at the board level, along with forced air cooling at the box level. For this range of heat fluxes, the use of cold plate technology, thermoelectrics, and flow-through cooling are also possible; other common techniques are the use of heat pipe or thermosyphon (see Chapter 4).

Besides the out-gassing of corrosive volatiles when subjected to heat, almost all known materials will expand or contract when their temperature is changed. This expansion and contraction causes problems with fit between product interface and interconnections. Local stress concentrations due to nonuniform temperature are especially damaging, because they can be so high. A familiar example is a hot water-glass that shatters when immersed in cold water. Metal structures, when subjected to cyclic heating and cooling, may ultimately collapse due to the induced stresses and fatigue caused by flexing. The thermocouple effect between the junction of two dissimilar metals causes an electric current that may induce electrolytic corrosion.

Plastics, natural fibers, leather, and both natural and synthetic rubber are all particularly sensitive to temperature extremes as evidenced by their brittleness at low temperatures and high degradation rates at high temperatures.

Shock and Vibration Control

Shock and vibration can harmfully flex leads and interconnects, dislodge parts or foreign particles into bearings, pumps, and electronics, cause acoustical and electrical noise, and lead to structural instabilities. Protective measures against shock and vibration stresses are generally determined by an analysis of the deflections and mechanical stresses produced by these environmental factors. This generally involves the determination of natural frequencies and evaluation of the mechanical stresses within components and materials produced by the shock and vibration environment. If the mechanical stresses so produced are below the acceptable safe working stress of the materials involved, no direct protection methods are required. If, on the other hand, the stresses exceed the safe levels, corrective measures such as stiffening, reduction of inertia and bending moment effects, and incorporation of further support members are indicated. If such approaches do not reduce the stresses below the acceptable safe levels, further reduction is usually possible by the use of shock-absorbing mounts.

Products are sometimes specially mounted to counter the destructive effects of shock and vibration. Shock mounts often serve this purpose, but effective means for attenuating shock and vibration simultaneously are complex. Isolation of a product against the effects of vibration requires that the natural frequency of the product be substantially lower than the undesired frequency of vibration.

Three basic kinds of isolators are available:

- Elastomers made of natural or synthetic rubber, used in a shear mode or in a diaphragm to dampen the induced shock or vibration.
- Metallic isolators, such as springs, metal meshes, or wire rope. Springs lack good damping qualities, but meshes and rope provide smooth friction damping.

- Viscous dampers (similar to the type used on automobiles) which are velocity-sensitive, tend to become ineffective under high-frequency vibration. Resilient mounts must be used with caution, since they can amplify the intensity of shock and vibration if improperly placed. The ideal goal is to design a product to be resistant to shock and vibrations, rather than to provide complete isolation.

Another vibration protective technique is potting or encapsulation of small parts and assemblies. The potting material should be compliant enough to dampen vibrations. Some materials polymerize exothermically, and the self-generated heat may cause cracking of the casting or damage to heat-sensitive products. Another problem is shrinkage of the potting material. In some cases, molds are made oversized to compensate for shrinkage.

One factor that is not often considered is that the vibration of two adjacent components, or separately insulated subsystems, can cause a collision if maximum excursions and sympathetically induced vibrations are not accounted for in design. Another failure mode, fatigue (the tendency for a material to break under cyclic stress loads considerably below its tensile strength), includes high cycle fatigue, acoustic fatigue, and fatigue under combined stresses such as temperature extremes, temperature fluctuations, and corrosion.

In addition to using proper materials and configuration, it may still be necessary to control the amount of shock and vibration experienced by the product. Damping products are used to reduce peak oscillations and special stabilizers can be employed when unstable configurations are involved. Typical examples of dampeners are viscous hysteresis, friction, and air damping. Vibration isolators commonly are identified by their construction and material used for the resilient elements (rubber, coil spring, woven metal mesh, etc.). Shock isolators differ from vibration isolators in that shock requires a stiffer spring and a higher natural frequency for the resilient element. Some of the types of isolation mounting products are underneath, over-and-under, and inclined isolators. In some cases, however, even though an item is properly insulated and isolated against shock and vibration damage, repetitive forces may loosen the fastening products. If the fastening products loosen enough to permit additional movement, the product will be subjected to increased forces and may fail. Many specialized self-locking fasteners are commercially available.

Chemical Action Control

The earth's environment contains numerous deteriorators including oxygen, carbon dioxide, nitrogen, snow, ice, sand, dust, salt water spray, organic matter, and chemicals in general. Product specifications often state limits on temperature, humidity, altitude, salt spray, fungus, sunshine, rain, sand, and dust.

A material or structure can chemically change in a number of ways. Among them are interactions with other materials (i.e., metal migration, diffusion) and modifications in the material itself (recrystallization, stress relaxation, phase changes, or changes induced by irradiation). In addition to the deterioration problems associated with the external environments to which products are subjected, adhesives, batteries, and certain types of capacitors are susceptible to chemical aging and biological growths.

Materials widely separated in the electromotive series are subject to galvanic action, which occurs when two dissimilar metals are in contact in a liquid medium. The most active metal dissolves, hydrogen is released, and an electric current flows from one metal to the other. Coatings of zinc are often applied to iron so that the zinc, which is more active, will dissolve and protect the iron (a process known as *galvanization*). Galvanic action is known to occur within the same piece of metal, if one portion of the metal is under stress and has a higher free-energy level than the other. The part under stress will dissolve if a suitable liquid medium is present. Stress-corrosion cracking occurs in certain magnesium alloys, stainless steels, brass, and aluminum alloys. It has also been found that a given metal will corrode much more rapidly under conditions of repeated stress than when no stress is applied.

Corrosion-resistant materials should be used as much as possible. Although aluminum is excellent in this respect due to the thin oxide coating that forms when it is exposed to the atmosphere, it tends to

pit in moist atmospheres. Furthermore, aluminum may corrode seriously in a salt-laden marine atmosphere. Either anodization or priming with zinc chromate can provide additional protection.

Magnesium is sometimes used to minimize the weight of products but must be protected by surface additives against corrosion and electrolytic reaction. A coating of zinc chromate serves both purposes. Because magnesium is the most reactive metal normally used for structural purposes, the grounding of the magnesium structure requires careful selection of another metal for making the connection. Zinc- or cadmium-plated steel are the more commonly chosen connective materials.

Copper, when pure, is quite resistant to corrosion. However, under certain conditions, copper-made products will become chemically unstable. Transformers have occasionally failed due to impurities in the insulation and moisture. These conditions favor electrolytic reaction, which causes the copper to dissolve and eventually results in an open circuit.

Iron and steel are used in many products and structures because of their good magnetic and structural properties; however, only certain stainless steels are reasonably resistant to corrosion. Various types of surface plating are often added, but since thin coatings are quite porous, undercoatings of another metal such as copper are often used as a moisture barrier. Cadmium and zinc plating in some marine environments corrode readily, and often grow "whiskers", which can cause short circuits in an electronic product.

Proper design of a product therefore requires trade-offs in

- Selecting corrosion-resistant materials
- Specifying protective coatings if required
- Avoiding use of dissimilar metallic contacts
- Controlling metallurgical factors to prevent undue internal stress levels
- Preventing water entrapment
- Using high-temperature resistance coatings when necessary
- Controlling the environment through dehydration, rust inhibition, and electrolytic and galvanic protective techniques

Biological Growth Control

High humidity and warm temperatures often favor the growth of fungus, which can lead to deterioration of many electrical and mechanical properties. Products designed for use under tropical conditions can receive a moisture- and fungus-proofing treatment, which consists of applying a moisture-resistant varnish containing a fungicide. Air-drying varnishes generally are not as moisture-proof as are the baked-on types. This treatment is usually quite effective, but the useful life of most fungicides is less than that of the varnish, and retreatments may be necessary.

Moisture Control

Moisture, and impurities that may be contained with it, are known to cause corrosion of many metal systems. In addition, mated products can be locked together, especially when moisture condenses on them and then freezes. Similarly, many materials that are normally pliable at low temperatures become hard and brittle if moisture has been absorbed and subsequently freezes. Condensed moisture acts as a medium for the interaction between many, otherwise relatively inert, materials. Most gases readily dissolve in moisture. The volume increase from water freezing (i.e., converting from a fluid to solid state) can also physically separate components, materials or connections. The chlorine released by PVC plastic, for example, forms hydrochloric acid when combined with moisture.

Although the presence of moisture may cause deterioration, the absence of moisture also may cause reliability problems. The useful properties of many nonmetallic materials such as leather and paper, which become brittle and crack when they are very dry, depend upon an optimum level of moisture. Similarly, fabrics wear out at an increasing rate as moisture levels are lowered, and fibers become dry and brittle. Environmental dust can cause increased wear, friction, and clogged filters due to lack of moisture.

Moisture, in conjunction with other environmental factors, creates difficulties that may not be characteristic of the factors acting alone. For example, abrasive dust and grit, which would otherwise escape, can be trapped by moisture. The permeability (to water vapor) of some plastics (PVC, polystyrene, polyethylene, etc.) is related directly to their temperature. The growth of fungus is enhanced by moisture, as is the galvanic corrosion between dissimilar metals. Some design techniques that can be used separately or combined to counteract the effects of moisture are

- Elimination of moisture traps by providing drainage or air circulation
- Using desiccant products to remove moisture when air circulation or drainage is not possible
- Applying protective coatings
- Providing rounded edges to allow uniform coating of protective material
- Using materials resistant to moisture effects, fungus, corrosion, etc.
- Hermetically sealing components, gaskets, and other sealing products
- Impregnating or encapsulating materials with moisture-resistant waxes, plastics, or varnishes
- Separation of dissimilar metals or materials that might combine or react in the presence of moisture or of components that might damage protective coatings.

The design team also must consider possible adverse effects caused by specific methods of protection. Hermetic sealing, gaskets, protective coatings, etc. may, for example, aggravate moisture difficulties by sealing moisture inside or contributing to condensation. The gasket materials must be evaluated carefully for outgassing of corrosive volatiles or for incompatibility with adjoining surfaces or protective coatings.

Sand and Dust Protection

In addition to the obvious effect of reduced visibility, sand and dust can degrade a product by abrasion leading to increased wear, friction causing both increased wear and heat, and clogging of filters, small apertures, and delicate products. Thus, products having moving parts require particular care when designing for sand and dust protection. Sand and dust will abrade optical surfaces, either by impact when being carried by air or by physical abrasion when the surfaces are improperly wiped during cleaning. Dust accumulations have an affinity for moisture and when combined may lead to corrosion or the growth of fungus.

In relatively dry regions, such as deserts, fine particles of dust and sand can readily be agitated into suspension in the air, where they may persist for many hours, sometimes reaching heights of several thousand feet. Thus, even though there is virtually no wind present, the speeds of vehicles or vehicle-transport product through these dust clouds can also cause surface abrasion by impact.

Although dust commonly is considered to be fine, dry particles of earth, it also may include minute particles of metals, combustion products, and solid chemical contaminants. These other forms may provide direct corrosion or fungal effects on products, because this dust may be alkaline, acidic, or microbiological.

When products require air circulation for cooling or removing moisture, the question is not whether to allow dust to enter, but rather how much or what size dust can be tolerated. The problem becomes one of filtering the air to remove dust particles above a specific nominal size. The nature of filters, however, is such that for a given working filter area, as the ability of the filter to stop increasingly smaller dust particles is increased, the flow of air or other fluid through the filter is decreased. Therefore, the filter surface area either must be increased, the flow of fluid through the filter decreased, or the allowable particle size increased (i.e., invariably, there must be a compromise). Interestingly enough, for aircraft engines, the amount of wear is proportional to the weight of ingested dust, but inversely proportional to dust size.

Sand and dust protection must be planned in conjunction with protective measures against other environmental factors. For example, it is not practical to specify a protective coating against moisture if sand and dust will be present, unless the coating is carefully chosen to resist abrasion and erosion or is self-healing.

Explosion Control

Protection against explosion is both a safety and reliability problem. An item that randomly exhibits explosive tendencies is one that has undesirable design characteristics and spectacular failure modes. This type of functional termination, therefore, requires extreme care in design and reliability analyses.

Explosion protection planning must be directed to three categories (not necessarily mutually exclusive) of products:

- Items containing materials susceptible to explosion
- Components located near enough to cause the explosive items to explode
- Product that might be damaged or rendered temporarily inoperative by overpressure, flying debris, or heat from an explosion

The first category includes products containing flammable gases or liquids, suspensions of dust in the air, compounds that spontaneously decompose in certain environments, product containing or subjected to high or low extremes of pressure (includes implosions), or any other products capable of creating an explosive reaction. Keep in mind that even basically inert materials such as glass or metals can explode when subjected to rapid environmental changes or extremes such as temperature or stress.

The second category includes many variations on methods for providing an energy pulse, a catalyst, or a specific condition that might trigger an explosion. A nonexplosive component, for example, could create a corrosive atmosphere, mechanical puncture, or frictional wear on the side of a vessel containing a high-pressure and thereby cause the container to explode.

The third category is important because a potentially explosive product (such as a high-pressure air tank) can be damaged or made to explode from another explosion. Thus, some reasoning must be applied when considering products not defined by the first two categories. From a practical standpoint, explosion protection for items in the third category should be directed to product that might possibly be near explosions.

The possibility of an explosive atmosphere leaking or circulating into other product compartments must also be recognized. Lead-acid batteries, for example, create hydrogen gas that, if confined or leaked into a small enclosure, could be exploded by electrical arcing from motor brushes, by sparks from metallic impacts, or by exhaust gases. Explosive environments, such as dust-laden air, might be circulated by air distribution products. Dust from common ingredients such as wheat flour has been the source of massive, catastrophic explosions.

Explosion protection and safety are very important for design and reliability evaluations and must be closely coordinated and controlled. Just as a safe product is not necessarily reliable, neither is a reliable product necessarily safe; but the two can be compatible.

Electromagnetic Radiation Control

Protection against the effect of electromagnetic radiation has become a sophisticated engineering field of electromagnetic compatibility design. The radiation environment in space near the earth is composed primarily of Van Allen, auroral, solar flare, and cosmic phenomena. Of lesser importance are solar wind, thermal energy atoms in space, neutrons, naturally occurring radon gas, albedo protons, plasma bremsstrahlung, and man-made nuclear sources. In the electromagnetic spectrum are gamma rays, X-rays, ultraviolet, and Lyman-alpha radiation. Damage near the surface of the earth is caused by the electromagnetic radiation in the wavelength range from approximately 0.15 to 5 m. This range includes the longer ultraviolet rays, visible light, and up to about the midpoint of the infrared band. The most direct approach to protection is to avoid the limited region in which high radiation levels are found. When exposure cannot be avoided, shielding and filtering are important protective measures. In other cases, material design changes or operating procedural changes must be instituted in order to provide protection.

High Vacuum Control

In a high vacuum, materials with a high vapor pressure will sublimate or evaporate rapidly, particularly at elevated temperatures. In some plastics, the loss of plasticizer by evaporation will cause cracking, shrinking, or increased brittleness. Metals such as magnesium, which would normally evaporate rapidly (1 g/cm^2/year at 250°C), can be protected by inorganic coatings with low vapor pressures.

In a high vacuum, adjoining solid surfaces can become cold-welded after losing adsorbed gases. Some form of lubrication is therefore necessary. Conventional oils and greases evaporate quickly. Graphite becomes unsatisfactory (actually an abrasive) because of the loss of absorbed water. However, thin films of soft metals, such as lead, silver, or gold, are effective lubricants in a high vacuum. Thin films of molybdenum disulfide are often sprayed over chrome or nickel plating, forming easily sheared layers. The film also releases sulfur at interfaces during sliding, performing the same function as water vapor does for graphite.

Human Factors and Operability

Humans are active participants in the operating of many products, and can be traded as an external "stress." Consequently, the interaction of humans with products must be weighed against safety, reliability, maintainability, and other product parameters to assess product reliability, maintainability time and performance, system and subsystem safety analyses, and specific human engineering design criteria. For convenience and clarity, four types of human interactions with a product are identified:

- Design and production of a product
- Operators and repairers as mechanical elements (human engineering)
- Operators and repairers as decision elements (human performance reliability)
- Bystanders (this classification is not considered further because it is largely a safety matter, unless they consciously or inadvertently operate or affect the operation of a product)

Designing for Operability

While product design teams, planners, and managers would hope to have well-above-average people in every position associated with product operation, product design must accommodate the realities of the product's life cycle use. The complexity of the product that is presented to the operator, the array of components, including ancillary product, cable connectors, patch panels, switches, knobs, controls, panel markings, meters, oscilloscopes, horns, bells, panel lights, etc., must be of concern in design. The various components comprising the product should be clearly labeled for easy identification. Those items that must be manipulated in order to connect the product (cable connectors, patch panels, switches) should be arranged in a simple and logical order and should be plainly marked. Typical constraints are that

- An operator should be within the physical capabilities of the complete range of potential operators.
- A person should not be required to do something that his or her coordination will not allow him or her to do.
- In times of psychological stress, people cannot easily use, read, and respond to controls and displays.

Mock-ups under realistic conditions are very helpful in uncovering potential problems early in the design process. For example, if a product must be used at night in extremely cold weather, have a person try to use it in a freezing, poorly lit room for several hours.

Designing for Maintainability

Few operational products can be perfectly reliable while meeting other product trade-offs such as rust. Maintenance can thus be an important consideration in the long-term effectiveness of a product.

Maintainability is the probability that, when a specified maintenance action is initiated, a failed product will be restored to operable conditions in a specified downtime. Thus, design features that will expedite maintenance will enhance maintainability. Designing for maintainability means inclusion in the design of those features that can be conceived to assist the maintenance technician. Specific features include the degree of accessibility for product replacement, facilities for fault isolation, special tools or test-product requirements, the level of servicing skills required, servicing documentation requirements, and spare-part stocking requirements.

Critical rating factors include grouping of components by electrical function, use of integral fault indication for basic modules; components or functional assemblies removable without interruption of permanent electrical connections; elimination of tool requirements for mechanical disassembly; direct access to removable assemblies; products commonality; and identification of replaceable components.

In checking out a product, the maintenance technician must often also operate it. Therefore, the design considerations relating to operability are, in general, also applicable to maintainability. In addition, many design considerations relating to maintenance are more complex than those relating to operation.

The design team needs to consider how the product actually will be repaired in the field, perhaps under the pressures of understaffed maintenance crews, many of whom are inexperienced, or by field service personnel with limited knowledge about the product. The design team must always keep in mind that a product may be used and repaired by people who have other things in mind than "babying" the product. The design team must also realize the difference between what people will probably do and what the design team thinks they ought to do.

The design team, in acknowledging that a product could fail if not maintained, should provide means for ease of maintenance, ease of removal of a failed unit, trouble shooting, access to the failed unit, and repair or replacement. Such means, in addition to improving maintainability, will also improve reliability through the averting of subsequent failure due to human errors during maintenance. Many design details are important to the maintenance technician. A list of "Do's" and "Don'ts" in designing for maintainability is presented in Table 11.6.3.

TABLE 11.6.3 Warning to Design Teams: Some Common Design Errors Affecting Maintainability

Don't place products or maintenance structures (e.g., oil filter) where they cannot be removed without removal of the whole unit from its case or without first removing other products
Don't put an adjustment out of arm's reach
Don't neglect to provide enough room for the technician to get his or her gloved hand into the unit to make an adjustment
Don't screw subassemblies together in such a way that the maintenance technician cannot tell which screw holds what
Don't use chassis and cover plates that fall when the last screw is removed
Don't make sockets and connectors for modules of the same configuration so that the wrong unit can be installed
Don't provide access doors with numerous small screws or attachments
Don't use air filters that must be replaced frequently; don't place these filters so that it is necessary to shut down power and disassemble the product to reach them
Don't omit the guide from a screwdriver adjustment, so that while the technician is adjusting and watching a meter, the screwdriver slips out of the slot
Don't design frequent failure or highly adverse impact failure items in least accessible locations
Don't unnecessarily subject nonwear components to failure stress caused by maintenance actions

The Design Team

During the entire period of design, reliability should be continuously monitored by the design team. The membership of the design team and the technical matters requiring their attention are shown in Table 11.6.4. A generic guiding list is given in Table 11.6.5.

TABLE 11.6.4 Design Team

Members	Technical Considerations
Project engineer	Products lists and application
Electrical, mechanical, chemical, manufacturing, and design representations	Tolerance studies
Management representatives	Environmental and operational effects
Maintenance and logistics representatives	Drift, aging, and end-of-life parameters
User community representatives	Regression or worst-case analysis
	Reliability analysis
	Trade-off studies
	Maintenance factors
	Test data
	Availability and affordability analysis

TABLE 11.6.5 Design Reliability Check List Item

Are the requirements for performance, application, environment, and maintainability and reliability established for each product, interface, and structure?
Are the best available methods for reducing the adverse effects of operational environments on critical structures being utilized?
Have normal modes of failure and the magnitude of each mode for each component or critical product have been identified as to root cause?
Has shelf life of products chosen for final design been determined?
Have limited-life products been identified and inspection and replacement requirements specified?
Have critical products that require special procurement, testing, and handling been identified?
Are effective safety factors being used in the application of products, interfaces, and structures?
Have studies been made considering variability and degradation of parameters of products and structures?
Have all vital adjustments been classified as to factory, preoperational, or operator types?
Have adjustments been minimized?
Have stability requirements of all products and structures associated with each adjustment been established?
Are similar plugs and connectors adequately protected against insertion in wrong socket?
Are malfunction-indicating products being used extensively?
Are self-monitoring or self-calibration products installed in major products where possible?
Are mechanical support structures adequate?
Is there a concentrated effort to make the developmental model as near to the production model as possible?
Have packaging and mechanical layout been designed to facilitate maintenance and to reduce maintenance costs and downtime?

Summary

When a product is being designed, it is assumed that the required investment will be justified according to how well the product performs its intended function over time. This assumption cannot be justified when a product fails to perform upon demand or fails to perform repeatedly. It is not enough to show that a product can conduct a function but that it can do so repeatedly when needed by the product user.

The design of any product involves trade-offs, including, but not limited to, performance capability, size, weight cost, product maintenance activities, and other factors, depending upon the intended use. In cases where human life is at risk, reliability issues play a major role in design. However, it is equally important to ask how large can a safety factor be for a critical situation such as a spacecraft, where human life might be in jeopardy, yet weight is a governing element.

The demand for better performance over time, coupled with higher-life cycle costs and stricter legal liabilities, has made product reliability of great importance. As the attitude toward production of engineering products has changed, reliability has established itself as one of the key elements when designing and manufacturing a product. However, the application of reliability principles in the product design and manufacturing processes has not always kept pace with the evolution of highly complex and dynamic engineering products. The design team must understand reliability theory and techniques, in conjunction

with design and manufacturing processes as well as the scenario for use of the product. The goal is to provide a scientific basis for design decisions considering product effectiveness requirements and the scenario for use, create a product design that will satisfy those requirements, and document those design decisions for both the hardware product and the recommended support product for sustaining product effectiveness during the life of the product.

11.7 Design of Manufacturing Systems and Processes

Leonard D. Albano

Design of Manufacturing Systems

Section 11.6 provides strategies for the design of products that meet the customer's needs, based on findings derived from market studies and product research and development. The success of such designs depends on the enterprise's ability to manufacture it, without compromising product performance or incurring noncompetitive production costs. Thus, the design of a manufacturing system must be done with a clear understanding of the functional requirements and physical design parameters for a product and an appreciation for the business goals and constraints.

The structure of a manufacturing system is hierarchical. At the highest level, there are two distinct subsystems: the *direct manufacturing operations* and the *indirect overhead functions*. The direct manufacturing operations consist of the employees and equipment needed to convert the input resource materials into the final desired products. The indirect overhead functions refer to the people, vendors, and capital that support and supervise the direct operations. At the lowest level, the manufacturing system is defined by many basic manufacturing operations, such as metal cutting and plastic forming, that are used to create specific part types (see Chapter 13).

Manufacturing systems can be classified in terms of the volume and variety of products produced. One possible classification scheme is as follows:

1. *Continuous flow processes*: high volume production of a single product type, such as sugar and oil refineries
2. *Assembly lines*: high volume production of a limited product variety; relies on standardized components and interchangeable parts; common system for automotive industry
3. *Batch processing*: relatively low volume production of multiple product types; concepts of group technology used to batch workpieces to increase efficiency of process routing; example applications include the manufacture of aircraft and heavy construction equipment
4. *Job shop*: low volume production of a variety of nonstandard parts; focuses on customized products; early reliance on craftsmen has shifted to cellular manufacturing systems in order to maximize flexibility and throughput

The design of a manufacturing system may be defined in terms of the concept of domains, as presented in Section 11.3, and the integration of three or more design fields; they are product design, organization design, and software design.

1. *Product design*: The design of the direct manufacturing operations is based on mapping the product's design parameters in the physical domain to process variables in the process domain.
2. *Organizational design*: The human and financial resources, which contribute to both direct operations and the indirect overhead, are provided to satisfy the goals and structure of the business organization.
3. *Software design*: Computer-based tools are easily available to control production and inventory levels, support information flow for decision-making, and automate manufacturing operations.

Several important advances in manufacturing can be identified from the above design model. *Concurrent engineering* refers to the integration of product design and manufacturing system design. This approach involves continuous mapping from the functional domain to the physical domain to the process domain. If the organizational structure changes to adapt to new customer needs, the single enterprise may be replaced by a highly flexible, virtual enterprise that pools resources from several qualified organizations for rapid response. This virtual enterprise corresponds to *agile manufacturing*. With a rationalized approach to system design, the role of the computers, information technology, robots, etc. can be better understood (see Chapter 13). Mitchell (1991) defines *computer-integrated manufacturing*

as "the use of computers to achieve an integrated flow of manufacturing activities, based on integrated information flow that links together all organizational activities."

A designer may develop many different manufacturing systems that are technically feasible. The designer must have some basis for the selection of a system and the layout of the various elements that comprise it. The evaluation of economic feasibility is one criterion for decision-making. Economic feasibility depends on the planned production volume, fixed resources, total production costs, projected revenue stream, and the desired income. The latter case involves consideration of the investment strategy. Because of the interrelations among the system performance and the economic factors, the suggested design procedure is to consider economic feasibility as a design constraint. The ideal to be achieved is one involving the design and production of reliable products that satisfy the customer's needs, while providing the desired profit and return on investment. See Chapter 13 for information on manufacturing system and enterprise management.

Manufacturing Process Design

The principal objective of manufacturing process design is to produce an organized plan for converting raw materials into useful products. It involves the selection of timely and cost-effective methods to produce a product without compromising quality and reliability. As part of the product development process, good manufacturing process design contributes to the industrial competitiveness of a manufacturing enterprise, while poor process design contributes to cost and schedule overruns and the delivery of products that fail to meet some or all of the customers' needs.

Manufacturing process planning often requires the consideration of several manufacturing processes for a specific part. In addition to considering manufacturing costs and production time, rational planning should also involve evaluation of how well a particular process satisfies the design requirements, which are delineated within the engineering drawings and reference specifications. The suitability of a given process may be based on many factors, such as

1. The dimensions and geometric precision that can be obtained
2. The surface roughness that can be attained
3. The changes that may be produced in material properties and part performance

Unfortunately, there is seldom much time to conduct an exhaustive laboratory or computer-based study and evaluation of all solution alternatives. A systematic approach to design that provides fundamental principles for decision-making would facilitate process design while enabling the designer to consider all important factors. Section 11.4, for instance, describes the concept of axiomatic design as a scientific framework for design and decision-making.

Metals Processing

Metal shapes and components can be produced by various casting, forming, and metal-removal processes (see Chapter 13). In the case of metal removal, process planning involves selecting and sequencing the appropriate machine tools and operations so as to convert a solid piece part from its initial shape to a final, desired geometry. This involves matching machine and tool data to the design requirements, subject to certain constraints imposed by the manufacturing organization and facilities. In practice, it consists of five or more steps:

1. Interpretation of the engineering drawing and reference specifications
2. Selection of machining operations to form the specified surfaces
3. Selection of machine tools for each machining operation
4. Selection of jigs and fixtures to guide or facilitate machining
5. Sequencing the machining operations

The initial material shape may be one of any number of geometries, ranging from simple bar stock to a complex casting. Basic machine tool operations are discussed in Chapter 13, and they include shaping, planing, turning, grinding, sawing, and drilling.

Process planning is performed by either experienced planners or automated software systems. There are at least two major difficulties associated with relying solely on human planners. First, they often rely on intuition gained through experience, and this experience requires a significant period of time to accumulate. Consequently, the number of truly skilled planners in any given industry is limited from generation to generation. Second, the capacity of a manufacturing enterprise to adopt new processes and new systems is limited by the knowledge background and creativity of the process planner. In order to address these problems, computer-aided process planning (CAPP) systems (see Chapter 13) have been developed.

The two fundamental strategies for CAPP are the *variant approach* and the *generative approach*. The variant approach is closely related to the automation of manual process planning techniques, i.e., the process plan for a specific part is created by computer-based retrieval of an existing process plan from a data base. Thus, the role of the computer in a variant system is primarily to manage a data base of process knowledge. Generative systems synthesize manufacturing and planning knowledge to generate specific process plans for a specific part. The reasoning mechanism for generative systems varies from algorithms to decision trees to artificial intelligence techniques. The variant approach is the basis for the overwhelming majority of computer codes for process planning because of a number of practical computational advantages. Nevertheless, several generative systems have been developed by industry, and this area will continue to be the focus of research and development, especially for large companies that can support the considerable investment needed for software development and hardware procurement.

Polymer Processing

There are basically two major categories of plastics: *thermoplastics* and *thermosetting resins* (see Chapter 12). Each group contains thousands of specific formulations with a wide range of mechanical, physical, and chemical properties. The scope of applications includes the automotive, houseware, and packaging industries. For example, government regulations pertaining to energy conservation have forced the automotive manufacturers to reduce the weight of automobiles. As a result, lightweight plastics have replaced traditional steel bumpers and outer panels while fulfilling specific performance requirements for mechanical strength, thermal resistance, and environmental durability.

To design with plastics, the product engineer selects an appropriate plastic material by considering a number of technical and economic factors in concert with specific data compiled in manufacturer's literature and encyclopedic data sources, such as the *Modern Plastics Encyclopedia*. To select the appropriate manufacturing process, the process engineer must understand the impact of the processing techniques on the material structure and, ultimately, the desired technical performance, e.g., mechanical strength and physical properties. Although a considerable number of plastic material alternatives are available, the variety of forming operations is limited. The basic set of processes includes compression molding, transfer molding, injection molding, extrusion, cold molding, thermoforming, and blow molding. The reader is referred to Chapter 13 for specific information on the production of thermoplastics and thermosetting resins.

Because of the interrelations between processing techniques and material structure, the axiomatic approach to design (see Section 11.4) provides helpful information for concurrent engineering. In this context, the use of plastic parts involves the successive mapping of engineering properties in the functional domain, to a plastic material structure in the physical domain, to processing techniques in the process domain. This mapping can be written in the form:

$$\{\mathbf{FRs}\} = [\mathbf{A}]\{\mathbf{DPs}\} \tag{11.7.1}$$

$$\{\mathbf{DPs}\} = [\mathbf{B}]\{\mathbf{PVs}\} \tag{11.7.2}$$

in which the desired material properties are contained in the vector of functional requirements {**FRs**}, the material structure is described in {**DPs**}, and the manufacturing process is defined by {**PVs**}. Equations (11.7.1) and (11.7.2) display the general structure of the equations for materials design and process planning. The design matrices [**A**] and [**B**] are used to determine whether the proposed mapping satisfies the Independence Axiom: they must be either diagonal or triangular to allow for the manufacture of a plastic part with the desired properties (see Section 11.4.1, "The First Axiom: The Independence Axiom"). The production of microcellular plastics of MIT (Suh, 1990) is an example that illustrates design of a manufacturing process to yield the desired polymer structure.

References

Mitchell, F.H. 1991. *CIM Systems: An Introduction to Computer-Integrated Manufacturing.* Prentice-Hall, Englewood Cliffs, NJ.

Modern Plastics Enclycopedia. McGraw-Hill, New York.

Suh, N.P. 1990. *The Principles of Design.* Oxford University Press, New York.

11.8 Precision Machine Design

Alexander Slocum

Any machine, from a machine tool to a photocopier to a camera, is an assembly of components that are designed to work together to achieve a desired level of performance. Each machine has a budget for cost and performance, and achieving the best balance between the two, regardless of the function of the machine, is the essence of *precision machine design.*

In order to be able to effectively develop a design for a precision machine, the design engineer must simultaneously envision in his/her head the functions the machine must perform (e.g., milling, turning, or grinding) alongside a pictorial library of component technologies (e.g., bearings, actuators, and sensors), generic machine configurations (e.g., cast or welded articulated and/or prismatic structures), analysis techniques (e.g., back-of-the-envelope and finite element methods), and manufacturing methods (e.g., machine, hand, or replication finished). In addition, the machine design engineer must be aware of the basic issues faced by the sensor and electronics engineer, the manufacturing engineer, the analyst, and the controls engineer. Only by a simultaneous consideration of all design factors can a viable and effective design be rapidly converged upon. Awareness of current technological limitations *in all fields* can also help a design engineer to develop new processes, machines, and/or components.

The goal of the precision machine design engineer is to make all the components of a precision machine in proper proportion of each other, both in relation to their physical size and the capabilities of the servocontroller and power systems. If a component is oversized, it may increase the cost of the machine while performance may not be increased. If a component is too small, the rest of the machine's components may never reach their potential and machine performance will suffer. Note that size is a function of static and dynamic qualifiers. Components that behave well statically do not necessarily have good dynamic performance.

In today's world where rapid time to market is essential, the design of quality precision machines depends on the ability of the design and the manufacturing engineers to predict how the machine will perform before it is built. Kinematics of a machine are easily tested for gross functionality using mechanism synthesis and analysis software. Wear rates, fatigue, and corrosion are often difficult to predict and control, but for the most part are understood problems in the context of machine tool design. Hence perhaps the most important factors affecting the quality of a machine are the accuracy, repeatability, and resolution of its components and the manner in which they are combined. These factors are critical because they affect every one of the parts that will be manufactured using the machine. Accordingly, minimizing machine cost and maximizing machine quality mandate predictability of accuracy, repeatability, and resolution. As noted by Donaldson*: "A basic finding from our experience in dealing with machining accuracy is that machine tools are deterministic. By this we mean that machine tool errors obey cause-and-effect relationships, and do not vary randomly for no reason."

In this section, the design of precision machines will be studied by following a path of analysis of the overall system's potential for accuracy, as well as applying this perspective to the understanding of the operation of the components that make up a precision machine:

- Analysis of errors in a precision machine
- Structures
- Bearings

It is important to realize that the design of precision machines is like any other endeavor. The better one understands the fundamental principles and concepts, the better one will be able to integrate components into a precision system. In the light of understanding the science of design, one must always consider practical applied philosophy. A few broad design guidelines for the precision machine designer include:

* Donaldson, R. November 1972. The deterministic approach to machining accuracy. *SME Fabricat. Technol. Symp.* Golden, CO (UCRL Preprint 74243).

- Subject all decisions to an "is there a better way" value analysis based on system considerations.
- Always picture in your mind how the system will be manufactured, assembled, used, and maintained.
- Minimize the number of parts in an assembly and minimize their complexity.
- Maximize the number of instances where reference surfaces and self-locating "snap together" parts can be used.
- Whenever possible, take advantage of kinematic design principles.
- Utilize new materials and technologies to their fullest potential.
- Read continuously and familiarize yourself with technologies in many areas.
- Observe continuously and familiarize yourself with products in many areas.

Analysis of Errors in a Precision Machine

The first step in the design of a precision machine is to understand the basic definitions and principles that characterize and govern the design of all precision machines. Some of the more basic definitions and principles include

- The principle of reversal
- Modeling the errors in a machine
- Determination of the relative errors between the tool and the workpiece
- Linear motion system errors
- Estimating position errors from modular bearing catalog data
- Axis of rotation errors
- Error budgets

Accuracy, Repeatability, and Resolution

There are three basic definitions to remember with respect to how well a machine tool can position its axes: *accuracy, repeatability (precision),* and *resolution. Accuracy* is the maximum translational or rotational error between any two points in the machine's work volume. *Repeatability (precision)* is the error between a number of successive attempts to move the machine to the same position, or the ability of the machine to make the same motion over and over. Bidirectional repeatability is the repeatability achieved when the point is approached from two different directions. This includes the effect of backlash in a leadscrew. Accuracy is often defined in terms of the mean, and repeatability in terms of the standard deviation. For a set of N data points with a normal (Gaussian) distribution, the mean x_{mean} and the standard deviation σ are defined as

$$x_{\text{mean}} = \frac{1}{N}\sum_{i=1}^{N} x_i \qquad \sigma = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}\left(x_i - x_{\text{mean}}\right)^2} \tag{11.8.1}$$

The standard deviation is used in the determination of the probability of occurrence of an event in a system that has a normal distribution. Table 11.8.1 gives the percent chance of a value occurring within a number of standard deviations of its expected value.* One must be very careful not to confuse an offset of the mean with the allowed random variation in a part dimension. For example, precision-molded plastic lenses for cameras have their molds hand finished to make the mean size of lenses produced equal to the nominal size required.

* The equation for the computation of the percent chance was obtained from Drake, A. 1967. *Fundamentals of Applied Probability Theory.* McGraw-Hill, New York. p. 211. Values ϕ (k) were computed using Mathematics™. Another valuable reference to have is Natrella, M. *Experimental Statistics,* NBS Handbook 91.

TABLE 11.8.1 Chance of a Value Falling within kσ of its Expected Value

k	% chance of occurrence
1.0	68.2689
2.0	95.4500
3.0	99.7300
4.0	99.9937
5.0	99.9999
6.0	100.0000

It is interesting to note, however, Bryan's* observation of the issue of using probabilistic methods to characterize repeatability: "The probabilistic approach to a problem is only a tool to allow us to deal with variables that are too numerous, or expensive to sort out properly by common sense and good metrology." One must not belittle probability, however, for it is a mathematical tool like any other that is available to the design engineer. Required use of this tool, however, might be an indication that it is time to take a closer look at the system and see if the system can be changed to make it deterministic and therefore more controllable. Often the key to repeatability is not within the machine itself, but in isolating the machine from variations in the environment.**

Resolution is the larger of the smallest programmable step or the smallest mechanical step the machine can make during point-to-point motion. Resolution is important because it gives a lower bound on the repeatability. When a machine's repeatable error is mapped, the resolution becomes important if the mapped errors are to be compensated for by other axes.

Amplification of Angular Errors

Perhaps the most overlooked error in machine design is the amplification of an angular error over a distance to create a large translational error at some point in the machine. Mathematically, this error has a magnitude equal to the product of the distance between the point and the rotation source and the sine of the included angle. Hence this type of error can be referred to as a *sine error.* This principle extends to locating bearing surfaces far from the workpiece area of the machine tool.*** Errors in the bearing's motion can be amplified by the distance between the bearing and the workpiece, and can be transmitted to the workpiece. This can result in horizontal and vertical straightness errors and axial position errors. The same is true for the effects of all other types of errors on machine components. A *cosine error*, on the other hand, is the error made when the measurement of the distance between a point and a line is not made along a path orthogonal to the line.

With respect to dimensional measurements. In the late 1800s, Dr. Ernst Abbe noted *"If errors in parallax are to be avoided, the measuring system must be placed coaxially with the axis along which displacement is to be measured on the workpiece."* The Abbe principle can be visualized by comparing measurements made with a dial caliper and a micrometer as shown in Figure 11.8.1. The dial caliper is often used around the shop because it is easy to use, the head slides back and forth to facilitate quick measurement. However, note that the measurement scale is located at the base of the jaws. When a part is measured near the tip of the jaws, the jaws can rock back slightly, owing to their elasticity and imperfections in the sliding jaw's bearing. This causes the caliper to yield a slightly undersize

* Bryan, J. The Power of Deterministic Thinking in Machine Tool Accuracy, 1st Int. Mach. Tool Eng. Conf. November 1984. Tokyo (UCRL Preprint 91531).

** *"In designing an experiment the agents and phenomena to be studied are marked off from all others and regarded as the field of investigation. All others are called disturbing agents. The experiment must be so arranged that the effects of disturbing agents on the phenomena to be investigated are as small as possible."* James C. Maxwell.

*** See Bryan, J.B. 1989. The Abbe principle revisited — an updated interpretation, *Precis. Eng.* 1(3):129–132. An extension of the Abbe principle to this type of situation is referred to as the *Bryan principle*, which is discussed in Section 5.2.1 along with a discussion on where to mount sensors for various types of measurements.

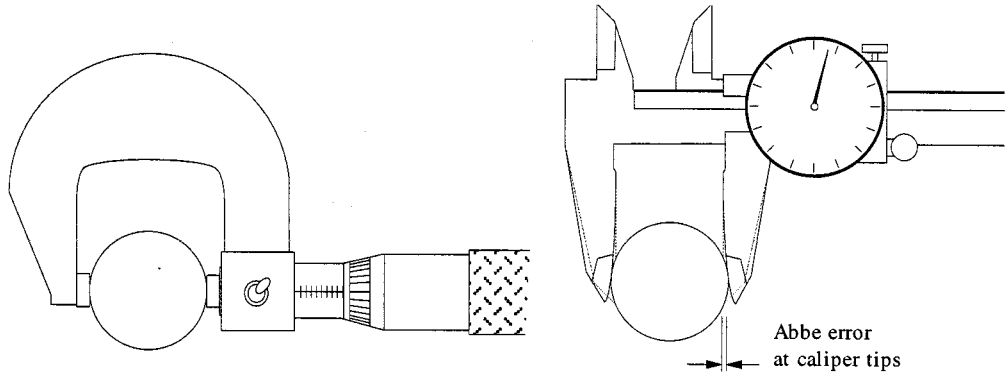


FIGURE 11.8.1 Abbe error illustrated through the use of a dial caliper and a micrometer.

measurement. The micrometer, on the other hand, uses a precision measuring device located in line with the part dimension, so there is no Abbe error. Both instruments are sensitive to how hard the jaws are closed on the part. A micrometer often has a torque limiting adjustment to provide a very repeatable measuring force. It is impossible to overstress the importance of Abbe errors.

The micrometer is more difficult to use and less versatile than the caliper, yet it has greater accuracy because of its more robust structural loop. So it is in the design of precision machines. Often one has to sacrifice ergonomics for performance. The goal, therefore, is to minimize the extent of the sacrifice.

Sensitive Directions

It is important to note that there are *sensitive directions* in a machine, and it is along these directions that the most effort must be expended to minimize errors. For example, as illustrated in Figure 11.8.2 an error motion ε of a tool tangent to the surface of a round part of radius r in a lathe results in a radial error in the workpiece of magnitude ε^2/r, which is much smaller than ε. Sensitive directions can be *fixed*, such as when the tool is stationary and the part is moving (e.g., a lathe), or *rotating*, such as when the tool is rotating and the part is fixed (e.g., a jig borer).

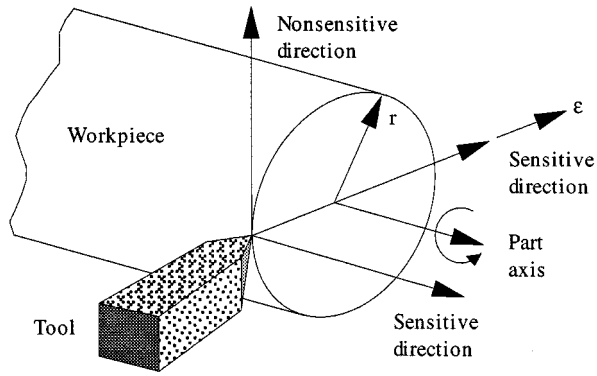


FIGURE 11.8.2 Illustration of sensitive directions.

The Reversal Principle

How were the first accurate machines developed? Machines were first made repeatable, which can be done by paying special detail to surface finish of bearings, prevention of lost motion (e.g., backlash), and by making the forces on the machine repeatable. Once a machine is made repeatable, it can be used to measure the accuracy of another component or machine using the principle of reversal. This is illustrated in Figure 11.8.3.

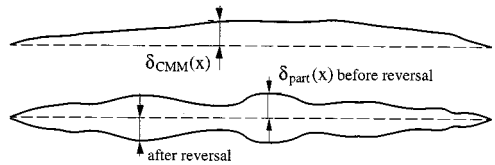


FIGURE 11.8.3 Illustration of the principal of reversal.

The machine is used to measure a part from one side, obtaining a reading of

$$Z_{\text{probe before reversal}}(x) = \delta_{CMM}(x) - \delta_{\text{part}}(x)$$

The part is then turned over, taking care to keep its axial position justified, then the machine is used to remeasure the part, obtaining a reading of

$$Z_{\text{probe after reversal}}(x) = \delta_{CMM}(x) - \delta_{\text{part}}(x)$$

By subtracting the measurements from each other, the repeatable error in the measuring machine is removed:

$$\delta_{\text{part}}(x) = \frac{-Z_{\text{probe before reversal}}(x) + Z_{\text{probe after reversal}}(x)}{2}$$

There are many variations (e.g., for roundness measurements), and the principle shows how repeatability is often the most important characteristic of a precision machine. In fact, the principle can be applied to the design of the machine itself. For example, two bearing rails can be ground side by side, and then placed end to end on a machine. If the carriage that rides on them has the same bearing spacing as the individual rail length, then as the carriage moves on the end-to-end rails, it will not pitch or yaw, it will only have a straightness error. This straightness error can then be more readily compensated for by the motion of an orthogonal axis via an error compensation algorithm in the controller.

Modeling the Errors in a Machine

In order to determine the effects of a component's error on the position of the toolpoint or the workpiece, the spatial relationship between the two must be defined. Figure 11.8.4 illustrates the issue: the tool is connected to the workpiece through the linkage that is the machine. Any errors between the machine's links create error between the tool and the workpiece. The essential goal of precision machine design is to be able to model and predict these errors, so they can be addressed during the design state of the machine.

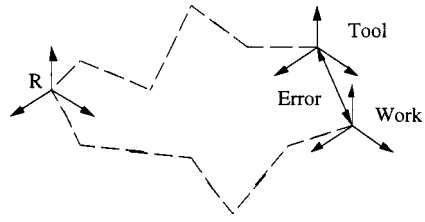


FIGURE 11.8.4 The goal is to be able to model the errors in the machine to predict the error between the tool and the work during the design phase so appropriate action can be taken before the machine is built.

To represent the relative position of a rigid body in three-dimensional space with respect to a given coordinate system, a 4×4 matrix is needed. This matrix represents the coordinate transformation to the reference coordinate system (X_R Y_R Z_R) from that of the rigid body frame (X_n Y_n Z_n), and it is called the *homogeneous transformation matrix* (HTM).* The first three columns of the HTM are direction cosines (unit vectors i, j, k) representing the orientation of the rigid body's X_n, Y_n, and Z_n axes with respect to the reference coordinate frame, and their scale factors are zero. The last column represents the position of the rigid body's coordinate system's origin with respect to the reference coordinate frame. P_s is a scale factor, which is usually set to unity to help avoid confusion. The presuperscript represents the reference frame you want the result to be represented in, and the postsubscript represents the reference frame from which you are transferring:

$$R_{\mathbf{T}_n} = \begin{bmatrix} O_{ix} & O_{iy} & O_{iz} & P_x \\ O_{jx} & O_{jy} & O_{jz} & P_y \\ O_{kx} & O_{ky} & O_{kz} & P_z \\ 0 & 0 & 0 & P_s \end{bmatrix} \tag{11.8.2}$$

Thus, the equivalent coordinates of a point in a coordinate frame n, with respect to a reference frame R, are

$$\begin{bmatrix} X_R \\ Y_R \\ Z_R \\ 1 \end{bmatrix} = R_{\mathbf{T}_n} \begin{bmatrix} X_n \\ Y_n \\ Z_n \\ 1 \end{bmatrix} \tag{11.8.3}$$

For example, if the X_1 Y_1 Z_1 coordinate system is translated by an amount x along the X axis, the HTM that transforms the coordinates of a point in the X_1 Y_1 Z_1 coordinate frame into the XYZ reference frame is

$$XYZ_{\mathbf{T}_{X_1Y_1Z_1}} = \begin{bmatrix} 1 & 0 & 0 & x \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{11.8.4}$$

If the X_1 Y_1 Z_1 coordinate system is rotated by an amount θ_x about the X axis, the HTM that transforms the coordinates of a point in the X_1 Y_1 Z_1 coordinate frame into the XYZ frame is

$$XYZ_{\mathbf{T}_{X_1Y_1Z_1}} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & \cos\theta_x & -\sin\theta_x & 0 \\ 0 & \sin\theta_x & \cos\theta_x & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{11.8.5}$$

Machine structures can be decomposed into a series of coordinate transformation matrices describing the relative position of each axis and any intermediate coordinate frames that may assist in the modeling

* The HTM representation of structures has existed for many decades. See, for example, Denavit, J. and Hartenberg, R. June 1955. A kinematic notation for lower-pair mechanisms based on matrices. *J. Appl. Mech.* Perhaps the most often referenced work with respect to its application to manufacturing tools is Paul, R. 1981. *Robot Manipulators: Mathematics, Programming, and Control.* MIT Press, Cambridge, MA.

process, starting at the tip and working all the way down to the base reference coordinate system (n = 0). If N rigid bodies are connected in series and the relative HTMs between connecting axes are known, the position of the tip (Nth axis) in terms of the reference coordinate system will be the sequential product of all the HTMs:

$$ {}^{R}\mathbf{T}_{N} = \prod_{m=1}^{N} {}^{m-1}\mathbf{T}_{m} = {}^{0}\mathbf{T}_{1}\,{}^{1}\mathbf{T}_{2}\,{}^{2}\mathbf{T}_{3}\ldots \tag{11.8.6} $$

It can be difficult to determine how a part modeled as a rigid body actually moves; thus care must be taken when evaluating the error terms in the HTMs of systems with multiple contact points. Nonserial link machines (e.g., a four-bar linkage robot) require a customized formulation to account for interaction of the links.

Determination of the Relative Errors Between the Tool and the Workpiece

Ideally, the HTM products (Equation 11.8.6) for the position of the point on the workpiece the tool contacts and the location of the toolpoint with respect to the reference frame will be identical. The relative error HTM $\mathbf{E}_{\text{rel}}$, representing position and orientation errors between the tool and workpiece, is determined for ${}^{R}\mathbf{T}_{\text{work}} = {}^{R}\mathbf{T}_{\text{tool}}\mathbf{E}_{\text{rel}}$:

$$ E_{\text{rel}} = {}^{R}\mathbf{T}_{\text{tool}}^{\;-1}\,{}^{R}\mathbf{T}_{\text{work}} \tag{11.8.7} $$

The relative error HTM is the transformation in the toolpoint coordinate system that must be done to the tool in order to be at the proper position on the workpiece. For implementation of error correction algorithms on numerically controlled machines, as illustrated in Figure 11.8.5 one must consider how the axes will be required to move in order to create the desired motion, specified by Equation 11.8.7 in the tool reference frame. For the general case of a machine with revolute and translational axes (e.g., a five-axis machining center), one would have to use inverse kinematic solutions such as those developed for robot motion path planning. Most machine tools and CMMs have only translational axes, and thus the error correction vector ${}^{R}\mathbf{P}_{\text{correction}}$ with respect to the reference coordinate frame can be obtained from

$$ {}^{R}\begin{bmatrix} P_x \\ P_y \\ P_z \end{bmatrix}_{\text{correction}} = {}^{R}\begin{bmatrix} P_x \\ P_y \\ P_z \end{bmatrix}_{\text{work}} - {}^{R}\begin{bmatrix} P_x \\ P_y \\ P_z \end{bmatrix}_{\text{tool}} \tag{11.8.8} $$

Because of Abbe offsets and angular orientation errors of the axes, ${}^{R}\mathbf{P}_{\text{correction}}$ will not necessarily be equal to the position vector $\mathbf{P}$ component of $\mathbf{E}_{\text{rel}}$. ${}^{R}\mathbf{P}_{\text{correction}}$ does represent the incremental motions the X, Y, and Z axes must make on a Cartesian machine in order to compensate for toolpoint location errors.

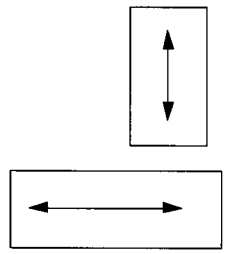


FIGURE 11.8.5 Repeatable error motions in one axis can be mapped and compensated for by motion of an orthogonal axis.

Compensating for errors in a machine can be a good thing; however, electronic compensation can only do so much. One has to start with a good design. The error in a machine can be minimized in general by maximizing the efficiency of the machine's *structural loop*. The structural loop is defined as the structure that joins the tool to the fixture to which the workpiece is attached. During cutting operations, the contact between the tool and workpiece can change the structural loop's characteristics. Maximizing the efficiency of the structural loop generally requires minimizing the path length of the mechanism. As can be seen from Abbe's principle or the HTM analysis method, the shorter the path length, the less the error amplification and the total end point error.

Linear Motion System Errors

Consider the case of an ideal linear motion carriage shown in Figure 11.8.6 with x, y, and z offsets of a, b, and c, respectively, All rigid bodies have three rotational (ε_X, ε_Y, ε_Z) and three translational (δ_X, δ_Y, δ_Z) error components associated with their motion. These errors can be defined as occurring about and along the reference coordinate system's axes, respectively. Often the errors will be a function of the position of the body in the reference frame.

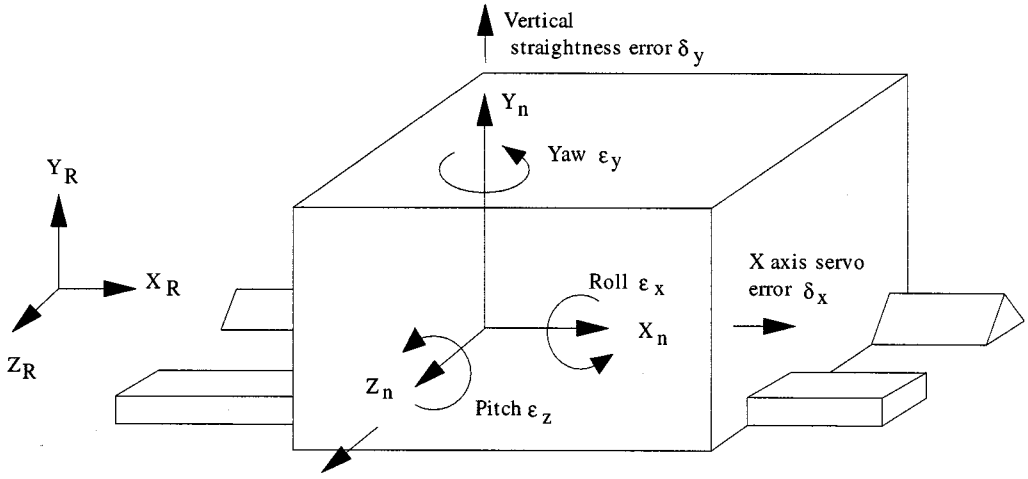


FIGURE 11.8.6 Motion and errors in a single-axis linear motion carriage (prismatic joint).

For the linear carriage, the HTM (see Equation (11.8.2)) that describes the effects of errors on carriage motion can be obtained in the following manner. Observing the right-hand rule, a rotation ε_X, about the X axis (*roll*) causes the tip of the Y-axis vector to move in the positive Z direction by an amount proportional to sin ε_X, and in the negative Y direction by an amount proportional to $1 - \cos \varepsilon_X$. Since the error terms are very small, at most on the order of minutes of arc, small-angle approximations are valid and will be used. Hence the element $o_{ky} = \varepsilon_X$. The roll error ε_X also causes the tip of the Z-axis vector to move in the negative Y direction, so that the element $o_{jz} = -\varepsilon_X$. A rotation ε_Y about the Y axis (*yaw*) causes the tip of the X-axis vector to move in the negative Z direction, so that $o_{kx} = -\varepsilon_Y$, and causes the tip of the Z-axis vector to move in the positive X direction, so $o_{iz} = \varepsilon_Y$. Similarly, a rotation ε_Z about the Z axis (*pitch*) causes the tip of the X-axis vector to move in the positive Y direction, so $o_{jx} = \varepsilon_Z$, and causes the tip of the Y axis vector to move in the negative X direction so that $o_{iy} = -\varepsilon_Z$. Translational errors δ_X, δ_Y, and δ_Z directly affect their respective axes, but care must be taken in defining them. Having neglected second-order terms, the HTM for the linear motion carriage with errors is

$$ {}^{R}\mathbf{T}_{\text{nerr}} = \begin{bmatrix} 1 & -\varepsilon_Z & \varepsilon_Y & a+\delta_X \\ \varepsilon_Z & 1 & -\varepsilon_X & b+\delta_Y \\ -\varepsilon_Y & \varepsilon_X & 1 & c+\delta_Z \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{11.8.9} $$

Estimating Position Errors from Modular Bearing Catalog Data

The HTM method is powerful, but from where does one get estimates of the errors? For the linear motion system, the HTM assumes that the errors occur at the center of stiffness of the carriage. The *center of stiffness* is the point at which, when a force is applied to the system, no net angular motion results. For a symmetrical system, it is located at the center of the system. For other systems, the center of stiffness can be found in the same way that the *center of mass* is found:

$$ x_{\text{center of stiffness}} = \frac{\sum_{i=1}^{N} K_i x_i}{\sum_{i=1}^{N} K_i} \tag{11.8.10} $$

Indeed, the *center of friction* can also be defined, and when locating the point, where the actuation force of an axis is to be applied, one can see that ideally, the center of mass, the center of stiffness, and the center of friction should all be coincident. In the best case, they will also be concurrent with external loads applied to an axis.

Once the center of stiffness has been found, the error motions about the center of stiffness can be determined using the dimensions of the system and estimates for the errors in individual bearing elements (e.g., linear ball bearing catalogs often include data on the vertical and horizontal error motions that will occur between the bearing block and the rail). Figure 11.8.7 illustrates the model of a linear motion system supported by modular bearings.

It can be assumed that the translational errors are based on the average of the straightness errors experienced by the bearing blocks:

$$ \begin{aligned} \delta_x &= \delta_{\text{servo}} \\ \delta_y &= \frac{\delta_{y1}+\delta_{y2}+\delta_{y3}+\delta_{y4}}{4} \\ \delta_z &= \frac{\delta_{z1}+\delta_{z2}+\delta_{z3}+\delta_{z4}}{4} \end{aligned} \tag{11.8.11} $$

The angular errors are based on the differences in the average straightness errors experienced by pairs of bearing blocks acting across the carriage:

$$ \begin{aligned} \varepsilon_x &= \frac{\frac{\left(\delta_{y2}+\delta_{y3}\right)}{2} - \frac{\left(\delta_{y1}+\delta_{y4}\right)}{2}}{W} \\ \varepsilon_y &= \frac{\frac{\left(\delta_{z3}+\delta_{z4}\right)}{2} - \frac{\left(\delta_{z1}+\delta_{z2}\right)}{2}}{L} \\ \varepsilon_z &= \frac{\frac{\left(\delta_{y1}+\delta_{y2}\right)}{2} - \frac{\left(\delta_{y3}+\delta_{y4}\right)}{2}}{L} \end{aligned} \tag{11.8.12} $$

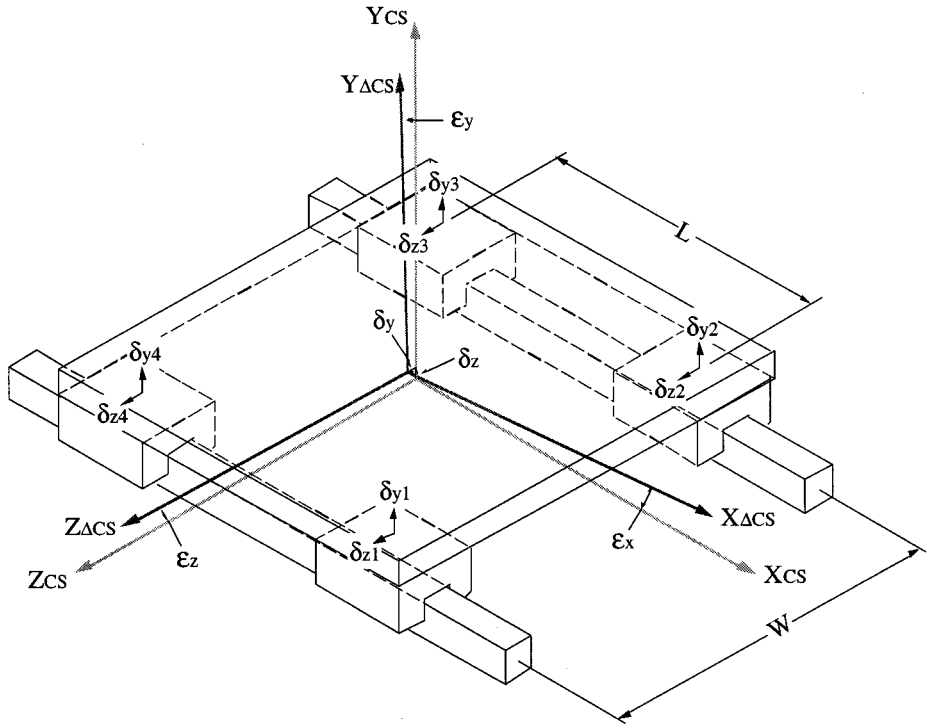


FIGURE 11.8.7 A linear motion system will have errors that act about the center of stiffness of the system.

When the bearing blocks are moving in unison to create the straightness error, they would not be creating the angular errors. However, to be conservative when modeling the errors in a machine, since it cannot be known which errors are going to occur, one should simultaneously incorporate both translational and angular error estimates into the error model of the machine.

Once a machine has been built, the error motions would be measured and analyzed to verify whether the machine meets its performance criteria. This is also a good time to evaluate the effectiveness of the model. For example, Figure 11.8.8 shows the straightness measurements for a linear motion axis supported by cam followers running on a vee and a flat. It is very difficult to tell what the source of the error is. Figure 11.8.9 shows a Fast Fourier Transform of the straightness data. Note that the FFT is plotted in terms of amplitude and wavelength (as opposed to the more commonly seen power and frequency plots used by electrical engineers). It can help identify the dominant sources of error, so design attention can be properly allocated. It can also show, as in this case, that the errors from a number of different sources are of similar magnitude, so if greater performance is required, one should probably seek an alternate design.

Axis of Rotation Errors*

Consider the rotating body shown in Figure 11.8.10. Ideally, the body rotates about its axis of rotation without any errors; however, in reality the axis of rotation revolves around an axis of the reference coordinate frame with radial errors δ_X and δ_Y, an axial error δ_Z, and tilt errors ε_X and ε_Y. All of these errors may be a function of the rotation angle θ_z. For a point in the spindle coordinate frame $X_nY_nZ_n$,

* Definitions used in this section were condensed from those provided in *Axis of Rotation: Methods for Specifying and Testing.* ANSI Standard B89.3.4M-1985, American Society of Mechanical Engineers, United Engineering Center, 345 East 47th Street, New York, NY 10017. This document also contains appendices which describe measurement techniques and other useful topics pertaining to axes of rotation.

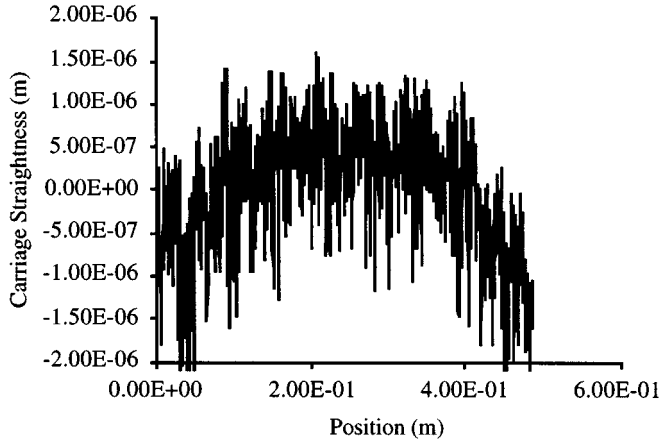


FIGURE 11.8.8 Strightness of a linear motion system.

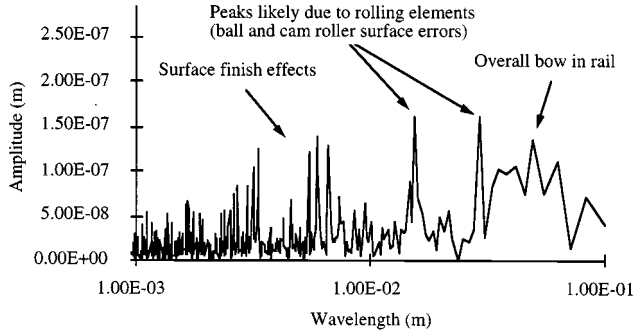


FIGURE 11.8.9 Fourier transform of the straightness of a linear motion system.

one would first use the rotation angle θ_z to transform the point roughly the reference frame. Then, since the other error motions are small, the order of multiplication of their HTMs would not be critical.

With the operators S = sine and C = cosine, the general result is

$$ {}^{R}\mathbf{T}_{\text{nerr}} = \begin{bmatrix} C\varepsilon_Y C\theta_Z & -C\varepsilon_Y S\theta_Z & S\varepsilon_Y & \delta_X \\ S\varepsilon_X S\varepsilon_Y C\theta_Z & C\varepsilon_X C\theta_Z - S\varepsilon_X S\varepsilon_Y S\theta_Z & -S\varepsilon_X C\varepsilon_Y & \delta_Y \\ -C\varepsilon_X S\varepsilon_Y C\theta_Z + S\varepsilon_X S\theta_Z & S\varepsilon_X C\theta_Z + C\varepsilon_X S\varepsilon_Y S\theta_Z & C\varepsilon_X C\varepsilon_Y & \delta_Z \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{11.8.13} $$

Note that this general result may also be used for the case of a linear motion carriage if ε_Z is substituted for θ_Z. Most often, second-order terms such as $\varepsilon_X\varepsilon_Y$ are negligible and small-angle approximations (i.e., cos ε = 1, sin ε = ε) can be used, which leads to

$$ {}^{R}\mathbf{T}_{\text{nerr}} = \begin{bmatrix} \cos\theta_Z & -\sin\theta_Z & \varepsilon_Y & \delta_X \\ \sin\theta_Z & \cos\theta_Z & -\varepsilon_X & \delta_Y \\ \varepsilon_X \sin\theta_Z - \varepsilon_Y \cos\theta_Z & \varepsilon_X \cos\theta_Z + \varepsilon_Y \sin\theta_Z & 1 & \delta_Z \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{11.8.14} $$

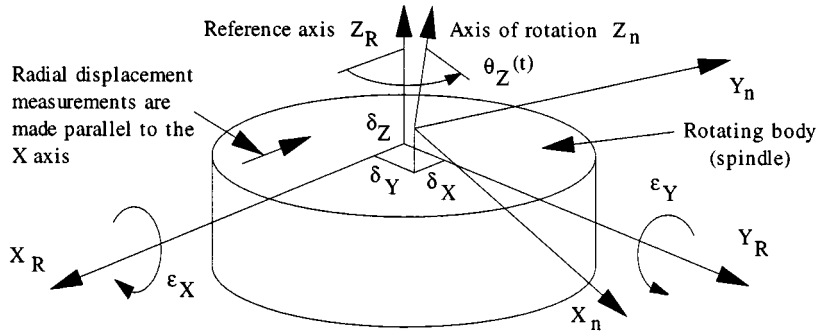


FIGURE 11.8.10 Motion and errors about an axis of rotation (revolute joint).

However, since this matrix is usually evaluated using a spreadsheet, it is best to use the exact form. When nanometer performance levels are sought, second-order effects can start to become important.

In the context of evaluating errors of rotating bodies, particularly when making measurements and discussing the results, it is necessary to consult the axis of rotation standard ANSI B89.3.4M. The "*error motion* of a spindle is the change in position, relative to the reference coordinate axes, of the surface of a perfect workpiece with its center line coincident with the axis of rotation." This definition excludes thermal drift errors. The *runout* is the total displacement measured by an instrument sensing against a moving surface or moved with respect to a fixed surface. The term total indicator reading (TIR) is equivalent to runout. Unfortunately, all too often an imperfect workpiece is eccentrically mounted to a spindle and used to evaluate the performance of a spindle; hence the runout can be a misleading measurement. The "*axial motion* is the error motion colinear with the Z reference axis." "Axial slip", "end camming", and "drunkenness" are nonpreferred terms which have been used in the past. The *tilt motion* is the error motion in an angular direction relative to the Z reference axis. Tilt motion creates sine errors on the spindle which is why the radial error motion is a function of Z position and face motion is a function of radius. Note that "coning", "wobble", and "swash" are sometimes used to describe tilt motion, but they are nonpreferred terms.

These errors are measured and then plotted using polar plots. The *error motion polar plot* is a polar plot of error motion made in synchronization with the rotation of the spindle. Error motion polar plots are often decomposed into plots of various error components. Some of the various types of error motion polar plots are shown in Figure 11.8.11. The *total error motion polar plot* is the complete error motion polar plot as recorded. The *average error motion polar plot* is the mean contour of the total error motion polar plot averaged over the number of revolutions. Note that asynchronous error motion components do not always average out to zero, so the average error motion polar plot may still contain asynchronous components. The average error motion value is a measure of the best roundness that can be obtained for a part machined while being held in the spindle (or the roundness of a hole the spindle is used to bore). The *fundamental error motion polar plot* is the best-fit reference circle fitted to the average error motion polar plots. The *residual error motion polar plot* is the deviation of the average error motion polar plot from the fundamental error motion polar plot. For radial error motion measurements, this represents the sum of the error motion and the workpiece (e.g., ball) out-of-roundness. The workpiece out-of-roundness can be removed using a reversal technique.* The *asynchronous error motion polar plot* is the deviation of the total error motion polar plot from the average error motion polar plot. Asynchronous in this context means that the deviations are not repetitive from revolution to revolution. Asynchronous

* The former was developed by Bob Donaldson at LLNL, and the latter by Spragg and Whitehouse. See Appendix B of ANSI B89.3.4M-1985 and the various cited references. Also see the ANSI standard *Measurement of Out-of-Roundness*, ANSI B89.3.1-1972(R1979).

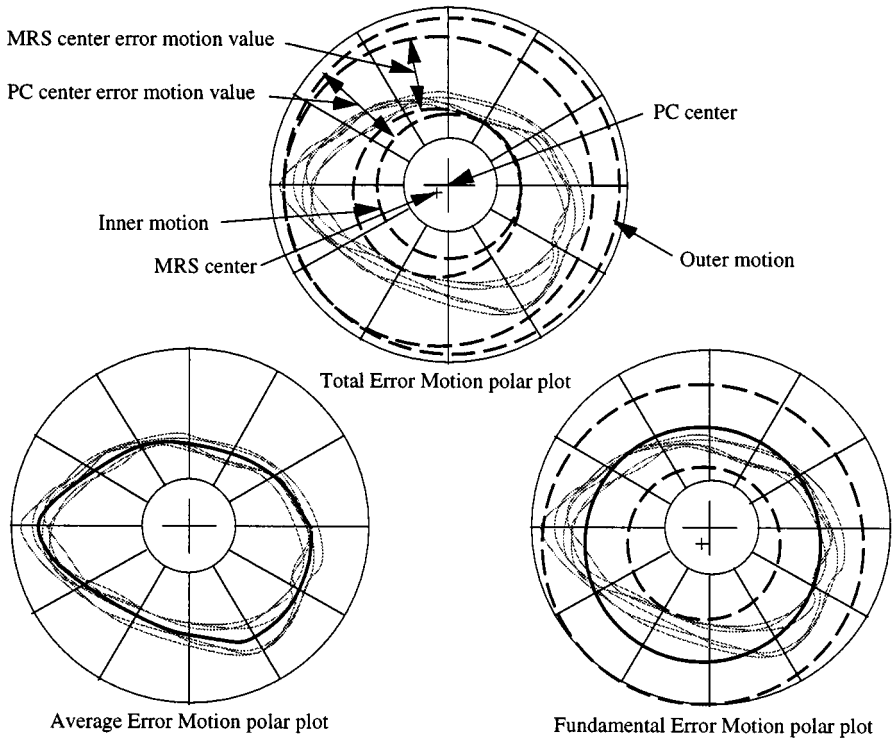


FIGURE 11.8.11 Examples of error motion and error motion component polar plots.

error motions are not necessarily random (in the statistical sense), but they do provide an indication of the attainable surface finish of the part.

Once again, the FFT is a vital tool for identifying the source of errors so that the designer can seek to minimize them. Figure 11.8.12 shows an FFT of a grinding spindle's error motion. The spindle speed was 1680 rpm (28 Hz), the bearing inner diameter was 75 mm, the outer diameter was 105 mm, the number of balls was 20, and the ball diameter was 10 mm. One of the dominant errors occurs at twice the rotation frequency, indicating that the bore in which the bearing is placed is most likely out-of-round.

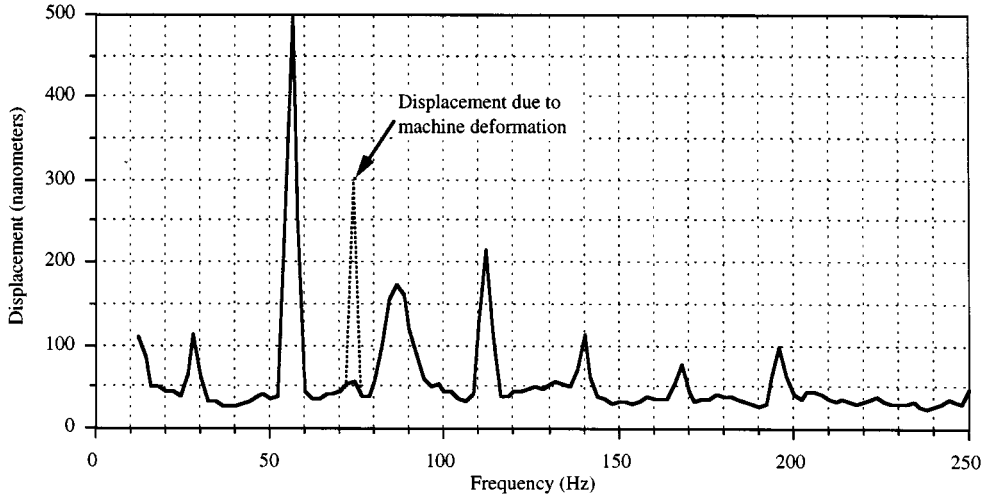


FIGURE 11.8.12 FFT of grinding spindle error motion.

Error Budgets

To define the relative position of one rigid body with respect to another, six degrees of freedom must be specified. To further complicate matters, error in each of the six degrees of freedom can have numerous contributing components. In fact, considering all the interacting elements in a typical machine tool, the number of errors that must be kept track of can be mindboggling. Therefore, the best way to keep track of and allocate allowable values for these errors is to use an *error budget*. An error budget, like any other budget, allocates resources (allowable amounts of error) among a machine's different components.* The goal is to allocate errors such that the ability of any particular component to meet its error allocation is not exceeded.

An error budget is formulated based on connectivity rules that define the behavior of a machine's components and their interfaces, and combinational rules that describe how errors of different types are to be combined. The first step in developing an error budget is to develop a kinematic model of the proposed system in the form of a series of homogeneous transformation matrices (HTM). The next step is to analyze systematically each type of error that can occur in the system and use the HTM model to help determine the effect of the errors on the toolpoint position accuracy with respect to the workpiece. The result is a list of all end point error components, their sources, and amplification-at-the-toolpoint factors (called the *error gains* or *sensitivities*). Different combinational rules can then be applied to yield upper and lower bound estimates of the total error in the machine.

Perhaps the most important step in assembling the error budget for a machine is the placement of the coordinate frames and the assignment of linear and angular errors corresponding to the axes. Angular motion errors suffer no ambiguity in their definition since they are unaffected by other errors and therefore can be defined with respect to any set of axes. Linear motion errors, on the other hand, must be carefully defined in terms of linear motion caused directly and linear motion that is the result of an Abbe error. During the design phase, the coordinate systems must be located at the origin of the angular errors, where pitch, yaw, and roll errors are the center of stiffness as defined above.

Combinational Rules for Errors

Different types of errors do not mix unless handled properly. Once all error components are multiplied by their respective error gains, a final combination of errors can then be made to yield an educated guess as to the machine's expected performance. There are three common types of errors, which are defined as:**

1. "*Random* — which, under apparently equal conditions at a given position, does not always have the same value, and can only be expressed statistically."
2. "*Systematic* — which always have the same value and sign at a given position and under given circumstances. (Wherever systematic errors have been established, they may be used for correcting the value measured.)" Systematic errors can generally be correlated with position along an axis and can be corrected if the relative accompanying random error is small enough.
3. "*Hysteresis* — is a systematic error (which in this instance is separated out for convenience). It is usually reproducible, has a sign depending on the direction of approach, and a value partly dependent on the travel. (Hysteresis errors may be used for correcting the measured value if the direction of approach is known and an adequate pretravel is made.)" Backlash is a type of hysteresis error that can be compensated for to the extent that is repeatable.

Systematic and hysteresis errors can often be compensated for to a certain degree using calibration techniques. Random error cannot be compensated for without real-time measurement and feedback into a correcting servoloop. Thus, when evaluating the error budget for a machine, three distinct *subbudgets*

---

* See Donaldson, R. Error budgets. In *Technology of Machine Tools,* Vol. 5. *Machine Tool Accuracy,* R.J. Hocken, (ed.), Machine Tool Task Force.

** These definitions are from the CIRP Scientific Committee for Metrology and Interchangeability, 1978. A proposal for defining and specifying the dimensional uncertainty of multiaxis measuring machines. *Ann. CIRP.* 27(2):623–630.

based on systematic, hysteresis, and random errors should be kept. Often inputs for the error budget are obtained from manufacturers' catalogs (e.g., straightness of linear bearings), and they represent the peak-to-valley amplitude errors e_{PV}. A peak-to-valley error's equivalent random error with uniform probability of occurrence is given by $\varepsilon_{\text{equiv. random}} = \varepsilon_{PV}/K_{PV}$. If the error is Gaussian, then $K_{PVrms} = 4$ and there is a 99.9937% probability that the peak-to-valley error will not exceed four times the equivalent random error.

In the systematic subbudget, errors are added together and sign is preserved so cancellation may sometimes occur. The same is true for the hysteresis subbudget. In the random subbudget, both the sum and the root-mean-square error should be considered, where the latter is given by

$$\varepsilon_{\text{irms}} = \left(\frac{1}{N} \sum_{i=1}^{N} \varepsilon_{\text{irandom}}^2 \right)^{1/2} \tag{11.8.15}$$

Note that in the random subbudget, all the random errors are taken as the 1σ values. For the final combination of errors, the 4σ value is typically used, which means that there is a 99.9937% chance that the random error component will not exceed the 4σ value. In this case the total worst-case error for the machine will be

$$\varepsilon_{\text{iworst case}} = \sum \varepsilon_{\text{isystematic}} + \sum \varepsilon_{\text{ihysteresis}} + 4 \sum \varepsilon_{\text{irandom}} \tag{11.8.16}$$

The best-case error for the machine will probably be

$$\varepsilon_{\text{ibest case}} = \sum \varepsilon_{\text{isystematic}} + \sum \varepsilon_{\text{ihysteresis}} + 4 \left(\sum \varepsilon_{\text{irandom}}^2 \right)^{1/2} \tag{11.8.17}$$

In practice, the average of these two values is often used as an estimate of the accuracy the design is likely to achieve.

Type of Errors

There are many types of errors that occur in a machine including:

Straightness. *Straightness* is the deviation from true straight-line motion. One generally thinks of the straightness error as primarily dependent on the overall geometry of the machine and applied loads. As shown in Figure 11.8.13 straightness error can be considered the deviation from a straight line motion along a linear axis. The unofficial term *smoothness* can be used to describe straightness errors that are dependent on the surface finish of the parts in contact, the type of bearing used, and the bearing preload. In other words, the smoothness of motion would be the deviation from the best-fit polynomial describing the straightness of motion. Smoothness is not intended to be a descriptor of surface finish, but rather a descriptor of high-frequency straightness errors whose wavelength is typically on the order of the magnitude of the error, normal to the surface of two moving bodies in contact.

Kinematic Errors. Kinematic errors are defined here as *errors in the trajectory of an axis that are caused by misaligned or improperly sized components.* For example, kinematic errors include orthogonality (squareness or perpendicularity) and parallelism of axes with respect to their ideal locations and each other. Translational errors in the spatial position of axes are also a form of kinematic error. The dimensions of an axis's components can also cause the tool or workpiece to be offset from where it is supposed to be and is also classified as a kinematic error.

As shown in Figure 11.8.14 given two machine axes of motion in the XZ plane with one machine axis aligned with the X reference axis, the orthogonality error is defined as the deviation ε_y from 90° between the machine's other axis and the reference X axis. This is a straightforward definition which is simple to specify on a drawing, but is not necessarily simple to measure or control in production. Parallelism between two axes has horizontal and vertical forms that define the relative taper and twist

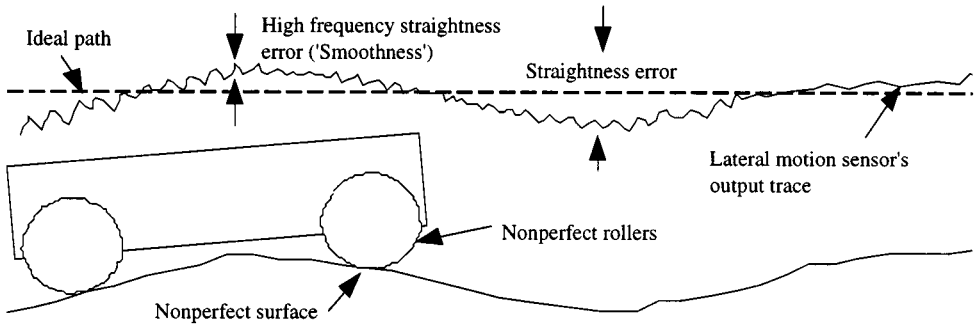


FIGURE 11.8.13 Strightness errors caused by surface form and finish errors.

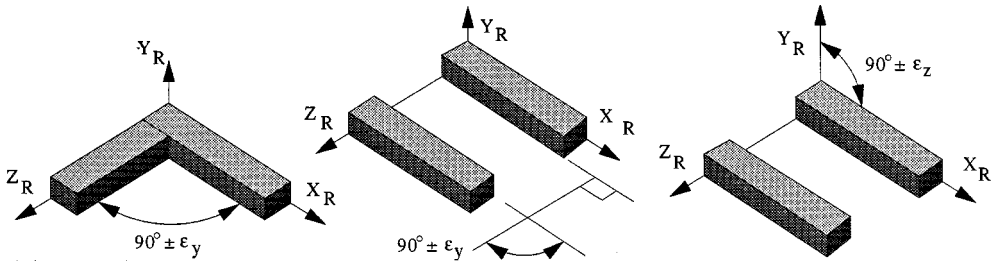


FIGURE 11.8.14 Orthogonality and horizontal and vertical parallelism errors.

between the two axes, respectively. An example of horizontal parallelism error (taper) would be an axis on a lathe that was not parallel to the spindle's axis of rotation. Using this axis to move a tool along the outer surface of a part would result in the part becoming tapered along its length. An example of vertical parallelism error is two axes used to support a milling machine bed. If one end of one of the axes is high, the bed itself will be warped when bolted to the axes, and parts machined while bolted to the bed will wobble.

Kinematic errors in a well-designed and manufactured machine would be very repeatable and can be compensated for if the controller is well designed. However, the fundamental principle of modern precision machine design is still to *maximize mechanical performance for a reasonable cost before using special controllers, algorithms, and sensors to correct for mechanical errors.*

External Load-Induced Errors. External loads that cause errors in a machine include gravity loads, cutting loads, and axis acceleration loads. The difficulty in modeling load-induced errors lies in their often distributed and/or varying effects. The types of errors discussed thus far have been geometrically induced and were a function of position. Thus, they could be relatively easily included in the HTM model of a machine. Load-induced errors, on the other hand, are often distributed throughout the structure. In order to incorporate them into the HTM model, a method for lumping them at discrete points must be devised. Depending on the structure, the bearing interface is often the most compliant part of the structure, and it can make sense to lump load-induced errors at the bearing interfaces. For more complex structures, it may be necessary to introduce additional coordinate frames into the HTM model.

In addition to increasing accuracy for enhancement of quality, machines are being required to move at greater speeds in order to increase productivity. Machine tools are usually thought of as big, bulky, slow-moving structures. The next generation of machine tools, however, will probably require axes to have acceleration capabilities in excess of 1 g. Deflections caused by accelerating the mass are acceptable for many drilling operations which would allow high spindle speeds and high axis feed rates to be used to increase productivity. Acceleration and deceleration rates of high-speed manufacturing equipment's

axes may one day routinely approach several g. In designing this type of equipment, accuracy along the path of motion may or may not be critical. However, final placement and settling time, where the maximum accelerations and inertial forces are present, will be important. Note that in the design of this type of machinery, cutting forces are often insignificant compared to inertial forces.

Another major contributor to load-induced errors in machine tools and some robots are cutting forces. Fortunately, high-speed cutting processes often generate low cutting forces, so the problems of high acceleration and high cutting force usually do not occur simultaneously. However, cutting forces are applied at the tool tip and act on every element in the machine. In order to estimate forces generated by the cutting process, actual cutting forces on machines with similar tools should be measured or appropriate handbooks consulted. In many cases, the rapid advance of new types of cutting tool materials and tool shapes will require the design engineer to consult with a tooling manufacturer or make experimental tests.

Load-Induced Errors from Machine Assembly. Even if all machine components are within required tolerance prior to assembly, additional load-induced errors can be introduced during assembly. The first type of error is forced geometric congruence between moving parts. An example is the bolting of a leadscrew to a carriage, where the nonstraightness of the leadscrew shaft creates a straightness error in the carriage with a period equal to the lead of the screw. A common example is mounting a mirror at its four corners, which often creates a visible distortion. The second type of error is the effect of the assembly process on the stiffness of the structure itself, and how the stiffness can be evaluated and incorporated into the HTM model. A third type of error, one that can be predicted, is the deformation of the machine when forces are applied to preload bearings and bolts. In addition, errors may also be caused by clamping or locking mechanisms.

Thermal Expansion Errors[*]. The need for ever-increasing accuracy and greater machine speeds makes thermal errors ever more important to control. Errors caused by thermal expansion are among the largest, most overlooked, and misunderstood form of error in the world of machine design. Thermal errors affect the machine, the part, and the tool. Even the warmth of a machinist's body can disrupt the accuracy of an ultraprecision machine. Figure 11.8.15 shows the thermal effects that must be accounted for in the design of a precision machine.[**] Thermal errors are particularly bothersome because they often cause angular errors that lead to Abbe errors.

Temperature changes induce thermal elastic strains ε_T, that are proportional to the product of the coefficient of expansion α, of the material and the temperature change, ΔT, experienced by the material:

$$\varepsilon_T = \alpha \Delta T$$

In addition, temperature gradients often cause angular errors, which lead to Abbe errors. If a machine, a tool, and a workpiece all expanded the same amount and could all be kept at the same temperature, then the system might expand uniformly with respect to the standard (always measured at 20°C) and everything would be within tolerance when brought back to standard temperature. However, different metals manufactured at different temperatures can experience serious dimensional metrology problems. Fortunately, standards have been developed (e.g., ANSI B89.6.2[***]) that define in great detail the effects of temperature and humidity on dimensional measurement and how measurements of these effects should be made. Although thermal strains can be minimized by using materials that do not expand very much,

[*] C° denotes temperature difference, whereas °C denotes an absolute temperature. This helps to avoid confusion, particularly when reference is made to temperature increases above ambient.

[**] J. Bryan, figure presented in the keynote address to the International Status of Thermal Error Research, *Ann CIRP*, Vol. 16, 1968. Also see McClure R. et al. 3.0 Quasistatic machine tool errors. October 1980. In *Technology of Machine Tools*, Vol. 5, *Machine Tool Accuracy*, NTIS UCRL-52960-5.

[***] Available from ASME, 22 Law Drive, Box 2350, Fairfield, NJ 07007-2350, (201)882-1167.

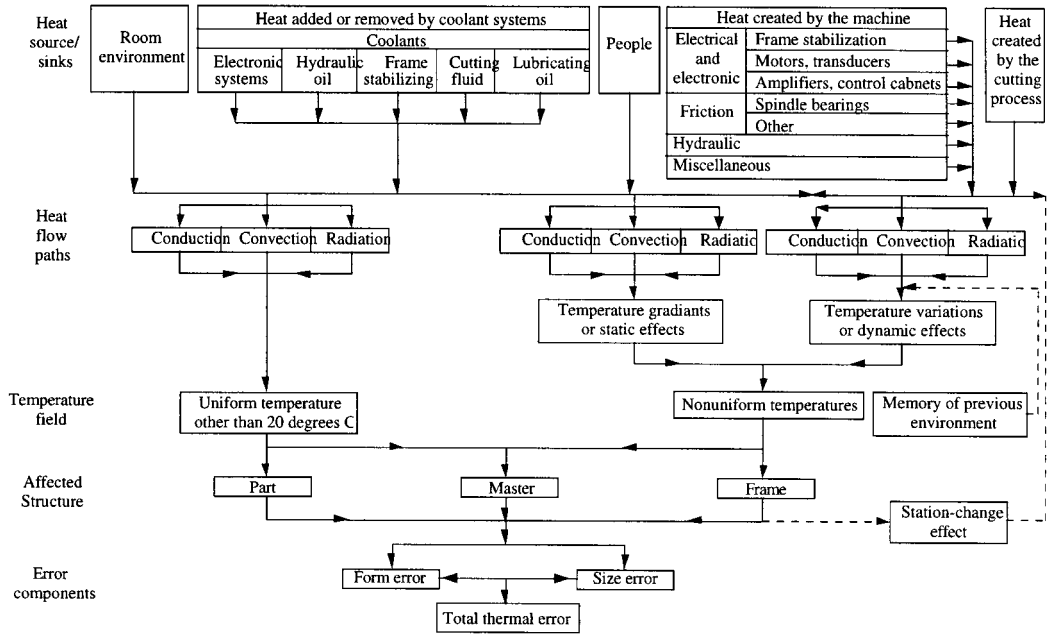


FIGURE 11.8.15 Thermal effects in manufacturing and metrology. (After Bryan.)

and by accurately controlling the environment; in practice, however, these can be difficult solutions to implement for economic reasons.

Structures

The detailed design of the machine structure requires an understanding of the errors in a machine and how they interact with the envisioned assemblage of components. When designing the structure, the engineer must have an overall philosophy in mind with respect to how to assemble the system and address principle types of errors.

For a precision machine, the engineer must first develop a strategy for attaining accuracy (these issues were addressed in the context of 11.8.1 “Analysis of Errors in a Precision Machine”):

- Accuracy obtained from component and assembly accuracy
  - Inexpensive once the process is perfected.
  - Accuracy is strongly coupled to thermal and mechanical loads on the machine.
- Accuracy obtained by error mapping a repeatable system
  - Inexpensive once the process is perfected.
  - Accuracy is moderately coupled to thermal and mechanical loads on the machine.
- Accuracy obtained from a metrology frame (measure the position of an inaccurate machine with respect to an accurate reference frame)
  - Expensive, but sometimes the only choice.
  - Accuracy is uncoupled to thermal and mechanical loads on the machine.

Next, a direction should be set regarding dynamic performance and how much is needed, by considering the following:

- System bandwidth requirements
- Effects of changing system parameters
- Methods to add damping
- Experimental modal analysis

Thermal errors are among the most difficult to predict and control, and an approach to address them should be established early on:

- Passive temperature control
- Active temperature control

With design approaches to the above, the next step is to consider structural materials and different machine configurations that are available:

- Materials
- Existing configurations

Finally, a decision should be made regarding the overall approach to be used with regard to how all the components are assembled:

- Elastically averaged design
- Kinematic design
- Bolted joint design
- Kinematic coupling design

These issues are discussed in greater detail below. Proper consideration of all the details takes a great deal of study* and practice, but these highlights should be useful as an introduction, or as a refresher.

Dynamic System Requirements

Engineers commonly ask "how stiff should the structure be?" A minimum specified static stiffness is a useful but not sufficient specification. For example, the machine can be made to not deflect too much under its own weight or the weight of a part. Fortunately, it can be predicted with reasonable accuracy. To obtain good surface finish or dynamic performance, the dynamic stiffness needs to be specified.

The dynamic stiffness of a system is the stiffness measured using an excitation force with a frequency equal to the damped natural frequency of the structure. The dynamic stiffness of a system is also equal to the static stiffness divided by the amplification at dynamic resonance. It takes a large amount of damping to reduce the amplification factor to a low level. There are several damping quantifiers that are used to describe energy dissipation in a structure. The quantifiers include:

η	Loss factor of material
η_s	Loss factor of material (geometry and load dependent)
$Q = A_r$	Resonance amplification factor
ϕ	Phase angle ϕ between stress and strain (hysteresis factor)
δ_{Ld}	Logarithmic decrement**
ΔU	The energy dissipated during one cycle
ζ	The damping factor associated with second-order systems

* See, for example, Slocum, A. 1997. *Precision Machine Design.* Society of Manufacturing Engineers, Dearborn, MI. Much of the contents of this section was derived from this book.

** Most texts on vibration refer to the log decrement as δ, however, to avoid confusion with discussions on displacement termed δ, the log decrement will be referred to here as δ_{Ld}.

The various damping terms are related (approximately) in the following manner:

$$\eta = \frac{1}{A_r} = \frac{\delta}{\pi} = \phi = \frac{\Delta U}{2\pi U} \tag{11.8.18}$$

The logarithmic decrement δ_{Ld} is an extremely useful measure of the relative amplitude between N successive oscillations of a freely vibrating system (one excited by an impulse):

$$\delta_{Ld} = \frac{-1}{N}\log_e\left(\frac{a_N}{a_1}\right) \tag{11.8.19}$$

The logarithmic decrement can also be related to the damping factor ζ, velocity damping factor b, mass m, and natural frequency ω_n of a second-order system model:

$$\zeta = \frac{\delta_{Ld}}{\sqrt{4\pi^2 + \delta_{Ld}^2}} \tag{11.8.20}$$

Note that the amplification at resonance of a second-order system is given by

$$Q = A_r = \frac{1}{2\zeta\sqrt{1-\zeta^2}} \quad (\zeta \le 0.707) \tag{11.8.21}$$

For example, Figure 11.8.16 shows the time decay of a fairly well-damped system. For this system, n = 4, a_5 = 0.5, a_1 = 1.5, δ_{ld} = 0.275, ζ = 0.44, and Q = 11.5.

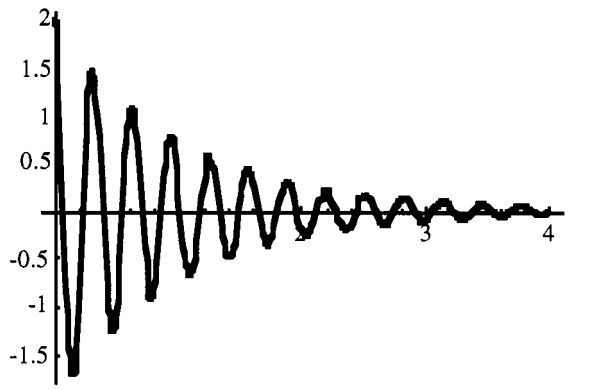


FIGURE 11.8.16 Time decay signal from a fairly well-damped system struck with an impulse.

Machines with good dynamic performance achieve damping by material selection and by the bolted joints and bearing interfaces in the structure. Unfortunately, material and joint damping factors are difficult to predict and are often too low. Thus the designer must consider the stiffness, mass, and damping of the system, and be prepared to alter them to achieve the desired level of performance. For high speed or high accuracy machines, damping mechanisms may have to be designed into the structure in order to meet realistic damping levels. The first step, however, is to try and determine the bandwidth required with an initial estimate of the stiffness, mass, and damping in the system. Then with these goals, one can try to change the parameter that has the greatest impact on performance.

System Bandwidth Requirements. An estimate of the system's servo bandwidth can be made by considering the motions the system is required to make. Most systems null high frequency disturbance forces with their own mass or added damping; however, lower frequency forces must be offset by forces from the controller/actuator. As a result, the servo bandwidth required is primarily a function of the types of moves that the system will be required to make. For start and stop moves (e.g., in a wafer stepper), the bandwidth should be on the order of

$$\omega(\text{hz}) = \frac{10}{2\pi t_{\text{settling time}}} \tag{11.8.22}$$

When contouring, the X and Y axis moves according to x = Rsinωt and y = Rcosωt, the frequency ω is a function of the linear velocity of the cutter through the material, and the radius of the contour:

$$\omega(\text{hz}) = \frac{v_{\text{linear}}}{2\pi r_{\text{path radius}}} \tag{11.8.23}$$

For example, for a large circle (e.g., 200 mm D) being cut at modest speeds (e.g., 0.1 m/sec), the bandwidth required is only 0.16 Hz. On the other hand, for a sharp turn in a corner, the radius of the cutter path trajectory may only be 1 mm, so ω > 16 Hz. The latter is a reasonable requirement for a large machine tool, where the spindle axis may weigh 1000 kg and thus require a critically damped system stiffness of 10 N/μm, which means a design stiffness on the order of 100 N/μm when one considers the limited damping that is obtainable.

In order to make a better assessment of what system parameters should be, a simple model can be used in the early stages of design. For example, a system with a motor driving a carriage can be modeled as shown in Figure 11.8.17.

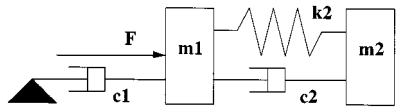


FIGURE 11.8.17 Effect of changing system mass on dynamic performance.

In this model:

- m_1 is the mass of a linear motor forcer or m_1 is the reflected inertia of the motor rotor and leadscrew (or just a linear motor's moving part):

$$M_{\text{reflected}} = \frac{4\pi^2 J}{l^2}$$

- m_2 is the mass of the carriage.
- c_1 is the damping in the linear and rotary bearings.
- c_2 is the damping in the actuator-carriage coupling and the carriage structure.
- k_2 is the stiffness of the actuator and actuator-carriage-tool structural loop.

The equations of motion of this system are

$$\begin{vmatrix} m_1 & 0 \\ 0 & m_2 \end{vmatrix} \begin{vmatrix} \ddot{x}_1 \\ \ddot{x}_2 \end{vmatrix} + \begin{vmatrix} c_1 + c_2 & -c_2 \\ -c_2 & c_2 \end{vmatrix} \begin{vmatrix} \dot{x}_1 \\ \dot{x}_2 \end{vmatrix} + \begin{vmatrix} k_2 & -k_2 \\ -k_2 & k_2 \end{vmatrix} \begin{vmatrix} x_1 \\ x_2 \end{vmatrix} + \begin{vmatrix} F(t) \\ 0 \end{vmatrix} \tag{11.8.24}$$

The transfer function x_2/F (dynamic response of the carriage) is

$$\frac{x_2}{F} = \frac{k_2 + c_2 s}{c_1 s\left(k_2 + c_2 s + m_2 s^2\right) + \left(m_1 + m_2\right) s^2 \left(k_2 + c_2 s\right) + m_1 m_2 s^4} \tag{11.8.25}$$

Note that the product of the masses term tends to dominate the system. To illustrate the design implications, consider four different calculated design options for a linear motion system shown in Figure 11.8.18.

Actuator	ballscrew	lin. motor	lin. motor	lin. motor
Bearings	linear ball	linear ball	air	air
Structural damping	no	no	no	yes
material damping zeta	0.005	0.005	0.005	0.1
actuator to ground zeta	0.05	0.03	0	0
m1 (actuator) (kg)	50	5	5	5
m2 (carriage), kg	50	50	50	50
c1 (N/m/s)	187	355	0	0
c2 (N/m/s)	19	19	19	374
k1 (N/m)	1.75E+08	1.75E+08	1.75E+08	1.75E+08
Bandwidth (Hz)	25	100	30	100

FIGURE 11.8.18 Calculated parameters of four possible linear motion systems.

As a guideline, the servo bandwidth of the system is generally limited by the frequency at which the servo can drive the system without exciting structural modes. Without special control techniques, the servo bandwidth can be no higher than the frequency found by drawing a horizontal line 3 dB above the resonant peak to intersect the response curve. This method is used only to initially size components. A detailed controls simulation must be done to verify performance and guide further system optimization. As an example, consider the response of the ballscrew-driven carriage supported by rolling element linear bearings, as shown in Figure 11.8.19. In this case, since preloaded linear guides and a ballnut are used, damping to ground will be high. The inertia of the ballscrew will lower the system frequency considerably (note the $m_1 m_2 s^4$ term in the transfer function). Going up 3 dB from the resonance peak and projecting to the left to intersect the response curve gives an estimate of the maximum bandwidth the machine can be driven at, without danger of exciting a resonance, of about 25 Hz.

Effects of Changing the System's Mass. Decreasing the mass of the system will enhance the ability of the machine to respond to command signals. However, the trade-off is that the system will lose the ability to attenuate high frequency noise and vibration as shown in Figure 11.8.20.

A system with decreased mass offers a higher natural frequency, which means that higher speed controller signals can be used without compromising the accuracy of the system. A low mass system also shows improved damping, a result of the increased loss factor [the loss factor $\zeta = c/(2m)$], but a low mass system shows less noise rejection at higher frequencies. This suggests that the machine will be less able to attenuate noise and vibration. Considering all the issues, in conclusion, the mass should be minimized to reduce controller effort and improve the frequency response and loss factor (ζ).

Effects of Changing the System's Stiffness. A system with higher stiffness will give a flatter response at low frequencies and give smaller displacements for a given force input. More importantly, the compromise of decreased noise attenuation is not as dramatic as is the case with lowering the system mass. This is shown by the similar shapes in the three curves at high frequencies in Figure 11.8.21. However, acoustical noise may be worsened by adding stiffness. In conclusion, the stiffness of the machine structure should be maximized to improve positioning accuracy.

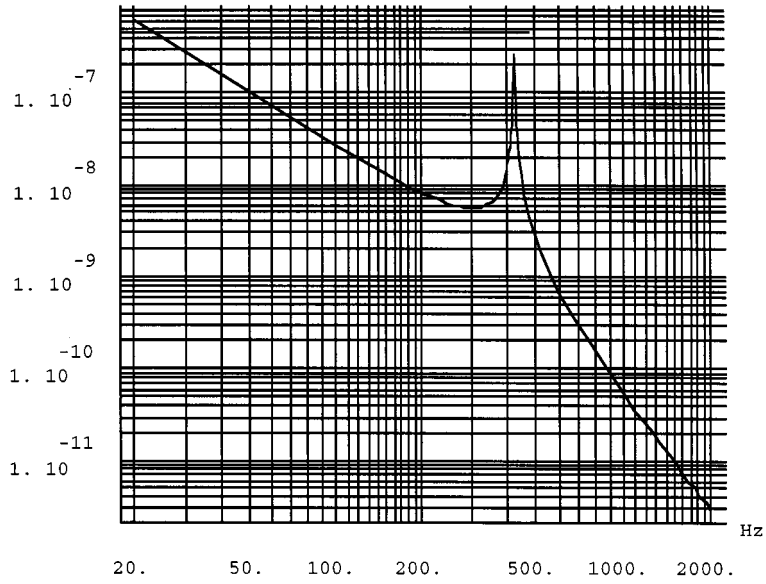


FIGURE 11.8.19 Frequency response of a ballscrew-driven system.

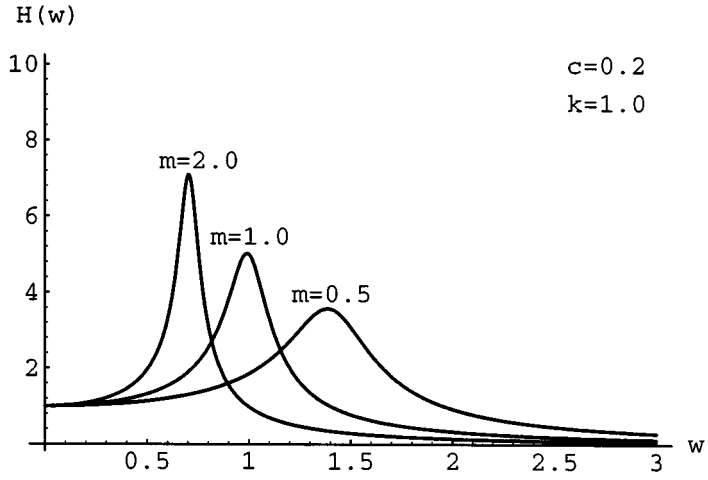


FIGURE 11.8.20 Effect of changing system mass on dynamic performance.

Effects of Changing the System's Damping. Increasing the system's damping can make a dramatic improvement in the system's response. The trend is for decreasing amplification of the output at resonance with increasing damping, although a damping coefficient of 0.4 may be difficult to obtain in practice. Figure 11.8.22 shows the dramatic improvement available by doubling the system's damping.

Methods of Achieving Damping

Although it has been extensively studied, the mechanism of damping in a material is difficult to quantify and one must generally rely on empirical results.* In fact, damping is highly dependent on alloy

* A discussion of the many different microstructural mechanisms that generate damping in materials is beyond the scope of this book. For a detailed discussion see Lazan, B.J., *Damping of Materials and Members in Structural Mechanics*. Pergamon Press, London.

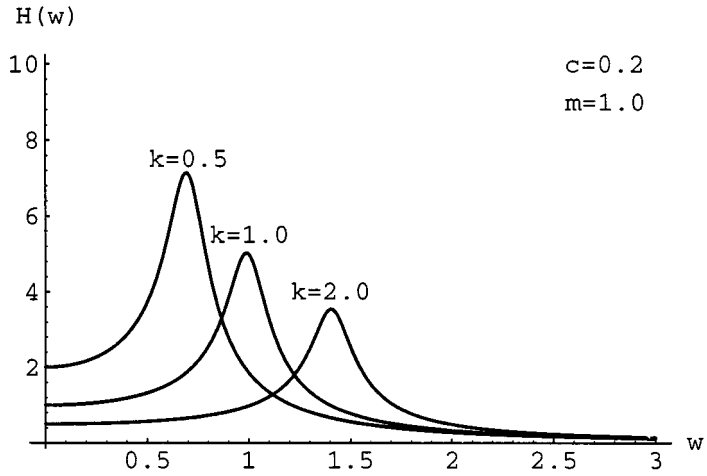


FIGURE 11.8.21 Effect of changing system stiffness on dynamic performance.

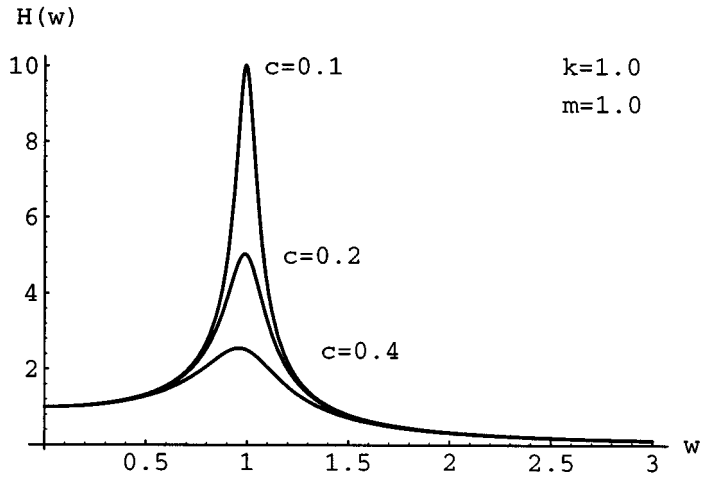


FIGURE 11.8.22 Effect of changing system damping on dynamic performance.

composition, frequency, stress level and type, and temperature. Structural damping levels are often quite low, and frequently the dominant source of damping is the joints in an assembly. In fact, one must be extremely wary of damping data that are presented in the literature, because often it is presented without a discussion of the design of the test setup.

The amount of damping one obtains from a material is very low compared to the amount of damping that one can obtain with the addition of a damping mechanism. Damping mechanisms can range from simple sand piles to more complex shear dampers or tuned mass dampers as discussed below. In general, cast iron is the best damped structural metal. Polymer concrete or granite make well-damped structures that tend to be more monolithic in nature; but ceramics have very poor damping.

Tuned Mass Dampers. In a machine with a rotating component (e.g., a grinding wheel), there is often enough energy at multiples of the rotational frequency (harmonics) to cause resonant vibrations in some of the machine's components. This often occurs in cantilevered components such as boring bars and some grinding wheel dressers. The amplification at a particular frequency can be minimized with the use of a tuned mass damper. A tuned mass damper is simply a mass, spring, and damper attached to a structure at the point where vibration motion is to be decreased. The size of the mass, spring, and damper

are chosen so they oscillate out of phase with the structure and thus help to reduce the structure's vibration amplitude.

Consider the single-degree-of-freedom system shown in Figure 11.8.23. The system contains a spring, mass, and damper. For a cantilevered steel beam, the spring would represent the beam stiffness, the mass would be that which combined with the spring yielded the natural frequency of the cantilevered beam, and the damper would be that which caused a 2% energy loss per cycle. As also shown in Figure 11.8.23, a second spring-mass-damper system can be added to the first to decrease the cantilevered beam's amplitude at resonance. The equations of motion of the system are

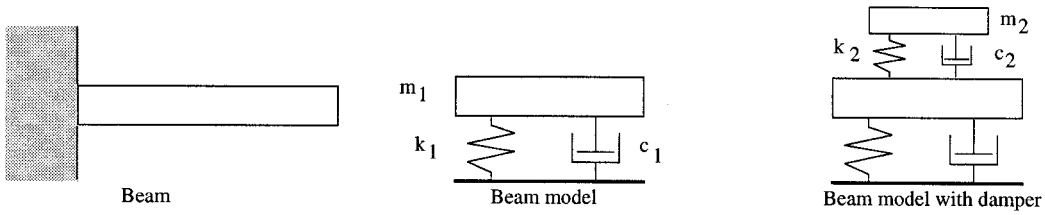


FIGURE 11.8.23 Cantilever beam, model, and model with tuned mass damper.

$$\begin{bmatrix} m_1 & 0 \\ 0 & m_2 \end{bmatrix} \ddot{x}(t) + \begin{bmatrix} c_1 + c_2 & -c_2 \\ -c_2 & c_2 \end{bmatrix} \dot{x}(t) + \begin{bmatrix} k_1 + k_2 & -k_2 \\ -k_2 & k_2 \end{bmatrix} x(t) \tag{11.8.26}$$

In the frequency domain, in order to present a solution for the motion of the system, the following notation is introduced:

$$Z_{ij}(\omega) = -\omega^2 m_{ij} + i\omega c_{ij} + k_{ij} \qquad i, j = 1, 2 \tag{11.8.27}$$

The amplitudes of the motions of the component and the damper as a function of frequency are given by*

$$X_2(\omega) = \frac{-Z_{12}(\omega) F_1}{Z_{11}(\omega) Z_{22}(\omega) - Z_{12}^2(\omega)} \tag{11.8.28}$$

The design of a tuned mass damper system for a machine component may involve the following steps:

- Determine the space available for the damper and calculate the mass (m_2) that can fit into this space.
- Determine the spring size (k_2) that makes the natural frequency of the damper equal to the natural frequency of the component.
- Use a spreadsheet to generate plots of component amplitude as a function of frequency and damper damping magnitude (c_2).

Constrained Layer Dampers. The structural joints in a machine tool have long been known to be a source of damping by the mechanisms of friction and microslip. A study of structural joint damping has shown that numerous theories are available for predicting damping by these mechanisms,** however, the amount of damping obtained is still less than what is required for critical damping, and controlling the surface interface parameters at the joint to achieve uniform results from machine to machine is difficult. In addition, as far as the accuracy of the machine is required, it would be best if the joints behaved as

* Ibid., p 115.

** See, for example, Tsutsumi, M. and Ito, Y. September 1979. Damping mechanisms of a bolted joint in machine tools, *Proc. 20th Int. Mach. Tool Des. Res. Conf.*, pp. 443–448; and Murty, A.S.R. and Padmanabhan, K.K. 1982. Effect of surface topography on damping in machine joints. *Precis. Eng.* 4(4):185–190.

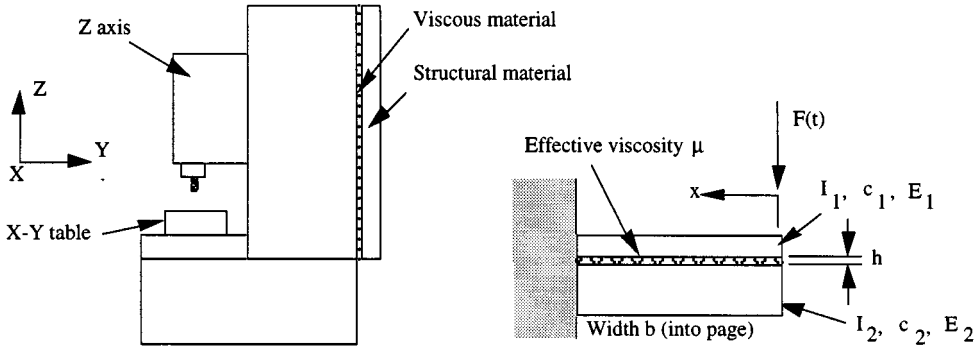


FIGURE 11.8.24 Increasing structural damping by adding alternate layers of viscous and structural materials.

a rigid interface which, as discussed below, can be obtained by bolting and grouting (or bonding) a joint. A better method is to add damping by applying alternate layers of viscoelastic and structural materials to a machine tool structure,* as illustrated in Figure 11.8.24.

Experimental Modal Analysis of a Machine's Dynamic Performance

The manner in which a machine behaves dynamically has a direct effect on the quality of the process. It is vital to be able to measure a machine's dynamic performance to enable engineers to determine exactly which parts of a machine need to be strengthened or made to be better damped.

Experimental modal analysis allows the study of vibration modes in a machine tool structure. An understanding of data acquisition, signal processing, and vibration theory is necessary to obtain meaningful results. For a modest-sized company (50 employees), it is feasible and reasonable to have one person dedicated to metrology, including experimental modal analysis. Companies routinely invest $100,000 or more in precision measurement devices (e.g., gauges, autocollimator, laser interferometer, ball bar, surface roughness, etc.). The dynamic performance of a machine is every bit as important, and requires only a modest investment (about $40K).

When a company's metrologist is trained in modal analysis, results can be rapidly obtained and presented to the design department. For example, suppose a new machine is being designed where axial stiffness of the drive system is of extreme importance (e.g., a centerless grinder). The design engineer wants to know if a rollerscrew will give better performance than a ballscrew. Not completely trusting catalog data, and not knowing the primary source of compliance (the mounting, the bearings, the shaft, or the nut), the engineer replaces a ballscrew — in a known good machine — with a rollerscrew. The metrologist does the modal analysis and the engineer learns in "real time" what the result is. Similarly, the engineer can look to other machines on the shop floor for physical models of elements to be used in a new design, and the metrologist can make the measurements in a short time.

All too often a modal analysis is done after a design is complete to help solve a problem. The results can often be used to help repair the problem; however, done before hand, modal testing of similar machines or test-subassemblies would have led to a better design approach that in the end would have resulted in a better integrated system. The results of a modal analysis yield important information regarding the machine's dynamic properties, including:

- Modal natural frequencies
- Modal damping factors
- Vibration mode shapes

* See, for example, Haranath, S., Ganesan, N., and Rao, B. 1987. Dynamic analysis of machine tool structures with applied damping treatment. *Int. J. Mach. Tools Manuf.* 27(1):43–55. Also see Marsh, E.R., Slocum, A.H. 1996. An integrated approach to structural damping. *Precis. Eng.*, 18(2,3):103–109.

This information is vital to a designer, who may use the information to

- Locate sources of compliance in a structure.
- Characterize machine performance.
- Optimize design parameters.
- Identify the weak links in a structure for design optimization.
- Identify modes which are being excited by the process (e.g., an end mill) so that structure can be modified accordingly.
- Identify modes (parts of the structure) which limit the speed of operation (e.g., in a coordinate measuring machine).

A manager may ask "why should we start doing this if we have not done it in the past, and were very successful?" The engineering answer is that faster and more accurate machines are needed in the future, and the materials being processed are getting harder and harder (e.g., ceramics). The business answer is that the problems are getting tougher to solve, and engineering time is expensive. Also, international competition is more widespread and much tougher. History has shown that application of new advanced tools virtually always guarantees an increase in performance and competitive edge.

Identification, Control, and Isolation of Heat Sources*

Heat that causes thermal errors is introduced into the machine from a number of sources including moving parts (e.g., high-speed spindles, leadscrew nuts, linear bearings, transmissions), motors, the material removal process, and the external environment (e.g., sunshine through a window, direct incandescent lighting, heating ducts, the floor, operators' body heat, etc.). Heat transfer mechanisms in a machine include conduction, convection, evaporation,** and radiation. They are executed by recirculating lubricating oil and cutting fluid, chips from the cutting process, conduction through the frame of the machine, convection of air in and around the machine, and internal and external sources of radiation.

There are three schools of thought regarding minimization of thermally induced errors. The first is to prevent thermal expansion in the first place. The second is to minimize the time it takes for the machine to reach its equilibrium temperature, or in some cases, bring the entire machine to a uniform temperature (e.g., circulate, temperature-controlled fluid through the machine), which can help to minimize differential expansion. The third is to disregard the effects of thermal errors and simply map them, which is a very difficult and often impractical thing to do. Regardless of the methodology followed, a thorough understanding of the effects of heat sources and transfer mechanisms is required, which is beyond the scope of this work.

The manner in which the temperature of a machine is to be controlled can have a large impact on the machine's design. In summary, there are many options available to the design engineer:

- Passive temperature control:
  - Minimize and isolate heat sources.
  - Minimize coefficient of thermal expansion.
  - Maximize thermal conductivity to minimize thermal gradients.
  - Maximize thermal diffusivity to quickly equilibrate transient thermal effects.
  - Minimize thermal emissivity of the structure to minimize radiant coupling to the environment, or maximize the emissivity to couple the structure to an environmental control enclosure.

* An important reference to have that discusses many of the issues discussed here in greater detail is *Temperature and Humidity Environment for Dimensional Measurement*, ANSI Standard B89.6.2-1973.

** Evaporation cooling occurs when aqueous fluids are used. Evaporative cooling is usually uneven and represents one of the biggest temperature control problems facing the precision machine design engineer.

- Active temperature control:
  - Air showers
  - Circulating temperature-controlled fluid within the machine
  - Oil showers
  - Thermoelectric coolers for localized temperature control of hot spots

Each conceptual design option must consider how the temperature within the machine will vary if each of these different temperature control strategies were applied.

Material Considerations

For the conceptual design phase, one should design the structure using differently available materials, and then also design a multimaterial hybrid. For example, cast iron can be made into virtually any shape, so the design engineer has greater freedom, but large sections are expensive to thermally anneal, which must be done to achieve material homogeneity and stability. Granite is usually used in the form of simple rectangular, circular, or planar shapes. Ships are made from welded steel plate and thus conceivably any size of machine tool could also be welded together. Polymer concrete can be cast into virtually any shape and requires no stress relief or prolonged aging cycle. With new ceramics and composite materials, the choices become even more varied, so one really must be alert and consider all options.

Cast Iron Structures

The good general properties of cast iron and the ease with which parts can be cast have made cast iron the foundation of the machine tool industry. Generally, when a machine tool component is smaller than a compact car, it is a candidate for being made of cast iron, although castings weighing hundreds of tons have been made. For larger parts or where economy is of prime importance, one should consider welding together plates and standard structural shapes (e.g., boxes, angles, I-beams, and channels), as discussed below.

Sand casting can be used to make virtually any size part, and it basically involves making a pattern out of a suitable material (e.g., foam, wood, or metal) and packing sand around it. Parting lines are used to allow the sand mold components to be disassembled from around the pattern and then reassembled after the part is removed. Sand cores are often inserted into the mold to form cavities inside the mold (e.g., the cylinders of an engine casting). Regardless of the design of the part, one must also consider that as the metal cools it shrinks (on the order of 5 to 10% for most metals), and that in order to remove the part from the mold without breaking the mold, a taper (draft) of about 1:10 is required. In addition, extra metal should be added to surfaces that will have to be machined (a machining allowance), and locating surfaces should be added so that the part can be fixtured to facilitate machining. Thus in order to specify a casting, there are a few basic guidelines one needs to know in order to minimize the work that a professional mold design engineer has to do to clean up your design. These guidelines are discussed below.*

Granite Structures**

Granite is used as a structural material in machines that are generally used in dry environments because granite can absorb moisture and swell. Thus it may not be appropriate to use granite in a machine where cutting fluid splashes all over it, although there is some debate as to how susceptible to swelling granite actually is. Granite can be sawed into a part of virtually any shape or size (deviations from round slabs

* See, for example, the handbook *Casting Design as Influenced by Foundry Practice.* Mechanite Metal Corp., Marietta, GA.

** A good source of design information about what shapes and sizes can be made is available from Rock of Ages Corp., Industrial Products Group, P.O. Box 482, Barre, VT 05641.

or rectilinear shapes can be expensive). Since it is a brittle material, sharp corners are not allowed, and most structures are built from pieces that are bolted (using inserts) and grouted or bonded together. Since granite is brittle, threaded holes cannot be formed, and thus threaded inserts must be glued or press-fit into round holes in the granite. Common applications of granite components in machine tools include structural members and air-bearing ways in coordinate measuring machines and other inspection machines. Note that the porosity of some granite makes it unsuitable for air bearings even after it has been polished.

The low thermal conductivity of granite makes it slow to absorb heat. This makes granite, particularly black granite, susceptible to thermal distortions caused by the top surface absorbing heat from overhead lights. Granite's coefficient of thermal expansion is less than that of most metals, so in the process of manufacture, shipping, and use, one must consider how differential thermal expansion will affect a machine's performance if metal components have been bolted to it. Although granite has been sitting in the ground for millions of years and thus may seem to possess the ultimate stability, one must consider the effects of relieving the stress upon the granite's dimensional stability after it has been quarried. There are suppliers of very stable granite components, so the design engineer must carefully shop around. A very desirable property of granite (or any other brittle material) is that it will chip when banged instead of forming a crater with a raised lip (Brinell). Granite is also relatively inexpensive to quarry, cut, and lap. Hence granite is often the material of choice for coordinate measuring machine tables.

Welded Structures*

Welded structures can be made from any weldable alloy (e.g., iron alloys such as 1018 structural steel or Invar). Welded structures (weldments) are commonly used, for example, where (1) the cost of a large structure is to be minimized, (2) a high degree of material damping is not needed or the structure will be filled with a damping material, and/or (3) the part is too large to be cast and thermally stress relieved economically. If welded properly, so that the weld material is in a metallurgically stable form, residual stresses can sometimes be removed using vibratory stress-relief methods.

A welded structure is similar to a cast structure, in that strength and stiffness are achieved through the use of sections that are reinforced with ribs; consequently, structural analysis can be difficult and beyond the conceptual design phase often requires the use of finite element methods. Damping and heat transfer characteristics across welded joints are also difficult to model because they are dependent on depth of weld penetration, composition of the weld material, and contact pressure between member surfaces at the joints where the weld does not penetrate. One solution is to specify full penetration welds. To minimize the cost of welded structures, the number of parts and linear meters of weld must be minimized. Furthermore, the more welds that are made, the greater the thermal distortion that is likely to occur as a result of the manufacturing process. However, if too few ribs are used, large plate sections can vibrate like drums. Damping and thermal performance of a welded structure can improve greatly if a viscous temperature-cooled fluid is recirculated throughout the structure, or if damping mechanisms are used as described above. A welded structure can also be used as a mold for a polymer concrete casting which creates a heavy but well-damped and stiff structure, but care must be taken to avoid bimaterial thermal deformation problems.

Ceramic Structures**

The first introduction of a machine that was made almost entirely of ceramics was in 1984 at the Tokyo machine tool show. Although all-ceramic machines generally perform admirably, they have yet to compete economically with machines made from cast iron or polymer concrete. However, in the future,

* A good reference to have is Blodgett, O. 1963. *Design of Weldments*. James F. Lincoln Arc Welding Foundation, P.O. Box 3035, Cleveland, OH 44104.

** See, for example, Ormiston, T. September 1990. Advanced Ceramics and Machine Design. *SME Tech. Paper* EM90-353.

as more and more ceramic components are made for use in consumer products (e.g., automobiles), it is likely that precision machines will contain more and more ceramic components.*

Hard materials (e.g., ceramics) offer advantages over conventional materials in terms of dimensional stability, strength, and stiffness over a wide range of temperatures. In applications ranging from adiabatic internal combustion engines for maximum efficiency to X-ray mirrors for X-ray photolithography, the ability to manufacture components from hard materials is clearly of vital importance to the future of the manufacturing industry. Unfortunately, most ceramic components are finish machined on machines built from cast iron and the abrasive nature of ceramics limits the life of these machines. Thus a new family of machine tools and machine tool components will have to be developed specifically for the manufacture of ceramic components. Consider several key properties of some ceramic materials that can help guide the design process for new machines:

- Most ceramic materials (e.g., aluminum oxide and silicon nitride) will not corrode in any fluid environment that might be used in the manufacture of ceramic components (i.e., fluids from oil to water).** Thus ceramics are key materials for water hydrostatic bearings.
- The more brittle a material is, the less plastic deformation that is generated during finish grinding or lapping; hence the surface is more likely to have a negative skewness. A surface with a negative skewness minimizes the need for a lubricant that has good wetting properties that allows for water to be considered as a lubricant. A surface with a negative skewness will also suffer less damage in the event that the lubricant is lost.
- The more brittle a material is, the less plastic deformation that is generated during finish grinding; hence the surface is less likely to contain high residual stress levels that can lead to dimensional instability. In addition, a precision machine should not be subjected to shock loads in the first place. For general machine tool applications, such as bearing rails, the element geometry generally provides more than enough strength to withstand impact loads generated when a machine crashes.
- Generous radii must be used in all corners, and threaded steel inserts must be used if parts are to be bolted to ceramic components. To bond two ceramic components together, conventional adhesives can be used, or for high-performance applications, ceramic parts can be frit bonded. In frit bonding, a glass powder is applied to the two mating surfaces and then fused by heating the parts above the melting point of the glass. The bond will not be as strong as the ceramic, but it will be almost as stiff.
- Unlike some metals, elements do not precipitate out of a ceramic material's microstructure with time (e.g., carbides do not form like in some iron alloys) so dimensional stability is enhanced. Thus ceramic components can also have virtually unmatched dimensional stability. For metrology masters (e.g., squares and straight edges), ceramics are much lighter than conventional materials, they are less likely to be damaged (dropped) in the first place, and they are less likely to become scratched or worn in everyday use.
- Most ceramics are pure and thus the achievable surface finish is limited only by the size of the grain structure formed during the sintering process; however, most advanced ceramic materials have submicron sizes and this effect is usually not a problem the way it can be with metals. Also note that metals contain discrete hard particles that are dragged across the softer surface during machining which degrades surface finish.

* Furukawa, Y. et al. 1986. Development of ultra precision machine tool made of ceramics. *Ann. CIRP.* 35(1): 279–282.

** Aluminum oxide components are subject to stress corrosion cracking in aqueous environments and thus care must be taken to minimize tensile stresses. Note that silicon-based ceramics, which have predominantly covalent molecules, are far less reactive with water. Only at high temperature and pressure do silicon-based ceramics become appreciably affected by aqueous environments.

- Ceramic materials have a high modulus that is good for machine stiffness, but some have poor thermal properties (e.g., alumina) that can lead to increased thermal deformations of the machine. Note that silicon carbide has very good thermal properties but it is considerably more expensive than alumina
- Ceramic materials in intimate contact will not gall or fret the way many metals often do; thus they make excellent bearing surfaces. However, when ceramic materials are in intimate sliding contact, traction forces can cause the local tensile strength to be exceeded, which produces surface cracks that lead to spalling. In such situations, it may be desirable to use a ceramic material with a high fracture toughness and tensile strength (e.g., silicon nitride). Thermal stresses can also initiate local or gross failure.
- Ceramic materials have a higher modulus of elasticity and lower density than do bearing steels; hence ceramic rolling elements have a smaller contact zone that leads to less heat generation. Hybrid bearings (e.g., metal rings and ceramic rolling elements) generate 30 to 50% less heat than do steel bearings. This means that grease can be used to lubricate the bearings at much higher speeds.* In general, for smaller diameter ultraprecision bearings, hybrid bearings can have up to three times the DN value of steel bearings (i.e., 4.5 million vs. 1.5 million).
- Aluminum oxide has good overall properties for structural applications such as fluidstatic bearing rails and CMM structures. Zirconia is very tough, and its coefficient of thermal expansion matches that of steel, so it can be used as a bearing rail liner for rolling element bearings without worrying about bimaterial expansion problems. However, most zirconias are multiphase materials and thus are not well suited for applications where high-dimensional stability is required (e.g., in precision bearings). Silicon nitride has the best overall properties including very high toughness, which makes it ideal for rolling element bearings, but it is too expensive to use for large structural components. Silicon carbide and tungsten carbide are extremely hard and wear-resistant and they are often used in cutting tool applications.
- Aluminum oxide components can be made by cold pressing followed by machining, firing, and grinding. Note that there is significant volume shrinkage during the firing process. Hence designing ceramic structural parts often requires assistance from the manufacturer. In general, the same shape design rules apply as for metal castings, and the wall thickness should not be greater than about 25 mm. Ceramics components (e.g., those made from silicon nitride) can also be made by hot pressing followed by grinding and lapping.

Polymer Concrete Castings**

Portland cement-based concrete is not dimensionally stable enough, due to its own internal structural variations with time and its hygroscopic nature, to allow it to be used for the main structure of a precision machine tool. Although in many applications, properly cured reinforced concrete on a stable, dry subgrade can provide a reasonably stable foundation for very large machines that are not self-supporting, unreinforced Portland cement concrete itself is not dimensionally stable due to (1) reaction shrinkage from cement hydration; (2) shrinkage due to loss of excess nonstoichiometric water, which leaves conduits for humidity-induced expansion or contraction; and (3) nonelastic dimensional changes (e.g., creep and microcracking in the inherent brittle/porous structure). Overall, strain variations with time may be as high as 1000 μm/m.

* Steel bearings require oil mist lubrication at higher speeds. Introducing an oil mist (oil dripped into a high-pressure air stream) into a bearing increases the chance that water and dirt particles might be introduced into the bearing, which can lead to premature failure. Actually, a precision bearing cooled by an oil mist should have its air cleaned and dried to a level usually associated with air bearings (e.g., 3-μm filter and H_2O < 50 to 100 ppm).

** See, for example, Capuano, T. September 10, 1987. Polymer concrete. *Mach. Des.* 133–135; Jablonowski, J. August 1987. New ways to build machine structures. *Am. Mach. Automat. Manuf.* 88–94; and McKeown, P.A. and Morgan, G.H. 1979. Epoxy granite: a structural material for precision machines. *Precis. Eng.* 1(4): 227–229.

Fortunately, a number of different types of polymer-based concretes have been developed which can be used to cast machine tool quality structures. For example, Fritz Studer AG, a prominent manufacturer of precision grinders in Switzerland, discovered that special polymers can be used to bind together specially prepared and sized aggregate to yield a stable, essentially castable, granite-like material with a damping coefficient much higher than that of cast iron.* By carefully controlling the manufacturing process and selection of binder and aggregate, properties can be varied somewhat to suit the user. The polymer concrete material and process developed by Studer is known as Granitan™ and its composition and manufacturing process was patented. Numerous companies have licensed the process and will make castings from Granitan™ to order. Other companies have developed their own proprietary polymer concretes with similar high performance properties.

For polymer concrete castings, the same rules for draft allowance apply as for metal castings if the mold is to be removed. Instead of ribs, polymer concrete structures usually use internal foam cores to maximize their stiffness-to-weight ratio. Unlike metal castings, a polymer concrete casting will not develop hot spots while curing even in thick, uneven sections. Polymer concrete castings can readily accommodate cast-in-place components such as bolt inserts, conduit, bearing rails, hydraulic lines, etc. It should be noted that a bolt will fail before a bonded-in-place insert.

With appropriate section design, polymer concrete structures can have the stiffness of cast iron structures and much greater damping than cast iron structures.** However due to polymer concretes' lower strength, heavily loaded machine substructures (e.g., carriages) are still best made from cast iron. Polymer concretes do not diffuse heat as well as cast iron structures and thus attention must be paid to the isolation of heat sources to prevent the formation of hot spots in a polymer concrete structure. In addition, their modulus of elasticity is about one fifth that of steel, and their strength is an order of magnitude less, so they are used primarily for machine bases.

Structural Configurations

For proper functioning of moving axes and operational stability, often it is important to minimize thermal and elastic structural loops. By this it is meant that the path length from the toolpoint to the workpiece through the structure should be minimal. The less material that separates the part of the structure that holds the tool and the part of the structure that holds the part, the more likely the entire system will quickly reach and maintain a stable equilibrium.

Open-Section (C or G) Frames

Most small machines are designed with an open frame, as shown in Figure 11.8.25 which greatly facilitates workzone access for fixturing and part handling. Note that the machine could be designed with the spindle oriented in the horizontal or the vertical direction. Unfortunately, open section frames are the least structurally and thermally stable. The lack of symmetry leads to undesirable thermal gradients and bending moments. The fact that a critical part of the structure is cantilevered means that Abbe errors abound; hence great care must be used when designing a precision machine with an open frame. Note that there are many different variations on this design for different types of machines (e.g., milling machines and lathes). The common feature of all is the fact that the structural loop is open.

Closed-Section (Bridge or Portal) Frames

Most large machines are designed with a closed frame as shown in Figure 11.8.26. When the Z motion is built into the bridge, a second actuator must often be slaved to the primary actuator that moves the

* Kreienbühl, R. September 19–20, 1990. Experience with synthetic granite for high precision machines. In *Proc. Symp. Mineralguss im Maschinenbau.* FH Darmstadt; and Renker, H.J. 1985. Stone based structural materials. *Precis. Eng.* 7(3): 161–164.

** See, for example, Weck, M. and Hartel, R. 1985. Design, manufacture, and testing of precision machines with essential polymer concrete components. *Precis. Eng.* 7(3): 165–170; and Salje, I. et al. 1988. Comparison of machine tool elements made of polymer concrete and cast iron. *Ann CIRP.* 37(1): 381–384.

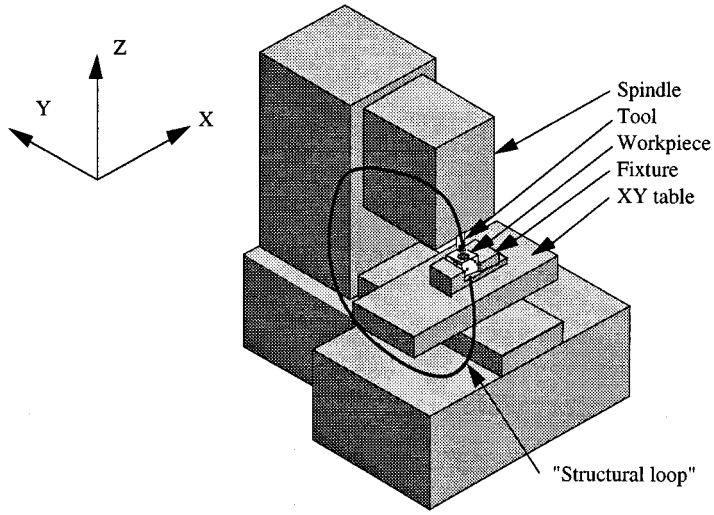


FIGURE 11.8.25 Structural loop in an open-frame machine tool.

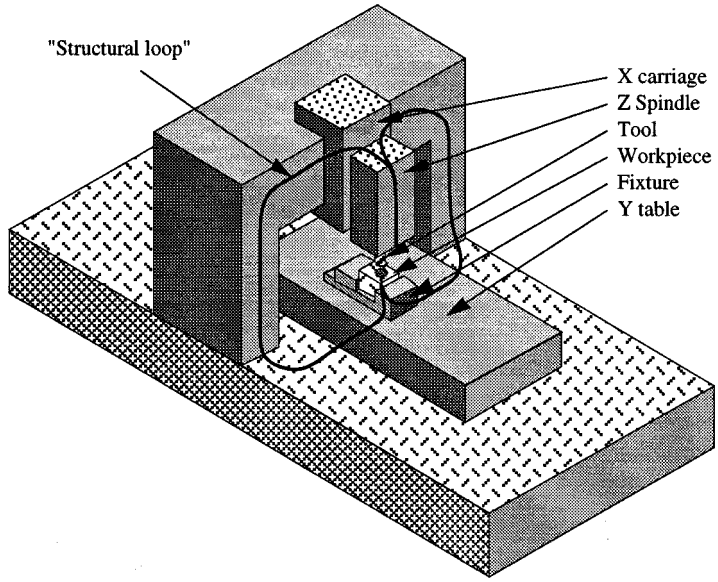


FIGURE 11.8.26 Structural loop in a closed-frame (bridge) machine tool.

bridge. This prevents the bridge from yawing (walking). Note that there are many different variations on this design for different types of machines (e.g., milling machines and lathes). The common feature of all is the fact that the structural loop is closed.

Tetrahedral Frames

Nature invented the tetrahedron and found it to be an immensely stable and powerful form (e.g., diamonds). In engineering and architecture, the tetrahedron represents the three-dimensional application of the age-old structure of stability, the triangle. Lindsey of NPL in England took these basic building blocks of nature and added well-engineered damping mechanisms to yield a major advancement in the

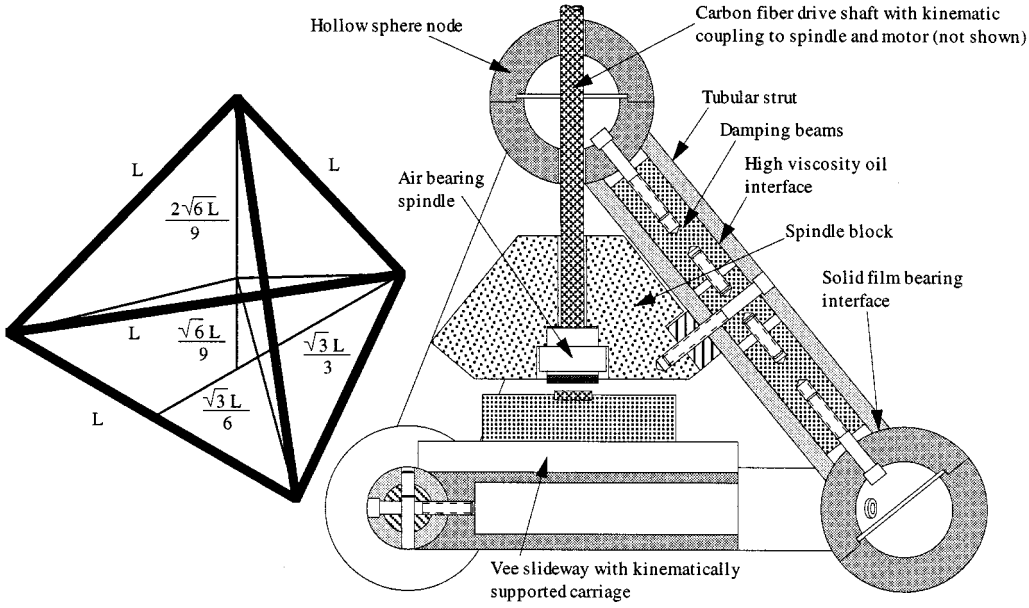


FIGURE 11.8.27 Tetraform structural concept for machine tools and instruments developed by Lindsey. (Courtesy of the National Physical Laboratory.)

structure of machine tools. The *Tetraform* machine tool concept is shown in Figure 11.8.27. The structures owes its exceptional dynamic performance to the following:*

- Damping in the legs is achieved through the use of inner cylinders which dissipate energy through viscous shear. Energy is dissipated via squeeze film damping and relative sliding motion damping.
- Damping at the joints is achieved by a sliding bearing (PTFE) interface.
- Microslip at the joints does not affect the dimensional stability of the machine because the minimum energy form of the tetrahedron wants to be preserved. Unlike a plane joint which can continue to slip and lead to dimensional instability, the tetrahedron's legs' spherical ends want to stay on the spherical joint nodes.

The latter point has even more profound consequences, in that it makes the use of composite materials in the structure an attractive alternative to metals or ceramics. Wound carbon fiber tubes can be designed to have a zero coefficient of thermal expansion along their length and they can have stiffness-to-weight ratios two times higher than are obtainable with metals. It would be difficult to design a conventional machine tool that made economical use of the desirable properties of carbon fibers.

Counterweights and Counterbalances

Counterweights, shown in Figure 11.8.28 can be used to minimize gravity loads. This helps minimize the holding torque required of servomotors, which in turn minimizes motor size and heat input to the machine. For a very high precision machine, however, the bearings used to guide the counterweights and support the pulley bearings should have negligible static friction. Counterweights also increase the mass of the system, which can decrease dynamic performance. In the case of quasistatic axes (e.g., large gantry-type surface grinders), the counterweight can increase the load on the structure that supports the axes, and if the structure is cantilevered, then as the counterbalanced axis moves, the deflection errors

* This concept is protected by worldwide patients. See, for example, U.K. patent 8,719,169, or contact the British Technology Group, 101 Newington Causeway, London, SE1 6BU, England.

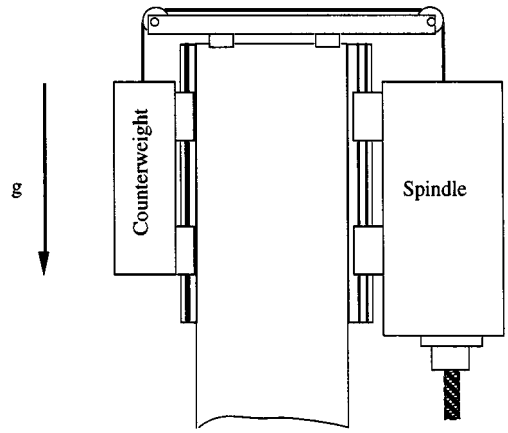


FIGURE 11.8.28 Use of a counterweight to balance the deadweight of an axis.

caused by the weight of the counterbalanced axis will change. Alternatively, a separate frame and low-accuracy slave carriage can be used to support the counterweights.

A counterbalance can be any passive means used to support the weight of an axis that moves in a vertical direction. Pistons have been used successfully but they can impart frictional and misalignment forces. A very effective counterbalance method is to use a float. This also provides viscous damping if the fluid used is a viscous oil.

Compensating Curvatures

All structures have finite stiffness, and when loads are applied, lateral and angular displacements will result. To compensate for these effects, it is possible to shape the otherwise straight ways of an axis so that the sum of the deflections and the intentional nonstraightness results in minimal net straightness error. This type of correction is known as a *compensating curvature*. Usually, it is very difficult to also compensate for angular errors. If the error budget is properly assembled, it can be used to plot the total error as a function of the position of an axis. Once the plot is found, it can be used to help design a shape (ideally, the inverse of the plot) that will cancel the lateral and hopefully also angular errors for the given bearing design. Compensating curvatures can easily be measured using an autocollimator.

When the load does not greatly change on parts moving along axes that are curvature compensated (e.g., axes that carry a measuring head), the method can be effective. If the loads do change greatly (e.g., a table carrying different weight parts and fixtures), the compensating curvature can sometimes decrease performance. One of its main attractions, however, is that once the correct compensated curvature and manufacturing process is found for a particular machine, it can be the least expensive way to correct for straightness errors.

With modern mapping techniques, compensating curvatures are used primarily on very large structures or when angular errors caused by deflection are too large and cannot be corrected for by another axis. For example, they might be used in a situation where otherwise the toolpoint would enter the workpiece at an angle rather than being perpendicular. A rotary axis would be required to compensate for this type of angular error, as opposed to just another linear axis, which can only be used to compensate for an Abbe error.

Structural Connectivity

The principle of kinematic design states that point contact should be established at the minimum number of points required to constrain a body in the desired position and orientation (i.e., six minus the number of desired degrees of freedom). This prevents overconstraint, and thus an “exact” mathematically continuous model of the system can be made. Kinematic locating mechanisms range from simple pins to Gothic arch-grooved three-ball couplings shown in Figure 11.8.29. Kinematic designs, however, are

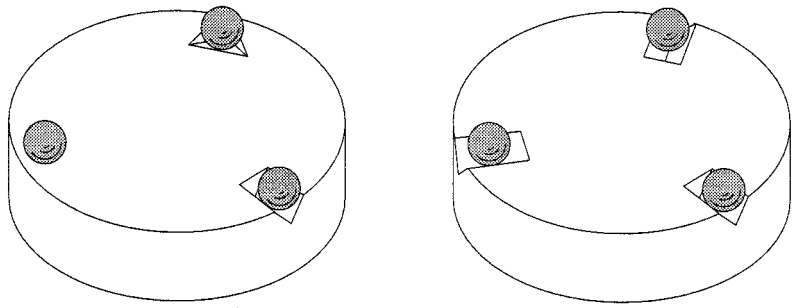


FIGURE 11.8.29 Flat-groove-tetrahendron (Kelvin clamp) and three-vee-groove kinematic couplings. In both cases, the balls are mounted to an upper plate (not shown) which is held in position with respect to the lower plate by the coupling.

subject to high-contact stresses, which often may require the use of ceramic components (e.g., silicon nitride balls and grooves) if the highest level of performance is to be achieved. If stress and corrosion fatigue are controlled, repeatability of a kinematic system can be on the order of the surface finish of the points in contact if the loads are repeatable or the preload high enough. Finite contact areas do exist, and they effectively elastically average out the local errors due to surface roughness. In addition, note that friction and microindentation will limit the accuracy of the kinematic model.

Kinematic support of a structure is often desired to ensure that the structure is not deformed by inaccuracies or instabilities of the mounting surface. For a small instrument, only one of the kinematic mounting points may be rigidly connected and the other two may consist of flexures that will allow for differential thermal growth between the instrument and the mounting surface. Friction does exist in kinematic couplings, and thus forces can be transmitted between a mounting surface and an instrument.

The principle of *elastic averaging* states that to accurately locate two surfaces and support a large load, there should be a very large number of contact points spread out over a broad region. Examples include curvic or Hirth couplings, which use meshed gear teeth (of different forms, respectively) to form a coupling. The teeth are clamped together with a very large preload. This mechanism is commonly used for indexing tables and indexing tool turrets. Figure 11.8.30 shows an indexing and clamping mechanism for a lathe tool turret. Note that many different types of clamping preload systems exist. However, this type of mechanism causes the system to be overconstrained; on the other hand, if an elastically averaged system is properly designed, fabricated, and preloaded, the average contact stress will be low, high points will wear themselves in with use, and errors will be averaged out by elastic deformation. The system itself will then have very high load capacity and stiffness. For a worn-in elastically averaged system, the repeatability is on the order of the accuracy of the manufacturing process used to make the parts divided by the square root of the number of contact points (i.e., teeth in a tooth coupling). Still, because of the large number of contact points, the chance of dirt contaminating the interface increases.

With respect to the connectivity between structural elements (e.g., a headstock and a bed), elements of a structure can either be connected together via:

- Kinematic design
  - Deterministic
  - Less reliance on manufacturing
  - Stiffness and load capacity limited
- Elastically averaged design
  - Nondeterministic
  - More reliance on manufacturing
  - Stiffness and load capacity not limited

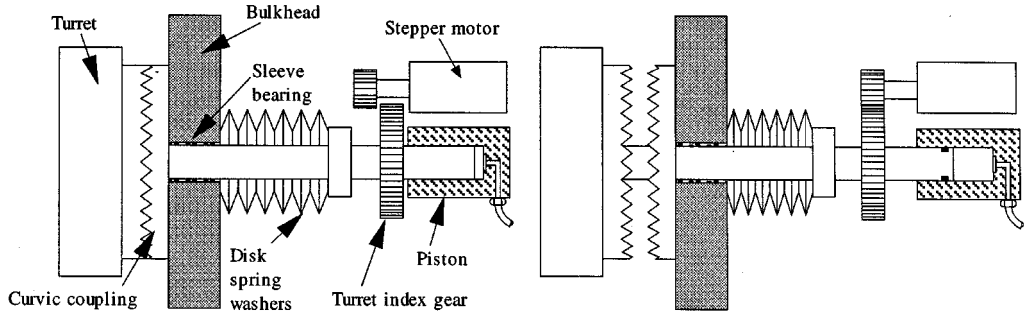


FIGURE 11.8.30 Typical turret indexing and locking mechanism in engaged and disengaged positions.

It should be noted that overconstrainted systems cannot accommodate differential thermal growth and hence are more prone to warping. Furthermore, finite deformation of the contact surface leads to micromechanical constraints that limit the true kinematicity of the structure. The use of hard materials (e.g., ceramics) helps to minimize the latter problem. For permanently connected components, one can align them using a kinematic location system and then inject epoxy or grout to create a bond between all the surfaces in close proximity, although with this method, one must make sure that the shrinkage of the bond material does not warp the components.

Bolted Joints*

Bolts can be used to prevent two parts from separating or sliding relative to one another. For the former, the tensile forces across the joint are transferred through the bolt shaft. For the latter, sliding motion is resisted by frictional forces generated from the normal load on the joint produced by tightening of the bolt and the coefficient of friction between the joint's parts. Because more than one bolt is usually used at a joint, it would be virtually impossible to ensure a tight fit of the bolt shafts in the holes, so it is not even worth trying. Sufficient lateral stiffness is usually provided by bolt preload and joint friction. For better resistance to shock loads, parts can be bolted in place and then holes drilled, reamed, and pinned with hardened steel dowels or roll pins. *In situ* drilling and reaming of the holes through both parts while they are bolted together maintains hole alignment, so multiple pins can be used.

Figure 11.8.31 illustrates the cross section of a typical portion of a bolted joint. A common bolt configuration used to bolt bearing rails for T-slides is shown in Figure 11.8.32. Many rails have a double row of bolts. In general, the cantilevered length should not exceed the bedded length. Ideally, the bedded length should be about 1.5 times the cantilevered length, but sadly this often takes up too much room. Furthermore, the stress cones under the boltheads should ideally overlap to maximize stiffness and minimize rail waviness. In practice, however, as shown in Figure 11.8.33 adequate stiffness can be obtained from a wider bolt spacing. The total stiffness for N segments of a bearing rail (e.g., N segments under a bearing pad) is given by**

$$K_{\text{Rail}} = \frac{N_{\text{Segments}}}{\dfrac{1}{K_{\text{Rail bend \& shear}}} + \dfrac{1}{K_{\text{Joint}} + \dfrac{1}{\dfrac{1}{K_{\text{Flange comp}}} + \dfrac{1}{K_{\text{Flange shear}}} + \dfrac{1}{K_{\text{Bed Shear}}} + \dfrac{1}{K_{\text{Bolt}}}}}}$$

* For a more detailed discussion, see, for example, Bickford, J.H. 1981. *An Introduction to the Design and Behavior of Bolted Joints.* Marcel Dekker, New York; as well as A. Blake's book.

** The theory and practical implementation implications of bolted joints for machine tool applications are beyond the scope of this work. These are discussed in greater detail in Slocum, A. 1997. *Precision Machine Design.* Society of Manufacturing Engineers, Dearborn, MI.

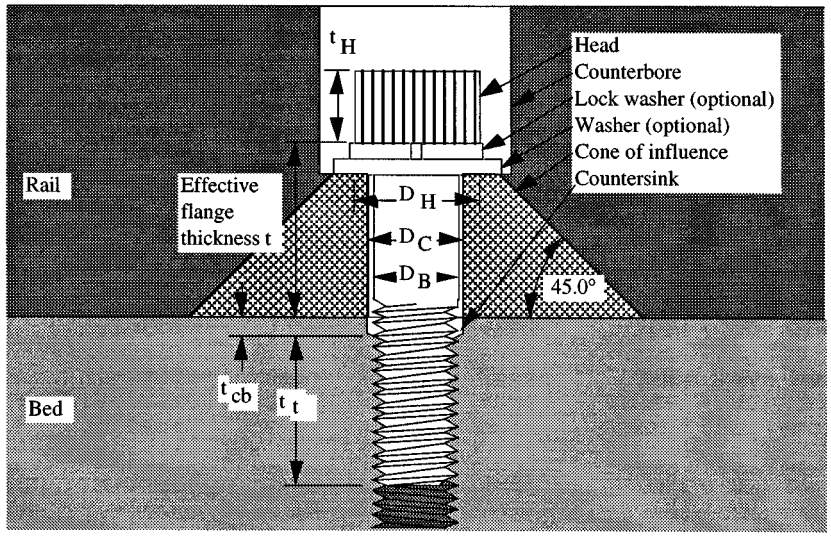


FIGURE 11.8.31 Typical components of a bolted assembly.

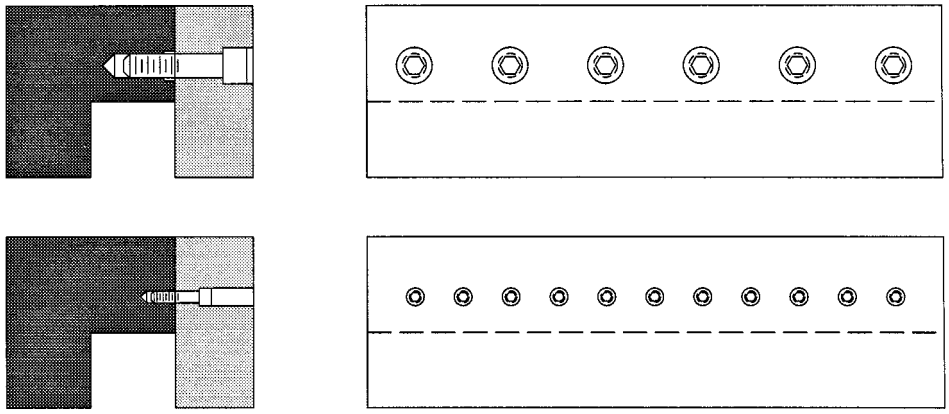


FIGURE 11.8.32 Counterbored and bolted bearing rails with equal stiffness.

Bolt diameter is not a sensitive parameter for stiffness when bolt spacing is made a function of the bolt diameter. When the length of the bolt and the cone are expressed as bolt lengths, bolt joint stiffness becomes linearly dependent on the bolt diameter. As a result, bolt diameter cancels out as shown in Figure 11.8.33.

Bearings

Since bearings are such a critical element in precision machines, one must *think* about the seemingly innumerable number of bearing design considerations that affect the performance of a machine, including:

- Speed and acceleration limits
- Range of motion
- Applied loads
- Accuracy
- Repeatability
- Sealability
- Size and configuration
- Weight
- Support equipment
- Maintenance

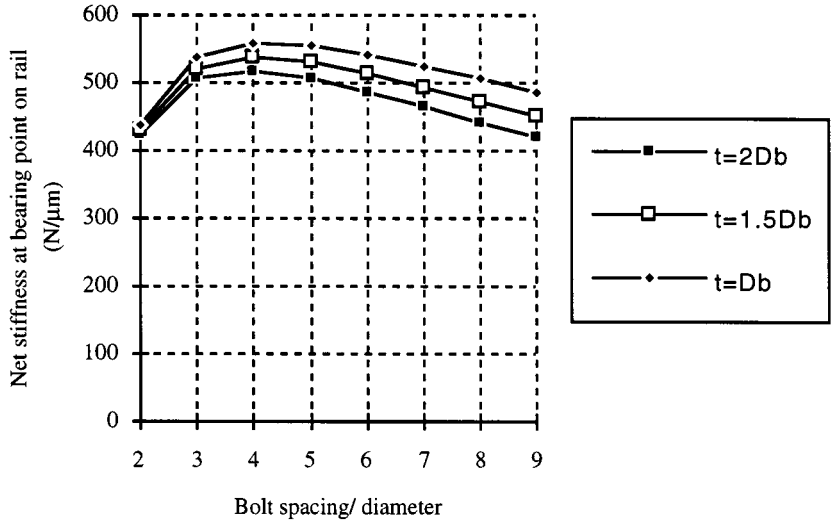


FIGURE 11.8.33 Relationships among stiffness and bolt spacing.

- Resolution
- Preload
- Stiffness
- Vibration and shock resistance
- Damping capability
- Friction
- Thermal performance
- Environmental sensitivity
- Material compatibility
- Mounting requirements
- Required life
- Availability
- Designability
- Manufacturability
- Cost

Because there are always design trade-offs in choosing a bearing for a precision machine, all of these factors must be considered simultaneously by the design engineer. In addition, there are several issues that are common to most types of bearings. These issues include surface roughness, preload, and replication of bearings in place.

Surface Roughness

Surface roughness is a characterization of the profile of the surface and often has an effect (although difficult if not impossible to characterize) on the smoothness of a bearing's motion. In terms of the manufacturing process, smoothness of motion of a bearing can only be quantified in terms of the surface roughness and bearing design:

- Sliding contact bearings tend to average out surface finish errors and wear less when the skewness is negative. A positive skewness (defined below) can lead to continued wear of the bearing.
- For rolling element bearings, if the contact area is larger than the typical peak-to-valley spacing, an elastic averaging effect will occur and a kinematic arrangement of rollers will produce smooth motion. If this condition is not met, the effect will be like driving on a cobblestone street. If numerous rolling elements are used, the effects of elastic averaging can help to smooth out the motion.
- Hydrostatic and aerostatic bearings are insensitive to surface finish effects when they are considerably less than the bearing clearance.
- Flexural and magnetic bearings are not sensitive surface finish.

There are three common parameters for specifying the surface finish or roughness*: the *root mean square* (rms or R_q), the *centerline average* (R_a), and the International Standards Organization (ISO) 10-point height parameter (R_z). The latter is with respect to the five highest peaks and five lowest valleys on a sample. A surface profile measurement yields a jagged trace. If a best fit straight line is drawn through a section of the trace of length L, then the R_q, R_a, and R_z, surface finishes are defined, respectively, from deviations y from the line as a function of distance x along the sample:

$$R_q = \sqrt{\frac{1}{L}\int_0^L y^2(x)\,dx} \qquad R_a = \frac{1}{L}\int_0^L y(x)\,dx \qquad R_z = \frac{\sum_{i=1}^{5} y_{\text{peak}}(i) - y_{\text{valley}}(i)}{5} \tag{11.8.29}$$

Unfortunately, these measures do not provide any information as to the topographical characteristics of the surface. As shown in Figure 11.8.34 the surface topography can be characterized by the *skewness*. The skewness is the ratio of the third moment of the amplitude distribution and the standard deviation σ from the mean line drawn through the surface roughness measurements. Hence the skewness provides a measure of the shape of the amplitude distribution curve. For a contact-type bearing, valleys separated by wide flat planes may be acceptable. This form would have a negative skewness value and is typically in the range –1.6 to –2.0 for bearing surfaces. Sharp spikes would soon grind off, creating wear debris and more damage, and hence positive skewness values are unacceptable for contact-type bearing surfaces. The skewness is defined mathematically from:

$$\text{skew} = \frac{\mu_3}{\sqrt{\mu_2}} = \frac{\mu_3}{\sigma} \qquad \mu_n = \int_{-\infty}^{\infty} (y-\mu)^n f(y)\,dy \qquad \mu = \int_{-\infty}^{\infty} y f(y)\,dy \tag{11.8.30}$$

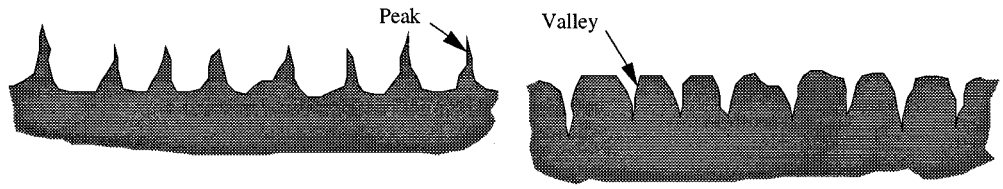


FIGURE 11.8.34 Surfaces with positive (left) and negative (right) skewness.

Numerous other methods exist for defining the shape and intensity of surface roughness features. For example, the autocorrelation function is used to check for the degree of randomness in a surface. This can be used to help track down periodic components (e.g., those caused by tool chatter), which can then sometimes be reduced in subsequently made parts. A frequency spectrum analysis can also be used to accomplish this. The science of surface metrology is constantly evolving as surface finish requirements increase, and the interested reader should consult the literature.**

* See, for example, Stout, K. May 1980. How smooth is smooth. *Prod. Eng.* The following discussion on surface roughness is derived from this article. For a detailed discussion of this subject, see *Surface Texture (Surface Roughness, Waviness, and Lay)*, ANSI Standard B46.1-1985, American Society of Mechanical Engineers, 345 East 47th St., New York, NY 10017. Also see Vorburger, T. and Raja, J. *Surface Finish Metrology Tutorial.* National Institute of Standards and Technology Report NISTIR 89-4088 (301-975-2000).

** See, for example, *Journal of Surface Metrology*, edited by K. Stout and published by Kogan Page, London.

Bearing Preload

A preloaded bearing is one where one bearing pushes against another, thereby squeezing the bearing rail. This allows the bearing to resist bi-directional loads without motion nonlinearity (e.g., backlash) when the load reverses direction. As the force on a pair of opposing pads preloaded against each other is applied, one bearing pushes harder on the rail by an amount equal to the product of the pad stiffness and the carriage deflection, and one pushes less by the same amount.

Insufficient preload is synonymous with at least some of the bearing points periodically losing contact with the bearing surface. This results in difficult-to-map error motions and decreased stiffness, possibly leading to chatter of the tool. Tool chatter, in turn, degrades the surface finish of the part. However, with contact-type bearings, preload can accelerate wear and can lead to stiction, which decreases controllability. These issues lead to the desirability of externally pressurized fluid film bearings (air or fluid) where preload will not change with time.

Unless the axes of the machine are designed to utilize the weight of the machine itself as a preload, in many cases the preload will change with time. This is particularly true if contact-type bearing surfaces wear and if the structure relaxes, due to internal stress needed to maintain the preload. In an effort to counter these effects, the classic method has been to use a device known as a gib to generate the preload. In general, a gib can be defined as a mechanical device that can apply preload to a bearing. In the classical sense, a gib was a wedge-shaped part that was advanced by the action of a screw. When a gibbed machine wears, the gibs, bearings, and sometimes the ways must be rehand-finished and the gibs adjusted to provide the proper preload.

In many machines it is common to use modular rolling bearing components whose preload is set by using oversized balls or rollers, or by tightening bolts that push on a plate contacting the rollers. When the latter type of system wears, it is often discarded and replaced with a new unit. Hence large economies of scale can be obtained by some bearing component manufacturers.

Replicated Bearings*

Replication is a process whereby a very low shrinkage polymer is poured or injected around a master shape that has been coated with mold release. When the polymer cures, it has the shape of the master, which can then be removed and used again. This process can be used to form a bearing surface itself, such as a sliding bearing or a hydrostatic bearing with a pocket, or to form a mating surface between a modular bearing carriage and rough surface of a structure such as a casting.

Because hardening of many polymer resins is an exothrmic reaction, one of the important aspects of this process is to minimize the amount of polymer used, and to maximize the stiffness and thermal diffusivity of the master and the part. This is required to prevent the heat of polymerization from heating the structure, deforming it, and then the polymer hardening to the thermally deformed shape. Unlike bolted assemblies, once the polymer cures to form the desired shape (e.g., a trough in which a linear bearing rail is placed), alignment is no longer possible.

Figure 11.8.35 shows schematic details of surface preparation required to ensure the debonding does not occur. This sawtooth pattern can be obtained in a number of different ways. In addition to the rough surface finish required, the surface must also be thoroughly cleaned and a mold release applied to the master.

Sliding Contact Bearings

Sliding contact bearings are the oldest, simplest, least expensive bearing technology, and they still have a wide range of applications, from construction machinery to machines with atomic resolution. Sliding bearings are thus a very important element in the machine design engineer's tool kit. Sliding contact bearings utilize a variety of different types of lubricants between various interface materials. Lubricants

* See, for example, Devitt, A. October 1989. Replication techniques for machine tool assembly. In *East Manuf. Tech. Conf.* Springfield, MA. Available from Devitt Machinery Co., Twin Oaks Center, Suite G, 4009 Market Street, Aston, PA 19104.

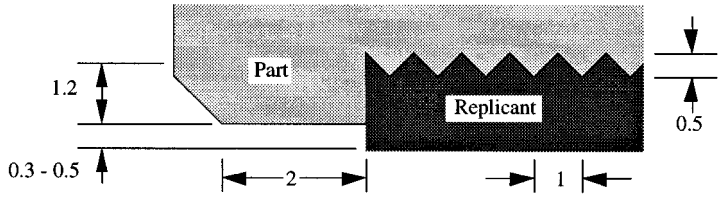


FIGURE 11.8.35 Details of surface and region for replication (dimensions are in mm).

range from light oil to grease to a solid lubricant such as graphite or a PTFE polymer. Because they often distribute loads over a large area, contact stresses and space requirements are often low, while stiffness and damping are usually high. In this section, general properties of sliding contact bearings are discussed followed by a discussion of design considerations.

There are numerous types of sliding contact bearings available and in the context of discussing their general properties, some specific categories will be discussed. In general, one should note that all sliding contact bearings have a static coefficient of friction that is greater (to some degree, no matter how small) than the dynamic coefficient of friction (static μ > dynamic μ). The difference between the static and dynamic friction coefficients will depend on the materials, surface finish, and lubricant.

Rolling Element Bearings

Figures 11.8.36 and 11.8.37 show representative types of ball and roller bearings. Typical linear rolling element bearing configurations are shown in Figure 11.8.38. Note that, in general, it is easier to make a ball spherical than a roller cylindrical; hence ball bearings are more commonly used in precision machines than are roller bearings, the exception being when very high loads must be withstood. A good roller bearing may be better than a moderate ball bearing, and hence in the end the most important thing is to compare manufacturers' specifications.

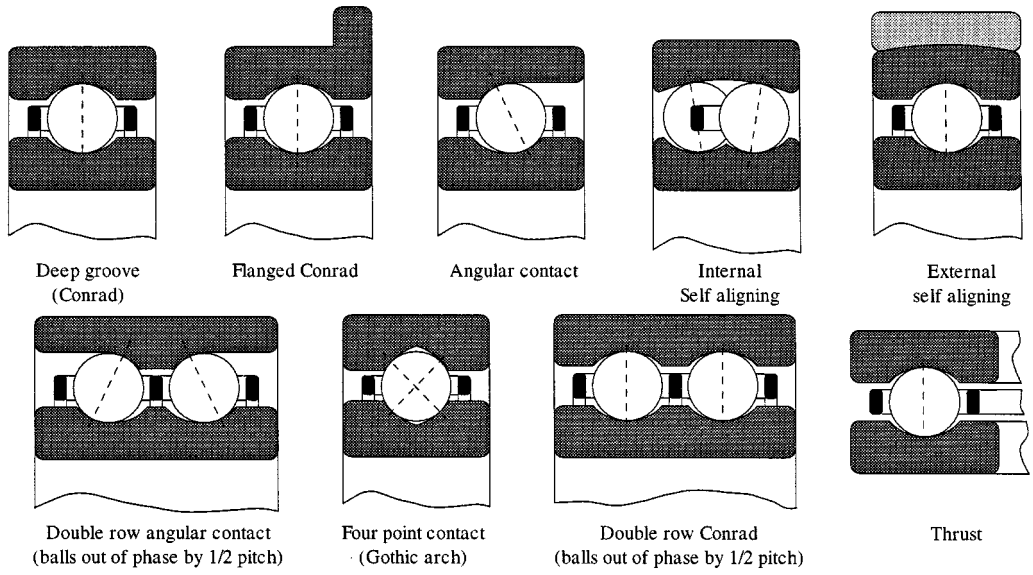


FIGURE 11.8.36 Typical ball bearing configurations for rotary motion bearings.

For linear motion applications, it is more difficult to maintain quality control of a curved surface a ball rides on than a planar surface a roller rides on because the latter can be self-checking. Note that machine-made linear bearing rails will have the same errors as the linear bearings on the machine that

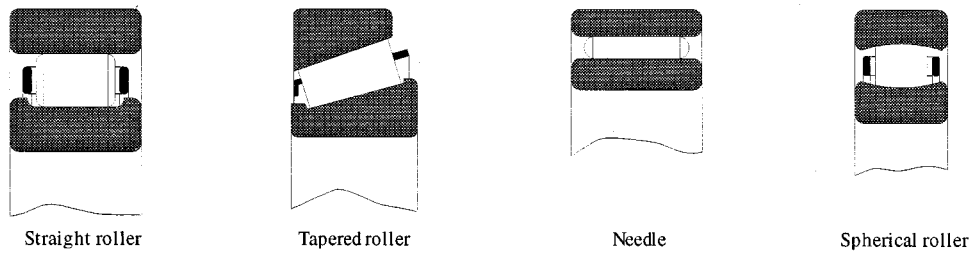


FIGURE 11.8.37 Typical roller bearing configurations for rotary motion bearings.

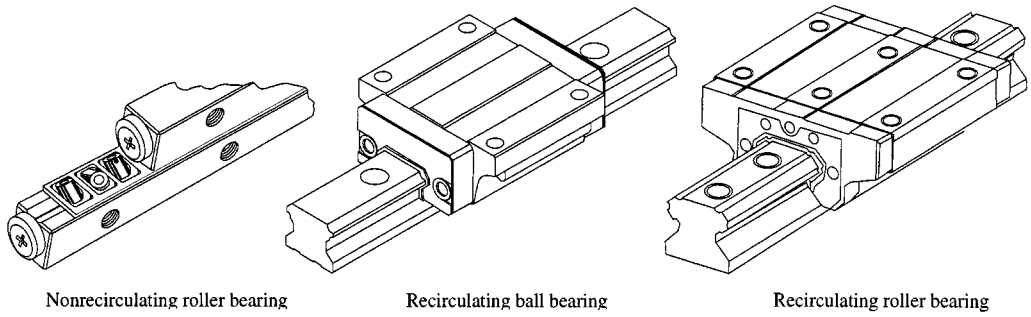


FIGURE 11.8.38 Typical linear rolling element bearing configurations.

was used to make them. For rotary axes, the raceway and the grinding tool can both be rotating at different speeds, which, combined with the random nature of precision spindle radial error motions, means that the form of the raceway can be uniform around is circumference. Thus rotary motion rolling element bearings can be made more accurate than linear motion rolling element bearings. Typical total radial error motion on a precision spindle is on the order of $^1/_4$ to 1 μm. For higher-performance linear or rotary motion bearings, one would typically move into the realm of aerostatic, hydrostatic, or magnetic bearings.

The design and production of a rolling element bearing requires careful analysis, materials selection, manufacturing quality control, and testing. Few companies other than bearing manufacturers have the resources for this type of effort. Whenever possible, one should use off-the-shelf bearing components. In addition, whenever possible, one should use modular components such as spindles and linear axes, particularly for nonsubmicron machines made in small lots (less than about 10 to 20 machines). The savings in design time, prototype testing, spare parts inventory, and repair and replacement costs often far outweigh the potential of saving a few dollars in manufacturing costs.

Designing with rolling element bearings can be intimidating because there are so many subtle details that can ruin a design if they are not considered. The best way to learn about how to handle these details is to think carefully about the physics of their design and the application and/or to work with others with experience, and if possible to experiment. In addition, manufacturers of precision bearing components are also usually willing to work with design engineers to integrate their bearing components into a design. Mapping and metrology frames can be used to increase accuracy given adequate repeatability, resolution, and controllability of the machine *if* the machine is designed with these error-reducing methods in mind.

Rolling Element Linear Motion Bearings

After the rotary motion rolling element ball bearing, the linear motion rolling element bearing ranks on the list of major inventions. Although not found in consumer products with anywhere near the frequency

of rotary ball bearings, linear rolling element bearings are vital for the manufacturing industry. They are critical elements in the design of today's high performance machine tools and factory automation systems.

Flexural Bearings*

Sliding, rolling, and fluid film bearings all rely on some form of mechanical or fluid contact to maintain the distance between two objects while allowing for relative motion between them. Since no surface is perfect and no fluid system is free from dynamic or thermal effects, all these bearings have an inherent fundamental limit to their performance. *Flexural bearings* (also called *fllexure pivots*), on the other hand, rely on the stretching of atomic bonds during elastic motion to attain smooth motion. Since there are millions of planes of atoms in a typical flexural bearing, an average effect is produced that allows flexural bearings to achieve atomically smooth motion. For example, flexural bearings allow the tip of a scanning tunneling microscope to scan the surface of a sample with subatomic resolution.** There are two categories of flexural bearings, monolithic and clamped-flat-spring, as shown in Figures 11.8.39 and 11.8.40.

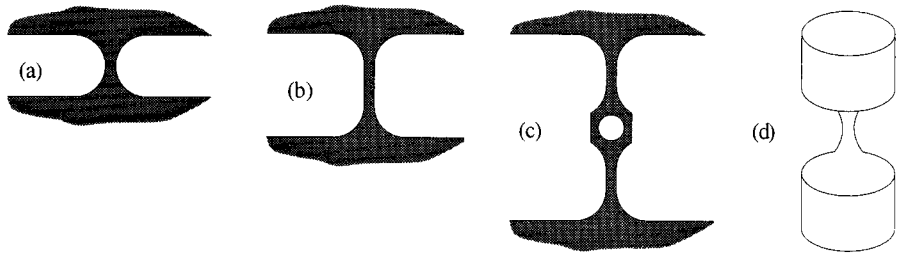


FIGURE 11.8.39 Monolithic flexural bearings.

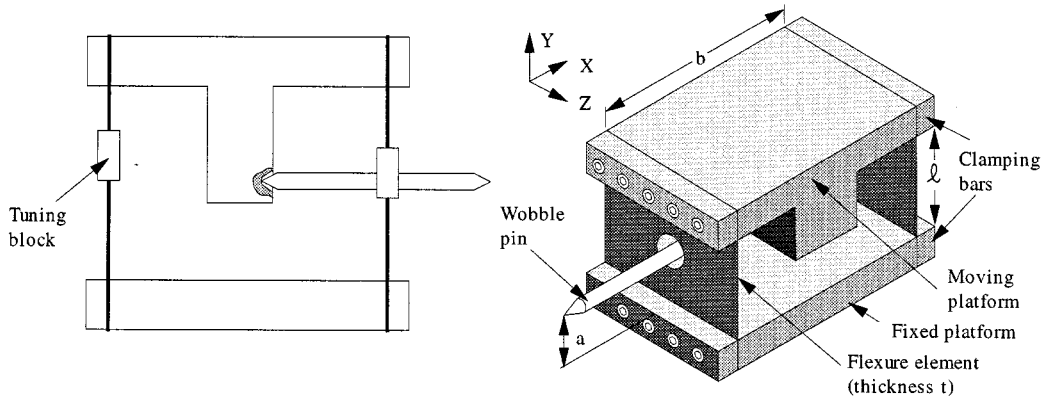


FIGURE 11.8.40 Clamped (blade) flexure.

* The reader may wish to consult the following references: Eastman, F.S. Nov. 1935. Flexure Pivots to Replace Knife Edges and Ball Bearings. *Univ. Wash. Eng. Exp. Sta. Bull.* No. 86; Eastman, F.S. Nov. 1937. The Design of Flexure Pivots. *J. Aerosp. Sci.* 5, 16–21; Jones, R.V. 1951. Parallel and Rectilinear Spring Movements. *J. Sci. Instrum.* 28, 38; Jones, R.V. and Young, I.R. 1956. Some Parasitic Deflections in Parallel Spring Movements. *J. Sci. Instrum.* 33, 11; Siddall, G.J. Sept. 1970. The Design and Performance of Flexure Pivots for Instruments. M.Sc. thesis, University of Aberdeen, Scotland, Department of Natural Philosophy.

** See, for example, Binnig, G. and Rohrer, H. 1982. Scanning Electron Microscopy, *Helv. Phys. Acta.* 55, 726–735.

Hydrostatic Bearings

Hydrostatic bearings utilize a thin film of high-pressure oil to support a load. In general, bearing gaps can be rather large, on the order of 5 to 100 μm. There are five basic types of hydrostatic bearings: single pad, opposed pad, journal, rotary thrust, and conical journal/thrust bearings. All operate on the principle of supporting a load on a thin film of high-pressure oil that flows continuously out of the bearing; hence a method is needed for supplying the pressurized oil and collecting and recirculating the oil that flows out of the bearing.

11.9 Robotics

Leonard D. Albano

The field of robotics is concerned with the design and development of mechanical devices that can be programmed to perform certain functions. Robots have been developed for a large variety of applications, such as material transport, automated assembly, and operations in controlled environments, such as high temperature and caustic surroundings. The semiconductor industry, for example, uses advanced robotic technology to fabricate IC chips. Many robots are simply relied on as a cost-effective alternative to human workers for certain highly repetitive tasks, such as spot welding and painting. For these applications, the robot must be able to undergo a range of accelerated motions while demonstrating the ability to position accurately its end effector. This must be done with minimum breakdowns and rapid repair. The control strategy involves solving the governing dynamic equations as the robot arm and end effector move through their operations. See Chapter 14 for further information.

11.10 Computer-Based Tools for Design Optimization

The range of computer applications in engineering design covers procedures from preliminary conceptual design to the production of manufacturing drawings and specifications (see Chapter 13). Most computer applications intended for production use can be classified into five or more major categories: analysis, computer-aided drafting and design, geometric modeling, data base management systems, and artificial intelligence. Traditional software for design optimization may be categorized as analytical applications, based on rational principles of mathematics and linear programming. Emerging computer-based tools for design optimization are an offshoot of research in artificial intelligence, capable of processing a variety of algorithmic, symbolic, deterministic, probabilistic, and fuzzy knowledge.

Design Optimization with Genetic Algorithms

Mark Jakiela

Introduction

One broad use of computer-based tools in design is to automate the design process itself. This is commonly done by modeling a design problem (or class of design problems) as a search problem. Such a model requires a representation and a search process. A representation is some set of parameters, either continuous or discrete, which can represent every possible design with appropriate values assigned to each parameter. We can call each possible parameter setting a design, and the set of all possible designs (possibly infinite in number) is called the design space. The search process is some method to investigate the design space and find some design that is acceptable or desired. When some acceptable designs are more desirable than others, the search process can be used to perform optimization.

Traditionally, the search and optimization technique that has been used most frequently for mechanical engineering problems has been a gradient-based search with continuous variables. When this representation and search method agrees with the problem being solved (the optimization model functions should be smooth and the objective preferably unimodal), this technique works exceedingly well. Currently, however, a variety of important problems are not amenable to this well-established technique because they cannot be modeled with only continuous variables (discrete or mixed discrete problems) and/or the objective and constraint functions do not exhibit appropriate characteristics. Other techniques are therefore required.

This subsection will provide an introduction to genetic algorithms ("GAs"), a search and optimization process that mimics the natural process of evolution. First, a brief tutorial on the fundamentals of GAs is given. This is followed by several examples drawn from a range of engineering problem domains. We hope that the variety of examples demonstrates the real strength and advantage of genetic algorithms: their versatility. They can be used to address almost any type of optimization problem. Requirements for their use are discussed in the tutorial. Some general and philosophical remarks conclude the subsection.

Genetic Algorithm Tutorial

Genetic algorithms are a simulation of a simplified process of evolution (Holland, 1975). The fundamental object of data used by a GA is referred to as a chromosome. Some number of chromosomes exist simultaneously in a population. The GA transforms one population into a succeeding population, creating a generation of children from a generation of parents. It does this with operators that are analogous to the biological operators of reproduction, recombination, and mutation. A measure of the quality of each chromosome, analogous to the fitness of a biological organism in an environment, is created and used to probabilistically select chromosomes that will serve as parents. With an increasing number of generations, the overall (e.g., average, single best) fitness of the population increases and the diversity represented in the population decreases.

For engineering purposes (see e.g., Goldberg, 1989), the chromosome is some type of encoding of the characteristics of a possible design. There is the notion, therefore, of a genotype (chromosome space) to phenotype (design space) mapping. Consider, as a simple example, a GA that is used to evolve optimal values of a set of parameters. A chromosome might be set up as a string of binary digits (i.e., bits), with concatenated substrings being binary encodings for each parameter. This is shown in Figure 11.10.1. Each bit position can be thought of as a gene and the value found there can be though of as an allele. The number of bits used to encode the parameter defines a resolution of the representation accuracy.

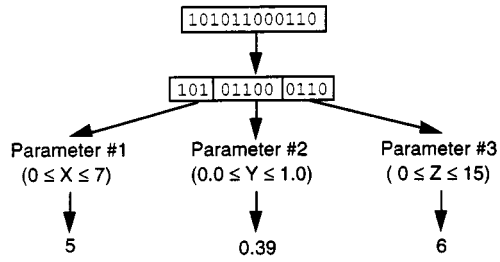


FIGURE 11.10.1 Binary chromosome representation.

Continuing with the explanatory parameter optimization example, the various processes of one iteration (i.e., generation) would proceed as follows. In a given generation, all chromosomes would be decoded to yield possible values for the parameters. These parameters would be used in an objective function of interest to yield a value measuring the quality of that set of parameters. This objective function value is used as the fitness measure of the chromosome. The probability of choosing a chromosome to serve as a parent is influenced by its fitness value: as in nature, the fittest are more likely to reproduce. Pairs of parents are randomly chosen with these weighting probabilities and mated. Copies of the parent chromosomes (i.e., reproduction) are used to build the chromosomes of the children. This is done by performing some type of recombination operation on them.

Figure 11.10.2 shows a simple one-point crossover recombination operation. First, a crossover point is randomly selected along the length of the chromosomes. Then, the front of one chromosome is appended to the back of the other chromosome and vice versa. This creates two children from two parents. Performing crossover with all the pairs of parents will create a tentative population of children that can replace the parents. The final step is to perform mutation on these children with a very low probability. One bit per 1000, for example, can be randomly chosen and inverted in value as an aid to maintaining diversity in the population. Finally, replacing the parents with the children brings the algorithm to the next generation.

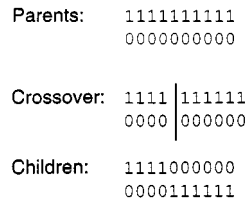


FIGURE 11.10.2 One-point crossover.

The procedure described above is commonly referred to as a “simple” GA (Goldberg, 1989) and can be thought of as a basic or canonical approach. In practice, several runtime parameters are required in its implementation. These include the probability of crossover (will two parents recombine, or will they proceed into the next generations unaltered?), the probability of mutation (on average, what fraction of genes are mutated?), a fitness scaling coefficient (an adjustment limiting the range of fitness values that

helps to prevent premature convergence), and population size (the number of chromosomes in a population). Note that a simple GA has the disadvantage of requiring a large number of function calls. This can be mitigated somewhat by using an "overlapping" population, in which a smaller number of parents will be chosen (i.e., fewer of the best) and the resulting children will replace the worst members of the parents' population. Regardless of the amount of overlap, note that a GA requires no derivative computations; the search is directed by the zeroth-order sampling of fitness function values. Because of this, a GA can perform well in ill-behaved multimodal search spaces (see Goldberg, 1989) and is relatively unlikely to become trapped in local suboptima. Equally important is the high degree of general flexibility of a chromosome representation. Although a binary encoding makes clear the general operation mechanisms and actually is used for a variety of problems, several other types of representations have been used and found to work well. In almost all cases, the evolutionary optimization strategy succeeds in producing improved designs. A variety of representations and fitness/objective functions will be described in the following examples.

Examples

Truss Parameter Selection. As a first example, we describe a problem that could readily be formulated as a continuous optimization problem (see Arora, 1989, pp. 23–31). The following genetic algorithm formulation, condensed from Wallace et al. (1995), will highlight the different approaches that are needed for a GA-based optimization.

Consider the symmetric truss shown in Figure 11.10.3. The truss is to be made of medium strength structural steel. We will assume that the load W and the truss height h are specified. The designer wishes to determine the truss base s and the inside and outside diameters of the members 1 and 2. A possible list of design variables and associated upper and lower limits for this problem are shown in Table 11.10.1. Note that the variables r_1 and r_2 allow the cross sections to vary continuously from solid rod stock to thin-walled pipe.

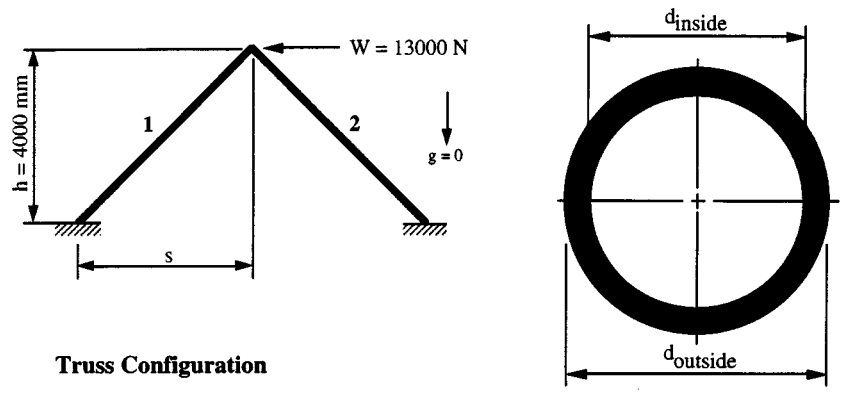


FIGURE 11.10.3 Definition of a simple two-member truss problem (gravity is assumed to be zero to simplify buckling calculations).

TABLE 11.10.1 Design Variables or Parameters to be Optimized in the Truss Problem

Design Variable	Limits
$d_{1(outside)}$	$0 < d_{1(outside)} < 100$ mm
$r_1 = d_{1(inside)}/d_{1(outside)}$	$0 < r_1 < 1$
$d_{2(outside)}$	$0 < d_{2(outside)} < 100$ mm
$r_2 = d_{2(inside)}/d_{2(outside)}$	$0 < r_2 < 1$
s	1000 mm < s < 3000 mm

In a continuous gradient-based formulation, the next step would be to model the relationships of interest as objective and constraint functions. Lagrange variables could then be introduced and the lagrangian would be subjected to the first- and second-order (constrained) optimality conditions to find a set of local optima. In addition to the upper and lower variable limits already listed, the other relationships of interest would include the normal stresses in members 1 and 2, the buckling stress in member 1, and the material cost as the objective function.

A genetic algorithm formulation does not use any derivatives. Instead, an objective function augmented with penalty terms for each potentially active constraint is used in an unconstrained optimization search. Wallace et al. (1995) formalize the notion of "design specifications" as a means to formulate such GA design optimization problems in a consistent and simplified way. The basic idea is that specifications indicate the probability that a particular performance level will be deemed acceptable. For the truss optimization the performance specifications considered are shown in Table 11.10.2. It is also possible to use other goal, objective, or preference formulations, such as the utility theory (Keeney and Raiffa, 1976).

The performance specifications on normal stress in the two members, $n_{\sigma 1}$ and $n_{\sigma 2}$, are modeled as safety factors. The specification distribution indicates that a safety factor for either member less than 1.5 is unacceptable (i.e., it will be accepted with a probability of zero, also see discussion below), and a safety factor greater than 2.0 is acceptable with certainty. To actually use this probabilistic specification, the design variables shown in Table 11.10.1 are decoded from a chromosome and used to compute values for $n_{\sigma 1}$ and $n_{\sigma 2}$, resulting in turn in two acceptance probabilities. A buckling safety factor for member 1 is treated in a similar manner (note that member 2 is always in tension). In addition to the upper and lower limits on r_1 and r_2, an additional performance specification is provided that indicates a strong preference for solid rod or thin-walled tube. This specification has, in effect, made the acceptable ranges of r_1 and r_2 discontinuous. Finally, the cost objective function is also provided to the problem as a probabilistic specification. If a full acceptance probability is achieved, the problem can be resolved with a lower material cost certainty point. In the solutions presented below, we assume that rod is 1/2 the price of pipe on a unit volume basis.

Once all individual acceptance probabilities are computed, a composite acceptance probability is used as the fitness function for the GA search. Intuitively, it seems as though the overall probability of acceptance is simply the product of the individual acceptance probabilities. Such a model does not work well in the GA-based search since a low score on a single characteristic will leave the overall design with a zero acceptance probability (in this case, the GA will consider "infeasible" designs with r_1 and/or r_2 between 0.2 and 0.7). The other characteristics of the designs could be very good and worthy of consideration in future generations. We therefore transform the multiplicative probability maximization into an additive information content minimization by taking the log of each probability term. The resulting objective is a generalized from of Suh's "Information Axiom" see Section 11.4.

$$I = \sum_{i=1}^{n} \log_2 \left(\frac{1}{p_i} \right) \tag{11.10.1}$$

where

n = number of design specifications or criteria
p_i = probability of being acceptable as defined by the i^{th} specification*
I = design information content in bits

Results for the solved truss problem are shown in Figure 11.10.4. Member 1 is under compression and thus is tubular (for buckling stiffness), while the tensile member 2 is made of less expensive rod stock. Convergence to a solution typically occurred in 80 generations requiring 2.5 sec on a Silicon Graphics

* For small values of p_i, a minimum nonzero value is used in Equation 11.10.1 to prevent division by zero.

TABLE 11.10.2 Performance Specifications and Distributions for the Truss Problem

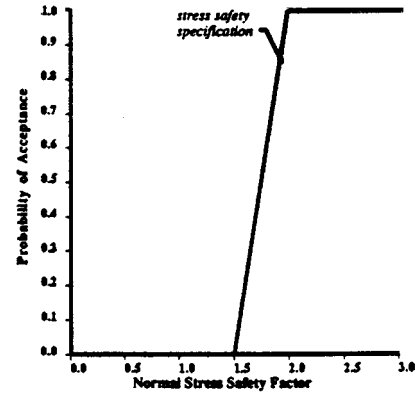
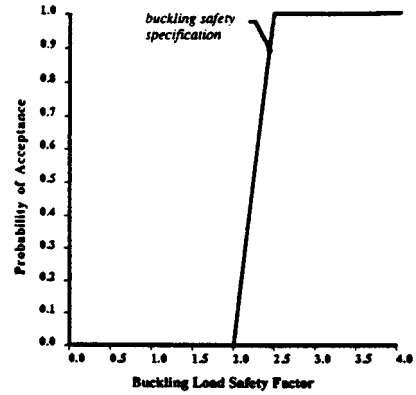
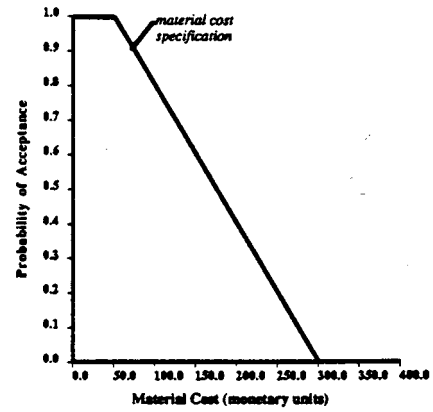
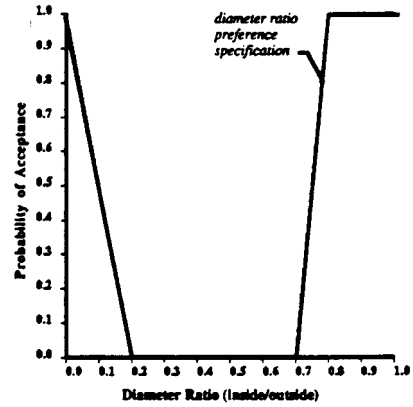
Performance Specifications	Specification Distribution
$n_{\sigma 1}$ (normal stress safety factor), member 1 $n_{\sigma 2}$ (normal stress safety factor), member 2	
$n_{\beta 1}$ (buckling safety), member 1	
C (cost)	

TABLE 11.10.2 Performance Specifications and Distributions for the Truss Problem (continued)

Performance Specifications	Specification Distribution
r_1 (diameter ratio), member 1	
r_2 (diameter ratio), member 2	

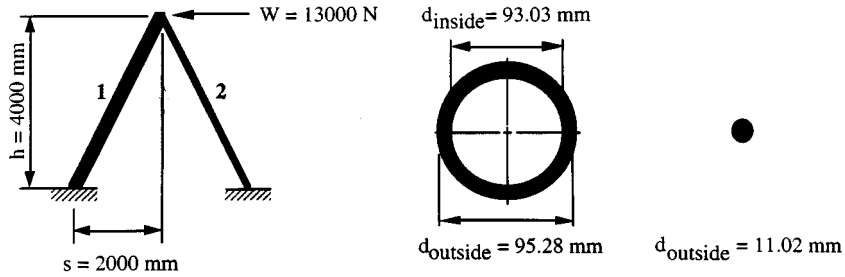


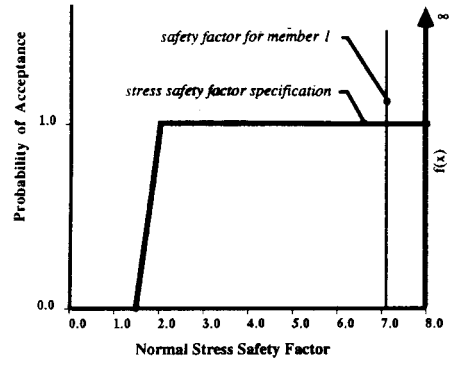
FIGURE 11.10.4 (a) Configuration of the optimized truss design.

Indigo[2] Extreme[2]. This search involved the evaluation of 2400 candidate designs (30 per generation). The search space for this problem contains $\approx 1.1 \times 10^{12}$ points (five parameters at 8-bit resolution gives 2^{40} points).

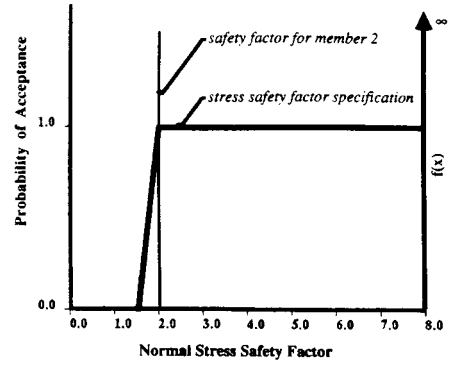
It must be noted that the search converged to this solution using rod stock for member 2 in only 17% of 60 consecutive optimizations. The more frequent solution shown in Figure 11.10.5 is nearly as good.

Topology Optimization. In this example, we will address planar stress/strain problems and the generation of optimal structural topologies. As shown in Figure 11.10.6, a design domain defines the allowable extents of any possible design. This domain is discretized into a rectangular grid, with each square element of the grid assigned a value of "material" or "void". Treating each grid element as a binary variable leads to a large combinatorial search space. The grid is used to create a finite element mesh for structural analysis. In our case, each material element contains four triangular finite elements made from the five nodes located at the four corners and the center of each material element. The genotype corresponding to this phenotype is a simple binary two-dimensional array chromosome, or a binary string, as described below.

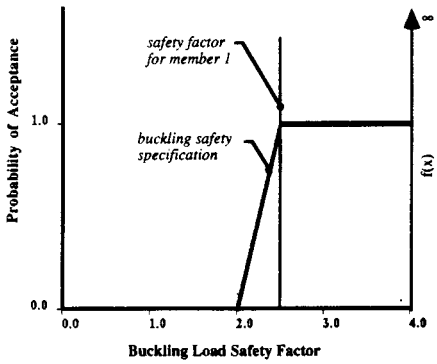
Phenotypically, certain material elements are required to always contain material, regardless of the bit value in the genome. We refer to these as material constraints. These are elements that serve as structural boundary conditions. In Figure 11.10.6, examples would be the highest and lowest elements bordering the wall, and the element that is used to apply the load. Thus, the genotype is not always a completely accurate representation of the actual structure.



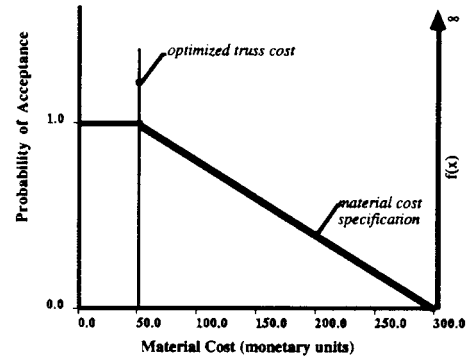
$n_{\sigma 1}$ = 7.09, $p_{acceptable}$= 1.0



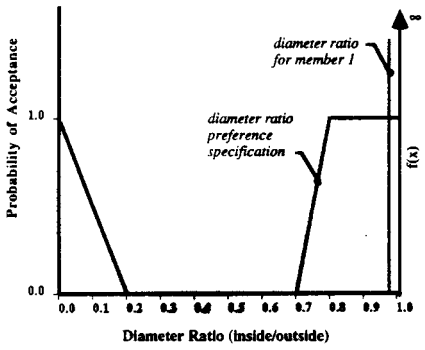
$n_{\sigma 2}$ = 2.03, $p_{acceptable}$= 1.0



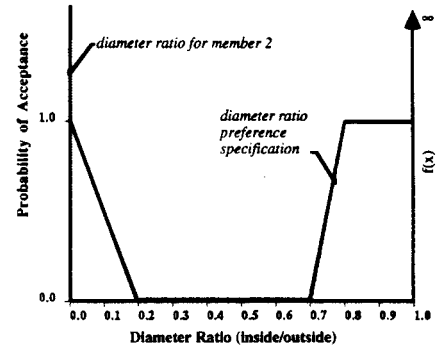
n_{B1} = 2.50, $p_{acceptable}$= 1.0



C = 51.02 units, p = .9958



r_1 = 0.9764, $p_{acceptable}$ = 1.0



r_2 = 0.0, $p_{acceptable}$ = 1.0

FIGURE 11.10.4 (b) Comparison of truss performance variables to design specifications.

The genotype can also differ from the phenotype because of connectivity analysis. Connectivity analysis requires that all elements that are included in the structural analysis be connected to other elements by at least one edge–edge connection, as opposed to only a corner connection. In addition, all elements considered connected must be linked to one of the material constraint elements by a path of edge connections. Any elements not so connected are removed from the mesh for the purposes of structural analysis: they are considered to have no stiffness and no mass. Since we are addressing planar problems, the rationale for connectivity analysis is that elements connected at corners cannot transfer

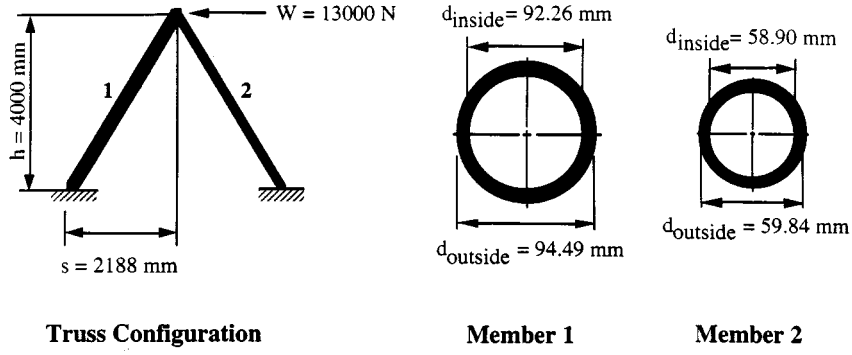


FIGURE 11.10.5 A result that did not find the better solution using cheaper rod stock for member 2. The cost is 56.73 units, giving $p_{acceptable}$ = 0.9728 (compared to 0.9958 for the solution using rod for member 2). $n_{\sigma 1}$ = 7.49, $n_{\sigma 2}$ = 2.01, $n_{\beta 1}$ = 2.50.

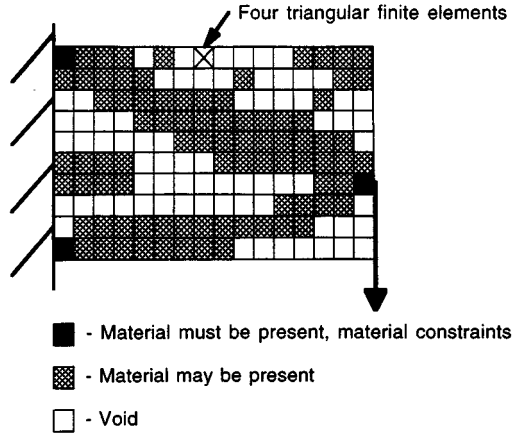


FIGURE 11.10.6 Typical design domain.

moments about those corners (since the finite element model treats corners as kinematic pin joints). It is less likely, therefore, that "disconnected" elements will contribute to enhanced structural performance. We have found that this is true empirically (see Chapman et al., 1993 for further discussion of this topic). Figure 11.10.7 shows the structure of Figure 11.10.6 after connectivity analysis. It is important to note that we do not penalize structures with more disconnected elements; we simply reward structures based upon the performance caused by their connected elements. This allows disconnected elements to in some sense be "recessive", in that crossover could cause them to combine with the (possibly disconnected) elements from another structure to yield a structure of connected elements of much improved performance.

As fitness, we are interested primarily in structural performance. In this study, this will be represented by two characteristics, mass and deflection. Generally, we wish the genetic algorithm to evolve structures that are both lightweight and deflect small amounts under a given load. The deflection, δ, will be measured with a finite element simulation and the mass m will be proportional to the number of connected material elements. These two characteristics can be used to create an unconstrained fitness, such as maximizing the following ratio:

$$\text{fitness} = \frac{1}{(\delta)(m)} \tag{11.10.2}$$

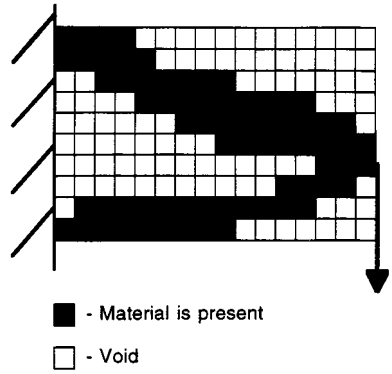


FIGURE 11.10.7 After connectivity analysis.

Alternatively, one of these characteristics can be optimized while treating the other as a constraint. Chapman and Jakiela (1995) provide some guidelines for using penalty functions for structural optimization that address many different characteristics (e.g., reducing the number of macroscopic holes in the topology).

In this example, we will use two different types of chromosomes to represent the material array. The first is a one-dimensional string chromosome made by concatenating the rows of the material array. The second is a two-dimensional binary array precisely matching the material array. For a two-dimensional array chromosome, a single-point crossover can be done as shown in Figure 11.10.8. A position is randomly selected along both dimensions, defining a point in the array. Complementary parts from both parents, diagonal about the point, contribute to the chromosomes of the children. Note that this approach can be extended to any dimension.

Parents:	1111111111 1111111111 1111111111 1111111111	0000000000 0000000000 0000000000 0000000000
Crossover:	111111 | 1111 111111 | 1111 111111 | 1111 111111 | 1111	000000 | 0000 000000 | 0000 000000 | 0000 000000 | 0000
Children:	1111110000 1111110000 0000001111 0000001111	0000001111 0000001111 1111110000 1111110000

FIGURE 11.10.8 One-point crossover on 2D arrays.

Using the binary string chromosome, Figure 11.10.9 shows results for three different material discretizations. The loading and material constraints were as described in Figure 11.10.6. All three cases display a uniformly distributed porosity with an overall shape arising implicitly from the topological optimization.

Hierarchial Shape Packing. Effective material utilization is important to virtually all industries. A specific problem in this area that can be modeled as a combinatorial optimization problem is the arrangement of planar shapes to be cut out from a piece of material so as to minimize scrap. This problem has great economic significance in the garment, sheet metal, shipbuilding, and other industries. A common instance involves a “blank” of stock material of fixed width, such as would result from a roll of material.

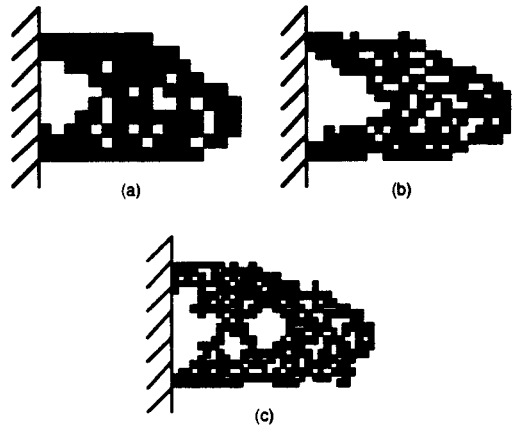


FIGURE 11.10.9 Results of (a) 10 × 16, (b) 15 × 24, and (c) 20 × 32 optimizations.

The part shapes must be arranged to minimize the length of material required from the roll. We have solved this problem with a two-level hierarchical genetic algorithm. A more complete description of this approach can be found in Dighe and Jakiela (1995a).

The approach of the lower level is shown in Figure 11.10.10. The blank of fixed width is considered analogous to a box into which the shapes will be packed. One by one, in a particular order, the shapes will be placed into the box so as to minimize the accrued height required. To pack a particular shape, a location and orientation is specified, as shown in Figure 11.10.10b, and the shape is "dropped" straight down until it contacts one of the previously packed shapes or the bottom of the box. The "dropping into a box" metaphor does not extend to include dynamic effects. Note, for example, that part D in Figure 11.10.10c would not rotate and slide to settle in a lower final position; it is in its final location for the values of x and θ chosen. Figure 11.10.10d shows a much better choice of parameters for x and θ. The lower level uses a genetic algorithm to optimize the values of x and θ and therefore is used each time a shape is packed.

The overall success of this packing strategy depends on the order in which the objects are placed. The higher-level GA searches the space of possible orders to find those that work well with the packing strategy of the lower-level GA. The chromosome representation for a packing order, and associated crossover and mutation operators are shown in Figure 11.10.11. The chromosome is simply a list of part labels, with the left-to-right order indicating the packing order. If we try to perform a single point crossover on these strings, Figure 11.10.11 shows that invalid offspring often result. Some labels are missing and others appear more than once. To remedy this problem, we use a simple order-based crossover as shown in Figure 11.10.11b. For positions to the left of the crossover point, each child receives the order from one of the parents. The order of the remaining parts is obtained from the order in which those remaining parts appear in the other parent's chromosome. Finally, mutation simply changes the location of a part label, usually by moving it to the left.

Note that this crossover operator tends to not disrupt the leftmost parts of the chromosomes. In particular, without mutation, the very leftmost position in any chromosome must have a label that was present in that position in the initial population. Hence, mutation is performed with a leftward bias.

Figure 11.10.12 shows some typical arrangements produced using this approach. Figures 11.10.12a and 11.10.12b show two configurations in the same optimization run. These ten objects were chosen such that their characteristic sizes varied by about a factor of 10. Figure 11.10.12 shows a jigsaw puzzle that we have used as a benchmark test for the packing system. Taking into account deliberate gaps between the parts, the density of the correct jigsaw arrangement is 0.92, not 1.00. Two arrangements from the same optimization run are shown in Figures 11.10.12d and 11.10.12e. It is clear that the simple "height-based" lower-level GA cannot match a human's visual processing and reasoning ability.

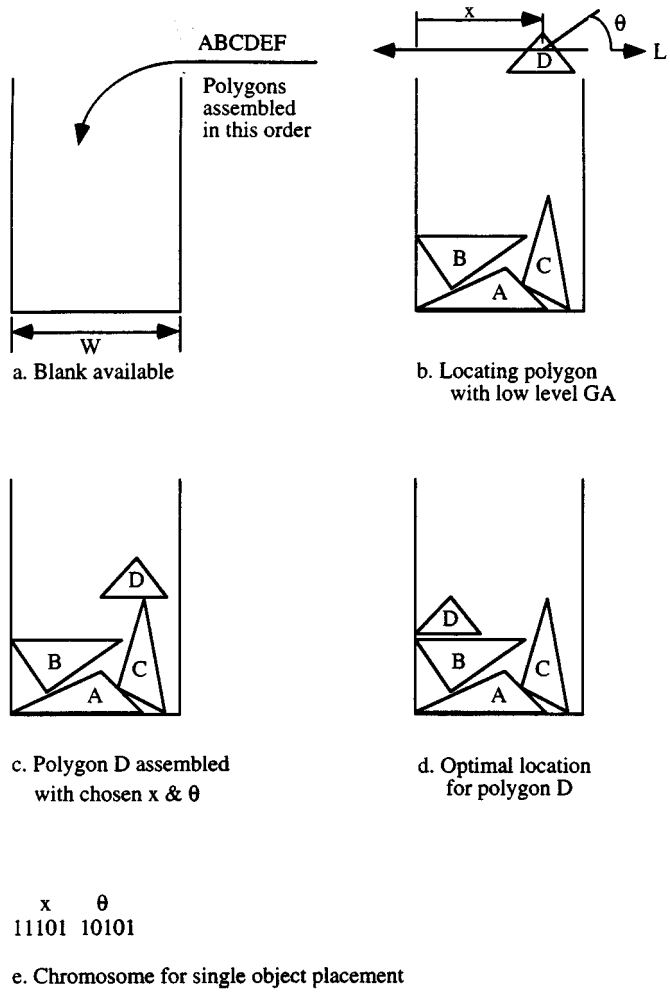


FIGURE 11.10.10 Lower-level GA for hierarchical shape packing.

Integrating a GA-based search with heuristics derived from observation of human experts is the subject of ongoing work in this area (see Dighe and Jakiela, 1995b).

References

Arora, J.S. 1989. *Introduction to Optimum Design.* McGraw-Hill, New York.

Chapman, C. and Jakiela, M. 1995. Genetic algorithm-based structural topology design with compliance and topology simplification considerations. *ASME J. Mech. Design.* to appear.

Chapman, C., Saitou, K., and Jakiela, M. 1993. Genetic algorithms as an approach to configuration and topology design. In *Proceedings of the ASME 19th Design Automation Conference: Advances in Design Automation,* Vol. 1. American Society of Mechanical Engineers, DE-Volume 65-1, New York, 485–498.

Dighe, R. and Jakiela, M.J. 1995a. Solving pattern nesting problems with genetic algorithms employing task decomposition and contact detection. In *Evolutionary Computation.* MIT Press. to appear.

Dighe, R. and Jakiela, M.J. 1995b. Automation of human patern nesting strategies. *ASME J. Mechan. Design.* submitted.

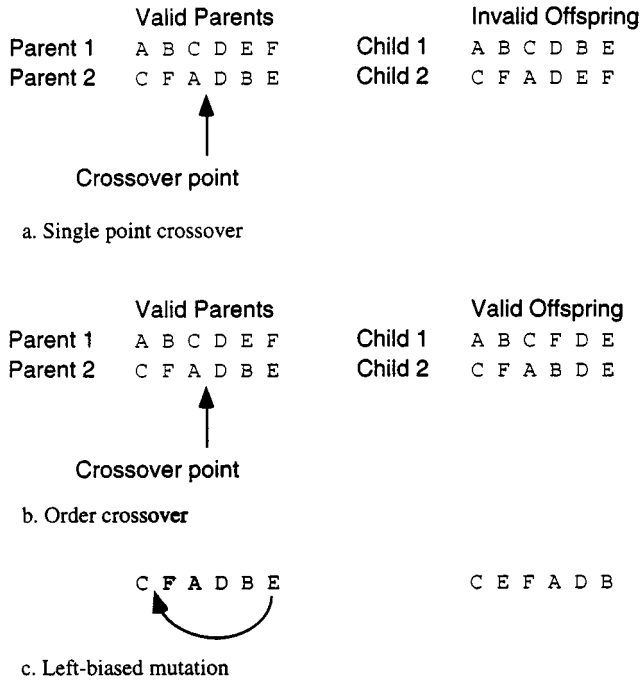


FIGURE 11.10.11 Crossover and mutation operators.

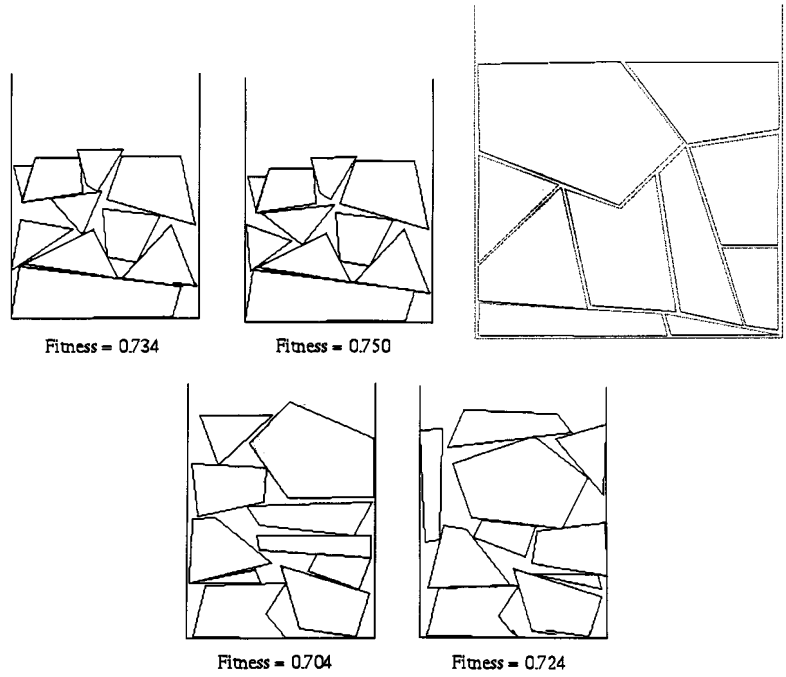


FIGURE 11.10.12 (a) Sample result from a GA run; (b) sample result from a GA run; (c) jigsaw puzzle attempted with sting-based representation; (d) layout of jigsaw puzzle using string-based representation and height-based fitness function; (e) layout of jigsaw puzzle using string-based representation and height-based fitness function.

Goldberg, D. 1989. *Genetic Algorithms in Search, Optimization, and Machine Learning.* Addison-Wesley, Reading, MA.

Holland, J. 1975. *Adaptation in Natural and Artificial Systems.* University of Michigan Press, Ann Arbor, MI.

Keeney, R.L. and Raiffa, H. 1976. *Decisions with Multiple Objectives: Preferences and Value Tradeoffs.* John Wiley & Sons, New York.

Suh, N.P. 1990. *The Principles of Design.* Oxford University Press, New York.

Wallace, D.R., Jakiela, M.J., and Flowers, W.C. 1995. Design search under probabilistic specifications using genetic algorithms. In *Computer-Aided Design.* Butterworth-Heinemann, to appear.

Optimization in Multidisciplinary Design

Kemper Lewis, Farrokh Mistree, and J. R. Jagannatha Rao

Introduction

The design of complex systems is a difficult task of integrating disciplines, each with their own analysis, synthesis, and decision process. Optimizing such a system on a global scale is realistically impossible, but finding a solution that is "good enough" and robust is achievable. Several approaches to formulating and solving a multidisciplinary design problem have arisen in a rather ad hoc fashion since the inception of multidisciplinary design optimization (MDO). These approaches include single-level and multilevel formulations, hierarchical and nonhierarchical system decomposition methods, and numerous optimization and analysis processes and approaches at the system and subsystem levels. Designers are referred to as decision makes and objectives as goals or rewards. With only one decision maker, the problem becomes a scalar or vector optimization problem. But in MDO, many decision makers may exist, and each decision maker's strategy to optimize his/her rewards could depend on the strategies and decisions of other decision makers. The modeling of strategic behavior based on the actions of other individuals is known as a *game*. Therefore, the focus in this section is on problems characterized by

- Single decision makers who have multiple rewards
- Multiple decision makers who have multiple rewards

This focus in optimization theory is shown as the shaded region in Figure 11.10.13. Typical courses in optimization focus on the upper-left quadrant, namely, scalar optimization problems with one objective and one decision maker. In this section, the focus is on the other three quadrants as a means to expand the application of optimization theory to problems that frequently occur in engineering design.

Classification of MDO Formulations

Simultaneous and Nested Analysis and Design. With the advent of MDO and its various applications, a structure of problem approaches and formulations is necessary. In Cramer et al. (1994) and, more recently, Balling and Sobieski (1994) and Lewis and Mistree (1995), various classes and classifications of problem formulations are presented. Figure 11.10.14 is a generic representation of a coupled, three-discipline system.

Depending on the level of analysis, the modules in Figure 11.10.14 may refer to disciplines, components, or processes. It is the decomposition of the system, subsystem coupling and solution, and system synthesis that pose major research and application problems in MDO. The terms used in the figure, as well as other common terms, and are defined below.

s_1, s_2, s_3: disciplinary *state variables* which comprise the *state equations*
y_1, y_2, y_3: the *state equations*
r_1, r_2, r_3: *residuals* in the state equations
y_{12}, y_{13}, y_{21}, y_{23}, y_{31}, y_{32}: *coupling functions*, y_{ij} contains those functions computed in discipline i which are needed in discipline j

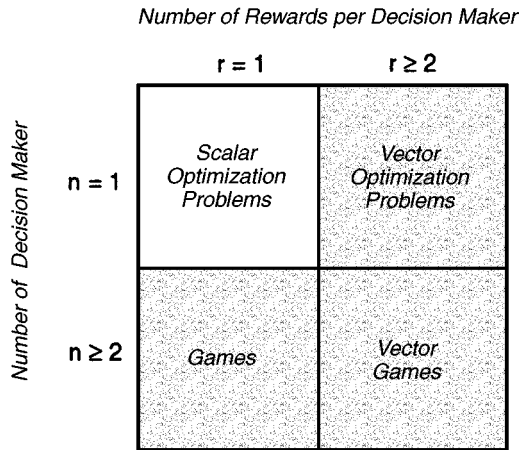


FIGURE 11.10.13 Various formulations in optimization theory.
t

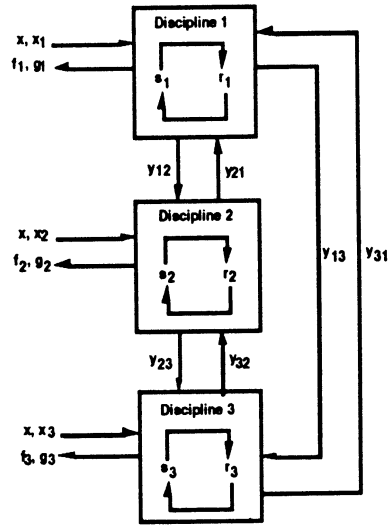


FIGURE 11.10.14 A three-discipline coupled system. (From Balling, R.J. and Sobieski, J. 1994. *5th AIAA/USAF/NASA/ISS/MD Symp. on Recent Advances in Multidisciplinary Analysis and Optimization.* Panama City, FL. 753–773.)

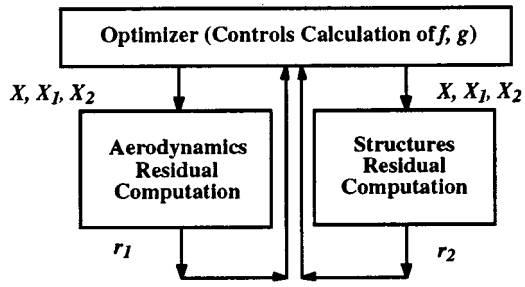


FIGURE 11.10.15 Single-SAND-SAND formulation.

$y^*_{12}, y^*_{13}, y^*_{21}, y^*_{23}, y^*_{31}, y^*_{32}$: *coupling variables*
x: *system design variables* needed by more than one discipline
x_1, x_2, x_3: *disciplinary design variables*
g_1, g_2, g_3: *design constraint functions*
f_1, f_2, f_3: *design objective functions*

The primary task at hand is summarized as follows: *Determine the values of the design, state, and coupling variables which satisfy the state equations, the coupling equalities, the design constraints, and the design objective functions.*

Based on this, six classifications for fundamental approaches to MDO problem formulation and solution are presented by Balling and Sobieski, which depend on three criteria:

1. System vs. multilevel decomposition
2. Simultaneous (SAND) vs. nested analysis and design (NAND) at the *system* level.
3. Simultaneous (SAND) vs. nested analysis and design (NAND) at the *subsystem* or *discipline* level.

At the discipline level, SAND implies that the disciplinary design and state variables are determined simultaneously by the optimizer, while NAND implies that the optimizer determines only the disciplinary design variables and requires determination of the state variables at each iteration. At the system level, SAND implies that the system design variables and coupling variables are determined simultaneously by the system optimizer, while NAND implies that the system optimizer determines only the system design variables and requires calls to a system analysis routine to determine the coupling variables at each iteration. The "optimizers" at the system level or discipline level could be gradient based or heuristic in nature, depending on the problem formulation. Further classifications can be generated if these approaches are combined or linked sequentially within one design problem.

Each approach has a three-part name consisting of the overall decomposition descriptor, the solution approach at the system level, and the solution approach at the subsystem level. The first part indicates whether the approach is a single-level or multilevel approach. The middle and last parts of the name indicate whether the SAND or NAND approach is used at the system and discipline levels, respectively.

Example

These terms are illustrated based on a simple example from aeroelastic MDO problem in Cramer et al. (1994). This aeroelastic problem involves two disciplines, aerodynamics and structures. This problem is extremely difficult, as the solution of either of these disciplines *alone* involves heavy analysis and computation. For simplicity, it is assumed that a *single-level* scheme is to be used. The first formulation choice is SAND or NAND at the system level. If a SAND formulation is used, the discipline-level formulation must be determined. These formulations are illustrated below.

Single-SAND-SAND. In this formulation, feasibility is *not* sought for the analysis problem in any sense until convergence in the solver is reached. The analysis "codes" for aerodynamics and structures in this formulation perform a simple function; they evaluate the *residuals* of the analysis equations, rather than solving some set of equations. This is illustrated in Figure 11.10.15. The *optimizer* controls the calculation of the objective functions, f, and the constraints, g, based on the residuals, r_i, from the disciplines. The optimizer sends system and disciplinary design variables, X and X_i, to the disciplinary analysis routines.

Single-SAND-NAND. In this formulation, feasibility *is* enforced for each individual discipline, while the optimizer drives the individual disciplines toward multidisciplinary feasibility and optimality by controlling the coupling functions between the disciplines. This is illustrated in Figure 11.10.16. The optimizer sends design and coupling variables, X and y^*_{ij}, to the disciplinary analysis routines, while these routines return coupling functions, y_{ij}, which are functions of the design variables and the disciplinary state variables, y_i.

Single-NAND-NAND. In this formulation, multidisciplinary feasibility is required at each iteration of the optimizer. Therefore, an aeroelastic analysis solver combines the information from the aerodynamics

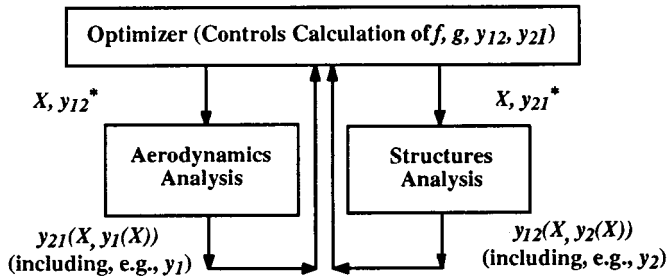


FIGURE 11.10.16 Single-SAND-NAND formulation.

and structures disciplines and ensures that the two disciplines are feasible concurrently. This is illustrated in Figure 11.10.17. The optimizer only controls the objective functions and constraints, while the coupling functions are handled by the aeroelastic analysis solver which becomes a "black box" from the perspective of the optimization code. In each discipline, an analysis code solves a set of equations. The resulting *state equations* are fed into the other disciplines. This process continues until convergence among the disciplines, and the optimizer is once again invoked.

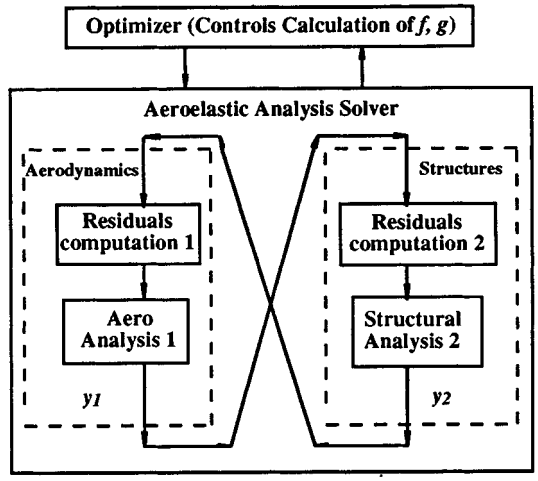


FIGURE 11.10.17 Single-NAND-NAND formulation.

Modeling Design Decisions: The Mathematical Programming Background

Decision-based design (DBD) is offered as a starting point for the creation of design methods that are based on the notion that the principal role of an engineer in the design of a product or process is to make decisions. Therefore, the process of design in its most basic sense is a series of decisions.

It is recognized that the implementation of DBD can take many forms; one implementation is the decision support problem (DSP) technique. Within the DSP technique there exist three types of decisions a designer can make: selection, compromise, or heuristic. In this section, the focus is on the compromise DSP, which is used to model multiobjective trade-offs in the solution of a mathematical model.

Compromise DSP. A compromise DSP is a hybrid formulation that incorporates concepts from both traditional mathematical programming and goal programming, and makes use of some new ones (Mistree et al., 1993). It is similar to goal programming in that the multiple objectives are formulated as system goals (involving both system and deviation variables) and the deviation function is solely a function of the goal deviation variables. This is in contrast to traditional mathematical programming where multiple objectives are modeled as a weighted function of the system variables only. The concept of system

constraints, however, is retained from the traditional constrained optimization formulation. Special emphasis is placed on the bounds on the system variables, unlike in traditional mathematical programming and goal programming. In effect the traditional formulation is a subset of the compromise DSP — an indication of the generality of the compromise formulation. The compromise DSP is stated in words as follows:

Given

An alternative that is to be improved through modification
Assumptions used to model the domain of interest
The system parameters
All other relevant information

n	Number of system variables
p + q	Number of system constraints
p	Equality constraints
q	Inequality constraints
m	Number of system goals
$g_i(X)$	System constraint function

$$g_i(\underline{X}) = C_i(\underline{X}) - D_i(\underline{X})$$

$f_k(d_i)$	Function of deviation variables to be minimized at priority level k for the preemptive case
W_i	Weight for the Archimedean case

Find

The values of the independent *system variables* (they describe the physical attributes of an artifact)

$$X_j \quad j = 1,\ldots, n$$

The values of the *deviation variables* (they indicate the extent to which the goals are achieved)

$$d_i^-, d_i^+ \quad i = 1,\ldots, m$$

Satisfy

The *system constraints* that must be satisfied for the solution to be feasible, there is no restriction placed on linearity or convexity

$$g_i(\underline{X}) = 0; \quad i = 1,\ldots, p$$

$$g_i(\underline{X}) \geq 0; \quad i = p+1,\ldots, p+q$$

The *system goals* that must achieve a specified target value as far as possible; there is no restriction placed on linearity or convexity

$$A_i(\underline{X}) + d_i^- - d_i^+ = G_i; \quad i = 1,\ldots, m$$

The lower and upper *bounds* on the system

$$X_j^{\min} \leq X_j \leq X_j^{\max}; \quad j = 1,\ldots, n$$

$$d_i^-, d_i \geq 0 \quad \text{and} \quad d_i^- \cdot d_i^+ = 0$$

Minimize

The *deviation function* which is a measure of the deviation of the system performance from that implied by the set of goals and their associated priority levels or relative weights:

Case a: Preemptive (lexicographic minimum)

$$Z = \left[f_1\left(d_i^-, d_i^+\right), \ldots, f_k\left(d_i^-, d_i^+\right) \right]$$

Case b: Archimedean

$$Z = \sum_{i=1}^{m} W_i\left(d_i^- + d_i^+\right); \qquad \sum W_i = 1; \quad W_i \geq 0$$

A graphical representation of a two-variable compromise DSP is shown in Figure 11.10.18. The difference between a system variable and a deviation variable is that the former represents a distance in the i^{th} dimension from the origin of the design space, whereas the latter has as its origin the surface of the system goal. The value of the i^{th} deviation variable is determined by the degree to which the i^{th} goal is achieved. It depends upon the value of $A_i(X)$ alone (Since G_i is fixed by the designer), which in turn is dependent upon the system variables X. The set of discrete variables can be continuous, Boolean, or a combination of types.

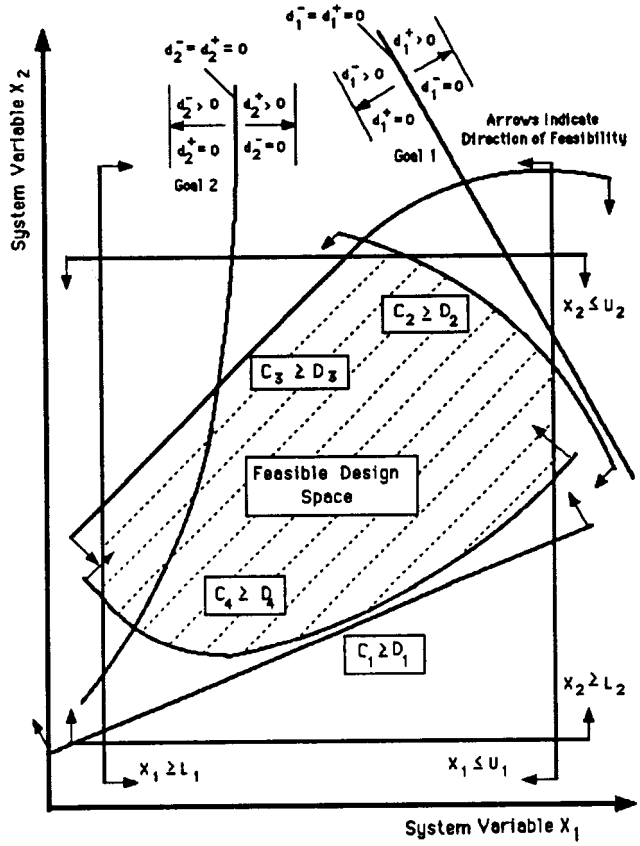


FIGURE 11.10.18 The compromise DSP.

The compromise DSP is fundamentally different from the traditional mathematical optimization model. This difference is depicted in Figure 11.10.19 where the optimization model on the left occurs when designer's aspirations can be met by the system achievement and therefore an optimum can be found. But when a designer's aspirations do not overlap with the actual achievements, a solution which is "good enough" or *satisficing*[*] must be found. This is depicted and modeled in the compromise DSP on the right of Figure 11.10.19. When the aspiration space overlaps with the feasible design space (center of Figure 11.10.19), the optimization model and compromise DSP are the same.

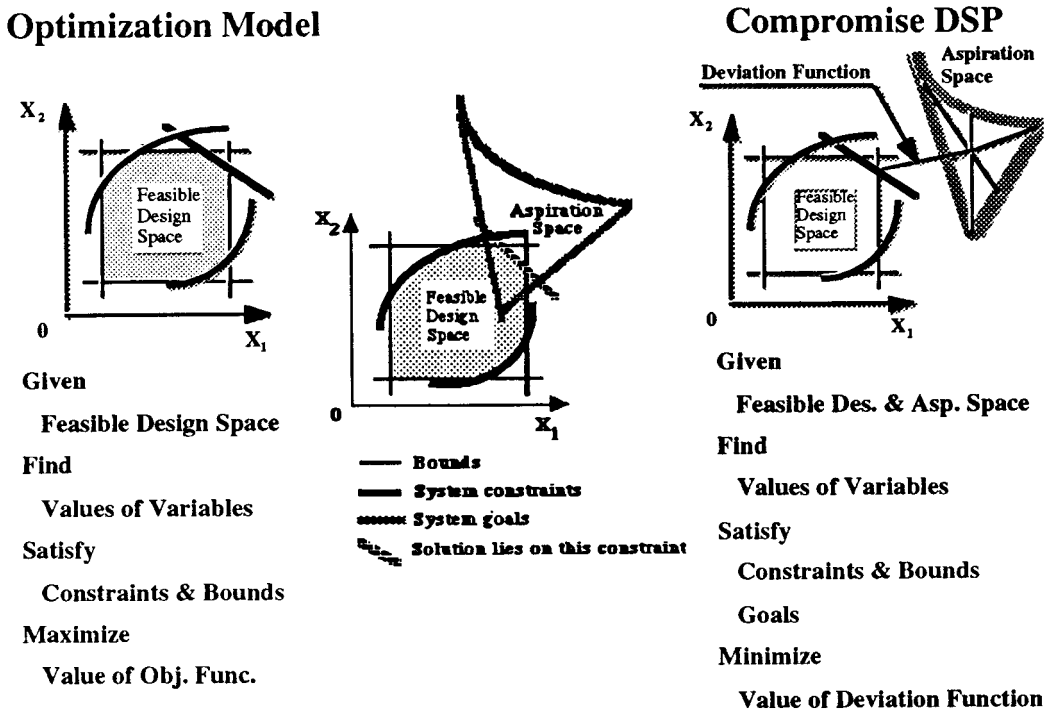


FIGURE 11.10.19 The optimizing and satisficing models.

Deviation Variables and Goals. In a **goal** one can distinguish the aspiration level, G_i, of the decision maker and the actual attainment, $A_i(X)$, of the goal. Three conditions need to be considered:

1. $A_i(X) \le G_i$; one wishes to achieve a value of $A_i(X)$ that is equal to or less than G_i.
2. $A_i(X) \ge G_i$; one wishes to achieve a value of $A_i(X)$ that is equal to or greater than G_i.
3. $A_i(X) = G_i$; one wishes to achieve a value of $A_i(X)$ equal to G_i.

Next, the concept of a **deviation variable** is introduced. Consider the third condition, namely, one would like the value of $A_i(X)$ to equal G_i. The deviation variable is defined as

$$d = G_i - A_i(\underline{X})$$

The deviation variable d can be negative or positive. In effect, a deviation variable represents the distance (deviation) between the aspiration level and the actual attainment of the goal. Considerable simplification

[*] Satisficing — not the "best", but "good enough" (the first use of this term, in the context of optimization, is attributed to Herbert Simon (Simon, 1982).

of the solution algorithm is affected if one can assert that all the variables in the problem being solved are positive. Hence, the deviation variable d is replaced by two variables:

$$d = d_i^- - d_i^+$$

where

$$d_i^- \cdot d_i^+ = 0 \quad \text{and} \quad d_i^-, d_i^+ \geq 0$$

The preceding ensures that the deviation variables never take on negative values. The product constraint ensures that one of the deviation variables will always be zero. The system goal becomes

$$A_i(\underline{X}) + d_i^- - d_i^+ = G_i; \quad i = 1, 2, \ldots, m \tag{11.10.3}$$

subject to

$$d_i^-, d_i^+ \geq 0 \quad \text{and} \quad d_i^- \cdot d_i^+ = 0 \tag{11.10.4}$$

If the problem is solved using an algorithm that provides a vertex solution as a matter of course, then the constraint is automatically satisfied making its inclusion in the formulation redundant. Since some algorithms use solution schemes which provide a vertex solution, it is assumed that this constraint is satisfied. For completeness this constraint is included in the mathematical forms of the compromise decision support problem given previously in this chapter and for brevity will be omitted from all subsequent formulations.

Note that a goal (Equation 11.10.3) is always expressed as an equality. When considering Equation (11.10.3), the following will be true:

if $A_i \leq G_i$ **then** $d_i^- > 0$ and $d_i^+ = 0$.
if $A_i \geq G_i$ **then** $d_i^- = 0$ and $d_i^+ \geq 0$.
if $A_i = G_i$ **then** $d_i^- = 0$ and $d_i^+ = 0$.

How are the three conditions listed using Equation (11.10.3)?

1. To satisfy $A_i(X) \leq G_i$, the positive deviation d_i^+ must be equal to zero. The negative deviation d_i^- will measure how far the performance of the actual design is from the goal.
2. To satisfy $A_i(X) \geq G_i$, the negative deviation d_i^- must be equal to zero. In this case, the degree of overachievement is indicated by the positive deviation d_i^+.
3. To satisfy $A_i(X) = G_i$, both deviations, d_i^- and d_i^+, must be zero.

At this point it is established what is to be minimized. In the next section means for minimization of the objective in goal programming are introduced.

The Lexicographic Minimum and the Deviation Function. The objective of a traditional single objective optimization problem requires the maximization or minimization of an objective function. The objective is a function of the problem variables. In a goal programming formulation, each of the objectives is converted into a goal (Equation 11.10.3) with its corresponding deviation variables. The resulting formulation is similar to a single objective optimization problem but with the following differences:

The objective is always to minimize a function.
The objective function is expressed using deviation variables only.

The objective in the goal programming formulation is called the achievement function. As indicated earlier the deviation variables are associated with goals and hence their range of values depend on the

goal itself. Goals are not equally important to a decision maker. Hence to effect a solution on the basis of preference, the goals may be rank-ordered into priority levels.

One should seek a solution which minimizes all unwanted deviations. There are various methods of measuring the effectiveness of the minimization of these unwanted deviations. The *lexicographic minimum* concept is a suitable approach, and it is defined as follows (see Ignizio, 1982; Ignizio, 1985).

Given an ordered array $\boldsymbol{f} = (f_1, f_2, \ldots, f_n)$ of nonnegative elements f_k, the solution given by $f^{(1)}$ is preferred to $f^{(2)}$ iff

$$f_k^{(1)} < f_k^{(2)}$$

and all higher-order elements (i.e., $f_1, \ldots, f_{k-1}$) are equal. If no other solution is preferred to $\boldsymbol{f}$, then $\boldsymbol{f}$ is the lexicographic minimum.

As an example, consider two solutions, $\boldsymbol{f}^{(r)}$ and $\boldsymbol{f}^{(s)}$, where

$$\boldsymbol{f}^{(r)} = (0,\ 10,\ 400,\ 56)$$

$$\boldsymbol{f}^{(s)} = (0,\ 11,\ 12,\ 20)$$

In this example, note that $\boldsymbol{f}^{(r)}$ is preferred to $\boldsymbol{f}^{(s)}$. The value 10 corresponding to $\boldsymbol{f}^{(r)}$ is smaller than the value 11 corresponding to $\boldsymbol{f}^{(s)}$. Once a preference is established, then all higher-order elements are assumed to be equivalent.

The Use of Coupled DSPs in Modeling and Solving Systems

Coupled compromise DSPs can be used to model multiobjective problems. These types of problems may involve the analysis and synthesis of problems from multiple disciplines, but in this section, the primary notion is *shared* design variable vectors. It is common in the design of complex systems, such as aircraft, for multiple design teams to include the same design variables in their design process. For example, consider the design of aircraft. The design variable, wing area, is used by the aerodynamics group to control the lift force. It is used by the structural group to control the structural framework of the wing. It is used by the propulsion group to control the amount of fuel needed. It seems advantageous to model systems this way, but this introduces an added level of complexity in a system model. Different disciplines are also dependent on the decisions made in other disciplines. Therefore, another focus is modeling the coupling of state variables between disciplines.

To begin the discussion, suppose that a designer (i.e., a decision maker) is investigating a typical engineering system such as a load-bearing structure, a high-speed mechanism, or an automotive transaxle. Assuming that this system is unambiguously described by a finite-dimensional variable vector $x \in \mathcal{R}^n$ (includes both the so-called "state" and "design" variables), then the general descriptive mathematical model can be characterized by a set of *parameters* p, which are fixed quantities over which the designer has no control, and the multifunction X, which represents the state equations and variational inequalities (derived from the natural laws such as the conservation principles) as well as the performance constraints for this system. Numerical approaches to such problems have been studied in Outrata (1990) and Outrata and Zowe (1995). If the only goal of the designer is to obtain a satisfying or a feasible design, then his/her task is simply to find a design x^o that satisfies the constraints and meets the goals as close as possible for a given p. On the other hand, if the designer wishes to find the design which minimizes the deviation between the goal achievement and aspiration with respect to several objective functions such as $f_i(x,p)$, i = 1, …, r, then the problem is to find a design x* that solves the following multiobjective formulation:

$$\underset{x \in X(p) \subset \Re^n}{\text{minimize}}\, d(x,p) = \left\{ d_1(x,p), \ldots, d_r(x,p) \right\} \tag{11.10.5}$$

where d_i represents the deviation from the objective f_i, to its target value, f_{iTV}. In other words, the problem is to minimize the deviation between what a designer wants and what a designer can achieve.

At this general level of discussion, one readily asserts that a model such as in Equation (11.10.5) is the typical starting point for much of the current education, research, and practice in mechanical systems modeling and applied optimization, and yet in specific design instances, this assertion should be boldly challenged. For example, since the designer only controls *x* and another foreign party controls *p*, how is *p* chosen? Can the designer assume that the foreign decision maker will always select the vector that is most advantageous to the design (as is tacitly assumed, it is argued later, by most justifications of existing engineering models)? If not, how should the designer respond to this conflict? A two-player strategic game has just been described (see, e.g., Von Neumann and Morgenstern, 1944; Aubin, 1979; Dresher, 1981) where one player controls *x* and the other player controls *p* and where *p* represents all decisions which are outside the scope of the designer. By the classical definition, a "game" consist of multiple decision makers or players (or designers, in this case) who each control a specified subset of system variables and who seek to minimize their own deviation functions subject to their individual constraints. To continue this argument further, consider a typical schedule of a product design and manufacture, as shown in Figure 11.10.20.

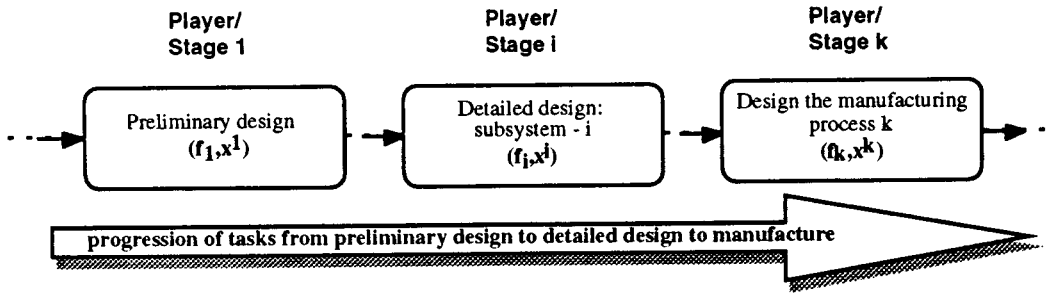


FIGURE 11.10.20 Different stages in a design time-line.

The decision makers at each of the several stages or stations in the time line are referred to as different players, even though they could be the same human designer. Each of the players controls and minimizes his own deviation function.

In the context where multiple decision makers collectively influence the final design, one must pose the following question: what mathematical formulation is appropriate for modeling the system in order to study the decisions made by *more than one player along the above time line?* Clearly, the conventional model of Equation (11.10.5) does not adequately represent the multiplayer scenario of the time line shown in Figure 11.10.20. This inadequacy is most acute when the players engage in strategic behavior which is anything other than total cooperation (e.g., dominance and noncooperation, among others). Thus, the major point of departure can be summarized as: (1) instead of formulating a single mathematical model at a single stage of the time line, as is done conventionally, these models represent one or more decision makers who determine the final design; and (2) each of the r decision makers has his own objective functions and one player's decision is influenced by the other. Thus, the final design could very well depend on whether the players compete, cooperate, or dominate one another — thus necessitating a study of the strategic interaction between the players. In short, a goal here is to model the design of Figure 11.10.20 as a *strategic game*, this figure perhaps being representative of a general engineering system design process at a high level of abstraction.

Mathematical modeling of such strategic behavior, where one decision maker's action depends on decisions by others, is well-established in wide-ranging applications from economics, business, and military (Aubin, 1979; Dresher, 1981; Fudenberg and Tirole, 1991; Mesterton-Gibbons, 1992). Such models are widely used in designing markets and auctions, and in predicting the outcome of encounters between competing companies or trading nations. As specific examples, consider automobile manufac-

turers who interact strategically in order to set prices and production schedules of cars; or the strategic behavior that occurs between drivers as they try to pass each other on a busy undivided highway.

References

Aubin, J.P. 1979. *Mathematical Methods of Game and Economic Theory.* North-Holland, Amsterdam.

Balling, R.J. and Sobieski, J. 1994. Optimization of coupled systems: a critical overview of approaches. In *5th AIAA/USAF/NASA/ISSMO Symp. on Recent Advances in Multidisciplinary Analysis and Optimization.* Panama City, FL. 753–773.

Cramer, E.J. Dennis, J.E., Frank, P.D., Lewis, R.M., and Subin, G.R. 1994. Problem formulation for multidisciplinary optimization. *SIAM J. Optimization.* 4(4): 754–776.

Dresher, M. 1981. *Games of Strategy.* Dover, New York.

Fudenberg, D. and Tirole, J. 1991. *Game Theory.* MIT Press, Cambridge, MA.

Ignizio, J.P. 1982. *Linear Programming in Single and Multi-Objective Systems.* Prentice-Hall, Englewood Cliffs, NJ.

Ignizio, J.P. 1985. *Introduction to Linear Goal Programming.* Sage University Papers. Beverly Hills, CA.

Ignizio, J.P. 1985. Multiobjective mathematical programming via the MULTIPLEX model and algorithm. *Eur. J. Operational Res.* 22: 338–346.

Lewis, K. and Mistree, F. 1995. On developing a taxonomy for multidisciplinary design optimization: a decision-based approach. In *The First World Congress of Structural and Multidisciplinary Optimization.* N. Olhoff and G.I.N. Rosvany, Eds., Elsevier Science, Oxford, UK: ISSMO. 811–818.

Mesterton-Gibbons, M. 1992. *An Introduction to Game-Theoretic Modeling.* Addison-Wesley, Redwood City, CA.

Mistree, F., Hughes, O.F., and Bras, B.A. 1993. The compromise decision support problem and the adaptive linear programming algorithm. In *Structural Optimization: Status and Promise*. AIAA, Washington, D.C., 247–286.

Outrata, J.V. 1990. On the numerical solution of a class of Stackelberg games. *Methods Models Operations Res.* 34: 255–277.

Outrata, J.V. and Zowe, J. 1995. A numerical approach to optimization problems with variational inequality constraints. *Math. Programming.* 68: 105–130.

Simon, H.A. 1982. *The Sciences of the Artificial.* MIT Press, Cambridge, MA.

Von Neumann, J. and Morgenster, O. 1944. *Theory of Games and Economic Behavior.* Princeton University Press, Princeton, NJ.

Lehman, R.L.; et. al. “Materials”
Mechanical Engineering Handbook
Ed. Frank Kreith
Boca Raton: CRC Press LLC, 1999

Materials

Richard L. Lehman
Rutgers University

Malcolm G. McLaren
Rutgers University

Victor A. Greenhut
Rutgers University

James D. Idol
Rutgers University

Daniel J. Strange
Alfred University

Steven H. Kosmatka
Portland Cement Institute

Weiping Wang
General Electric Corporate R&D

R. Alan Ridilla
General Electric Plastics

Matthew B. Buczek
General Electric Aircraft Engines

William F. Fischer, III
Lanxide Corporation

12.1 Metals

Victor A. Greenhut

Introduction — Nature and Properties of Pure Metals

Metals achieve engineering importance because of their abundance, variety, and unique properties as conferred by metallic bonding. Twenty-four of the 26 most abundant elements in the Earth's crust are

0-8493-9418-X/98/$0.00+$.50

metals, with only two nonmetallic elements, oxygen and silicon, exceeding metals in frequency. The two most abundant metallic elements, iron (5.0%) and aluminum (8.1%), are also the most commonly used structural metals. Iron is the most-used metal, in part because it can be extracted from its frequently occurring, enriched ores with considerably less energy penalty than aluminum, but also because of the very wide range of mechanical properties its alloys can provide (as will be seen below). The next 15 elements in frequency, found at least in parts per thousand, include most common engineering metals and alloys: calcium (3.6%), magnesium (2.1%), titanium (0.63%), manganese (0.10), chromium (0.037%), zirconium (0.026%), nickel (0.020%), vanadium (0.017%), copper (0.010%), uranium (0.008%), tungsten (0.005%), zinc (0.004%), lead (0.002%), cobalt (0.001%), and beryllium (0.001%). The cost of metals is strongly affected by strategic abundance as well as secondary factors such as extraction/processing cost and perceived value. Plain carbon steels and cast irons, iron alloys with carbon, are usually most cost-effective for ordinary mechanical applications. These alloys increase in cost with alloying additions.

A variety of metal properties are unique among materials and of importance technologically. These properties are conferred by metallic bonding, in which the “extra” outer valence electrons are “shared” among all metal ion cores. This bonding is different from other types of solids in that the electrons are free to acquire energy, and the metallic ions are relatively mobile, and quite interchangeable with regard to their positions in the crystal lattice, the three-dimensional repeating arrangement of atoms in a solid. This section of the chapter will concentrate on the mechanical properties of metals, for which metallic bonding provides ductile deformation, i.e., shows substantial permanent shape change under mechanical load prior to fracture. The ductility of metals at low and moderate temperature makes them formable as solids and also confers safety (fracture toughness) in mechanical applications, in that under impact loading the metal will absorb energy rather than break catastrophically.

Metals are good conductors of heat and electricity because thermal and electrical energy can be transferred by the free electrons. These two properties tend to parallel each other. For example, the pure noble metals (e.g., copper, silver, gold, platinum) are among the best electrical and thermal conductors. As a broad generalization, metallic elements with an odd number of valence electrons tend to be better conductors than those with an even number. These behaviors can be seen in Table C.6A of the Appendix. Thermal conductivity and electrical resistivity (inverse conductivity) have a reciprocal relationship and follow the indicted trends. As metals are alloyed with other elements, are deformed, contain multiple phases, and contain crystalline imperfections, their electrical and thermal conductivity usually decreases significantly from that of the pure, perfect, unalloyed metal. The specific values of thermal conductivity and electrical resistivity for several common engineering alloys is given in Table C.6B of the Appendix. Electrical and thermal conductivities tend to decrease proportionately to each other with increasing temperature for a specific metal. These conductivities may be altered if heating introduces metallurgical change during annealing (see subsection on mechanical forming).

Metals are opaque to and reflective of light and most of the electromagnetic spectrum, because electromagnetic energy is transferred to the free electrons and immediately retransmitted. This gives most metals a characteristic reflective “metallic color” or sheen, which if the metal is very smooth yields a mirror surface. At very short wavelengths (high energies) of the electromagnetic spectrum, such as X rays, the radiant energy will penetrate the material. This is applied in radiographic analysis of metals for flaws such as cracks, casting porosity, and inclusions.

Metals are almost always crystalline solids with a regular repeating pattern of ions. A number of atomic-level defects occur in this periodic array. A large number of atomic sites are “vacancies” (point defects) not occupied by atoms (Figure 12.1.1). The number and mobility of vacant sites increase rapidly with temperature. The number and mobility of vacancies in metals are quite high compared with other materials because there are no charge balance or local electron bond considerations. This means that solid metal can undergo significant changes with only moderate thermal excitation as vacancy motion (diffusion) provides atom-by-atom reconstruction of the material. Vacancies allow solid metals to homogenize in a “soaking pit” after casting and permit dissimilar metals to diffusion bond at moderate temperatures and within short times. In the process, substitutional metallic atoms (ions) shown in Figure

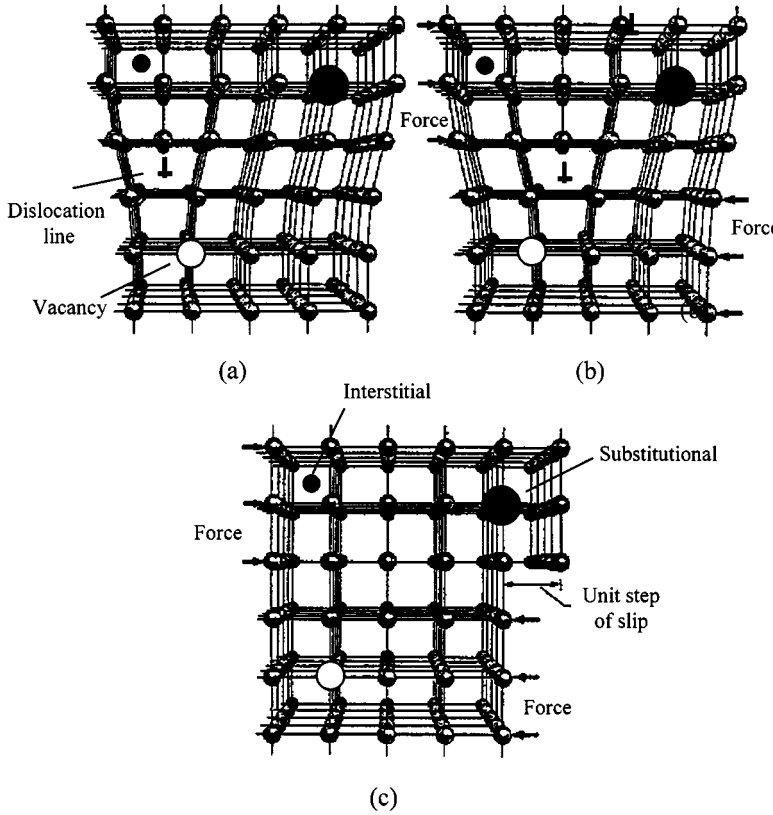


FIGURE 12.1.1 Point defects exist in the metal crystal structure: vacancies, substitutional atoms (ions), ions, and interstitial atoms (ions). A dislocation (⊥) line moves under an applied shear force until a surface step of plastic deformation is produced on the surface.

12.1.1 move via vacancy jumps while small interstitial atoms such as carbon (Figure 12.1.1) move from interstice to interstice. Vacancy mobility gives rise to major changes in mechanical properties during annealing (see subsection on mechanical forming) and is an important mechanism in creep deformation under load at elevated temperature (see subsection on corrosion and environmental effects).

At a slightly larger level, linear atomic packing defects known as dislocations, give rise to the ability of metallic materials to deform substantially under load. When a plane of atoms in the lattice ends, it gives rise to an edge "dislocation" such as that shown in Figure 12.1.1a. Such a dislocation can break and remake bonds relatively easily in a metal and thereby shift an atomic distance (Figure 12.1.1b). The process can continue until a surface step results. Many dislocations moving in this fashion can give rise to significant shape change in the material at moderate stresses. The onset of such massive dislocation motion in a metal is termed *yield* and occurs at the "yield stress" or "elastic limit" (see subsection on strength and deformation). Dislocations explain why the yield stress can be as low as about 100 Pa (10 psi) in a pure, pristine, single crystal of metal. Dislocations also explain how a fine-grained polycrystalline metal containing many microstructural features which interfere with dislocation motion may have a yield stress as great as 10 gPa (1000 ksi). Dislocations interact with each other in three dimensions and multiply. Therefore, dislocation motion can cause a major increase in dislocation density and yield stress, termed *cold work*. Vacancies can rearrange these dislocation tangles, restoring the metal to a condition closer to its original state, thereby lowering the yield stress. This can occur at moderate annealing temperatures (see subsection on mechanical forming).

The interaction of deformation, alloying elements, temperature, and time can cause a wide variety of microstructures in a solid metal down to near atomic levels with mechanical (and other) properties which

can vary over a very wide range. It is possible to manipulate the properties of a single metal composition over a very wide range in the solid state — a behavior which can be used to mechanically form a particular metal and then use it in a demanding load-bearing application. The use of minor alloying additions can provide a yet wider range of properties with appropriate thermal and mechanical treatment.

Casting

One of the important technological advantages of metals is their ability to incorporate a wide variety of secondary elements in a particular metal and thereby create alloys of the metal. Alloying can increase the strength of a metal by several orders of magnitude and permit the strength and ductility to be varied over a wide range by thermal and/or mechanical treatment, resulting in ease of mechanical forming or resistance to deformation.

Several metal phases may exist together in the solid as grains (crystals), or secondary phases may occur as smaller entities on grain (intercrystal) boundaries or within grains. Often the strengthening phase is submicroscopic and cannot be detected by optical metallography (reflection optical microscopy). The size and distribution of secondary phases is manipulated by thermomechanical (thermal and/or mechanical) treatment of the solid metal as well as the original casting procedure.

Casting methods include expendable mold casting (investment/precision, plaster mold, dry sand, and wet sand casting), permanent mold casting (ingot, permanent mold, centrifugal, and die casting), and continuous casting (direct chill and "splat" casting). These are listed in approximate order of cooling rate in Figure 12.1.2. As cooling rate increases, the grain (crystal) size tends to be smaller and the strength increases while compositional segregation decreases, providing more uniform properties. At the extremely high casting rates (10^5 to 10^6/sec) of continuous splat casting, it is possible to produce homogeneous metals not possible in terms of phase diagrams, and many metals have been produced in the amorphous state, yielding unusual metallic glasses. Ingot casting and continuous direct chill casting are primarily used to produce solid metal which will be extensively mechanically formed to final shape. The other casting methods are used to produce shapes near final dimensions, but to varying extends may receive extensive machining, forming, or finishing prior to use. For the latter group, grain refiners are frequently added to reduce solidification grain size. Metal tends to solidify directionally, with grains elongated in the direction of heat flow. This gives rise to directional mechanical properties which should be accounted for in design.

Comparison of Casting Methods

	Solidification Rate	Grain Size	Strength	Segregation
Expendable mold	Slow	Coarse	Low	Most
Investment	↓	↑	↓	↑
Plaster mold	↓	↑	↓	↑
Dry sand	↓	↑	↓	↑
Green sand	↓	↑	↓	↑
Reusable mold	↓	↑	↓	↑
Ingot	↓	↑	↓	↑
Permanent mold	↓	↑	↓	↑
Centrifugal	↓	↑	↓	↑
Die cast	↓	↑	↓	↑
Continuous — direct chill	↓	↑	↓	↑
Continuous — splat cast	Fast	Fine	High	Least

FIGURE 12.1.2 The effects of casting speed (solidification rate) are compared.

To obtain optimum properties and prevent flaws which may cause failure, the casting procedure must avoid or control compositional segregation, shrinkage cavities, porosity, improper texture (grain directionality), residual (internal) stresses, and flux/slag inclusions. This can be accomplished with good

casting practice. With the exception of investment (lost wax, precision) casting and to a lesser extent die casting, it is difficult to achieve very exacting tolerances and fine surface finish without postfinishing or forming of a casting.

Strength and Deformation, Fracture Toughness

Figure 12.1.3 shows a typical stress–strain diagram for a metal. The first portion is a linear, spring-type behavior, termed *elastic*, and attributable to stretching of atomic bonds. The slope of the curve is the "stiffness" (given for various metals in Table C.3 of the Appendix). The relative stiffness is low for metals as contrasted with ceramics because atomic bonding is less strong. Similarly, high-melting-point metals tend to be stiffer than those with weaker atomic bonds and lower melting behavior. The stiffness behavior is frequently given quantitatively for uniaxial loading by the simplified expressions of Hooke's law:

$$\varepsilon_x = \sigma_x / E \qquad \varepsilon_y = \varepsilon_z = -\upsilon \sigma_x / E \tag{12.1.1}$$

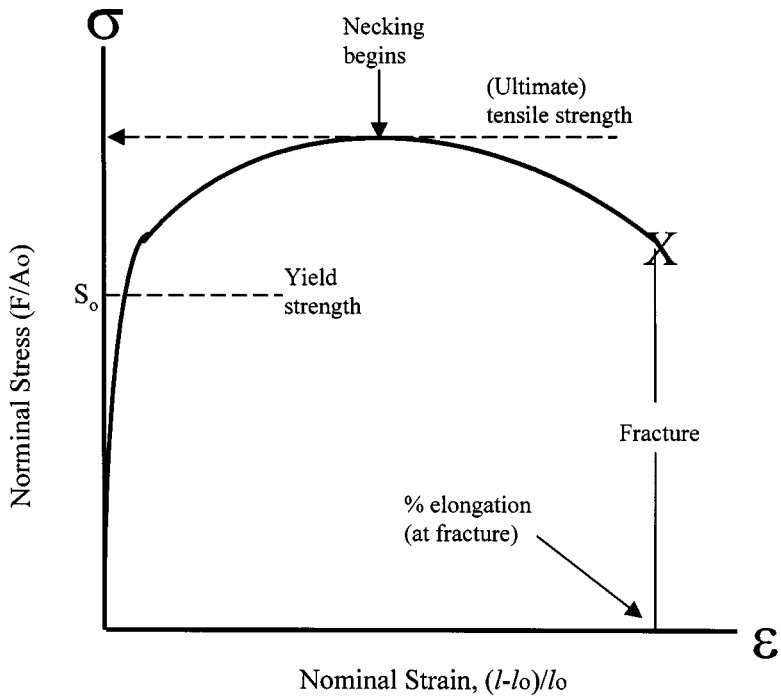


FIGURE 12.1.3 Typical engineering stress–strain curve for a metal.

Where σ_x is the stress (force per unit area, psi or Pa) in the *x* direction of applied unidirectional tensile load, ε_x is the strain (length per unit length or percent) in the same direction ε_y and ε_z are the contracting strains in the lateral directions, *E* is Young's modulus (the modulus of elasticity), and υ is Poisson's ratio. Values of the modulus of elasticity and Poisson's ratio are given in Table C.6A of the Appendix for pure metals and in C.6B for common engineering alloys. It may be noted that another property which depends on atomic bond strength is thermal expansion. As the elastic modulus (stiffness) increases with atomic bond strength, the coefficient of linear expansion tends to decrease, as seen in Table C.6.

The relationship of Equation (12.1.1) is for an isotropic material, but most engineering metals have some directionality of elastic properties and other structure-insensitive properties such as thermal expansion coefficient. The directionality results from directional elongation or preferred crystal orientation,

which result from both directional solidification and mechanical forming of metals. In most cases two elastic moduli and a Poisson's ratio are required to fully specify behavior. A principal modulus might be given in the rolling direction of sheet or plate with a secondary modulus in the transverse direction. A difference of 2 to 5% should ordinarily be expected, but some metals can show an elastic modulus difference as great as a factor of 2 in the principal directions of heavily formed material. Such directional differences should be accounted for when spring force or dimensional tolerance under load (or change of temperature) is critical in a design.

At a critical stress the metal begins to deform permanently, as seen as a break in the straight-line behavior in the stress-strain diagram of Figure 12.1.3. The stress for this onset is termed the yield stress or elastic limit. For engineering purposes it is usually taken at 0.2% plastic strain in order to provide a predictable, identifiable value. An extensive table of yield values and usual applications for commercial metals and alloys is given in Appendix C.5. In the case of steel a small yield drop allows for clear identification of the yield stress and this value is used. The onset of yield is a structure-sensitive property. It can vary over many orders of magnitude and depends on such factors as grain size and structure, phases present, degree of cold work, and secondary phases in grains or on grain boundaries as affected by the thermal and mechanical treatment of the alloy. The extension to failure, the ductility, and maximum in the stress–strain curve, the "ultimate stress" or "tensile strength" (see Appendix C.5) are also structure-sensitive properties. The strength and specific strength (strength-to-weight ratio) generally decrease with temperature.

The ductility usually decreases as the strength (yield or ultimate) increases for a particular metal. Reduction in the grain size of the metal will usually increase yield stress while decreasing ductility (Figure 12.1.4). Either yield or ultimate strength are used for engineering design with an appropriate safety factor, although the former may be more objective because it measures the onset of permanent deformation. Ductility after yield provides safety, in that, rather than abrupt, catastrophic failure, the metal deforms.

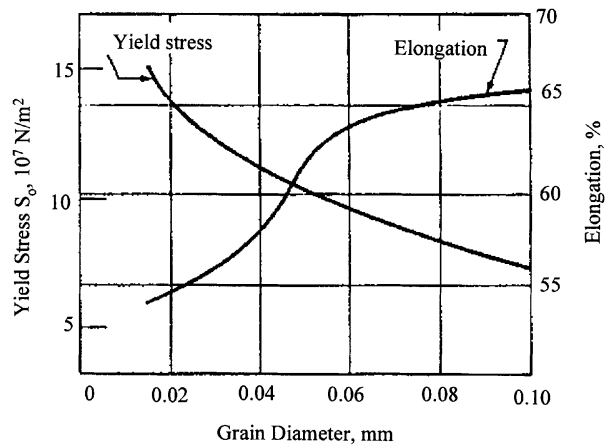


FIGURE 12.1.4 The effect of grain (crystal) size on yield stress and elongation to failure (ductility) for cartridge brass (Cu–30 Zn) in tension.

A different, independent measure is needed for impact loads — "toughness." This is often treated in design, materials selection, and flaw evaluation by extending Griffith's theory of critical flaw size in a brittle material:

$$\sigma_f = K_{1c} / \gamma c^{1/2} \tag{12.1.2}$$

where σ_f is the failure stress, K_{Ic} is a structure-sensitive materials property, the "fracture toughness" or "stress intensity factor" for a normal load, γ is a constant depending on orientation, and c is the depth of a long, narrow surface flaw or crack (or half that of an internal flaw). This is a separate design issue from that of strength. It is of particular importance when a metal shows limited ductility and catastrophic failure must be avoided. In some applications the growth of cracks, c is monitored to prevent catastrophic failure. Alternatively, as a performance test sufficient energy absorption as characteristic of a metal is determined when it is fractured in a Charpy or Izod impact test. Many metals will show a rapid decline in such energy absorption below a nil ductility temperature (NDT), which may establish a lowest use temperature for a particular metal in a particular state and for a particular application. Welds are often qualified by impact tests as well as strength testing. Care must be taken to apply the impact test appropriate to an application.

Hardness, the resistance of the near surface of a metal to penetration by an indentor, is also employed as a mechanical test. Increased hardness can often be correlated with an increase in yield and ultimate strengths. Typical hardness values for a large number of commercial metals and alloys are provided in Appendix C.5. A hardness indent is frequently done to "determine" the strength of a steel, using "equivalency" tables. Great caution must be taken in applying such tables because while hardness is an easy test to perform, it measures a complex and interactive set of properties, increasing with strength, elastic modulus, and work hardening rate. It is also an observation of surface properties which may not be characteristic of the bulk metal — particularly thick-gauge steel used in tension. Surface-hardening treatments can make the simplistic use of an "equivalency" table particularly dangerous. Application to nonferrous metals is also problematic. If a hardness tested part is to be put into service, the placement of hardness indents (surface flaws) can cause permanent failure.

A summary of important engineering metals can be found in Appendix C.5. This extensive table provides strength, hardness, and applications information for many commercial metals in varied heat treatments.

Mechanical Forming

Hot working is used when major shape change, cross-section reduction, or texture (directional) properties are desired. Cold working is preferred when close tolerances and fine surface finish are needed. The cold-worked form of a metal typically shows higher yield and tensile strength, as can be seen for several alloys listed in Appendix C.5. Rolling, forging, and extrusion are primarily done hot, while shape drawing, extrusion, deep drawing, stretching, spinning, bending, and high-velocity forming are more commonly performed cold. Hot rolling between parallel rollers is used to reduce ingots to plates, sheets, strips, and skelp, as well as structural shapes, rail, bar, round stock (including thick-walled pipe), and wire. Sheet metal and threads on round or wire stock may be rolled to shape cold. Closed die hot forging employs dies with the final part shape, while open die forging (including swaging and roll forging) uses less-shaped dies. Coining, embossing, and hobbing are cold-forging operations used to obtain precision, detailed surface relief or dimensions. Generally, extrusion and die drawing require careful control of die configuration and forming rate and, in the latter case, lubricant system. Impact extrusion, hydrostatic extrusion, and deep drawing (thin-walled aluminum cans) permit very large precise dimensional and cross-section changes to be made cold in a single pass. Stretching, spinning, bending are usually used to shape sheet or plate metal and the spring-back of the metal due to elastic modulus must be accounted for to obtain a precise shape.

Solute, Dispersion, and Precipitation Strengthening and Heat Treatment

Alloying additions can have profound consequences on the strength of metals. Major alloying additions can lead to multiphase materials which are stronger than single-phase materials. Such metal alloys may also give very fine grain size with further strengthening of material. Small alloying additions may also

substantially increase strength by solute strengthening as solid solution substitutional or interstitial atoms and or by particle strengthening as dispersion or precipitation hardening alloys.

Substitutional solute strengthening of copper by various atoms is shown in Figure 12.1.5. As the amount of an alloying element in solution increases, the strength increases as dislocations are held in place by the "foreign" atoms. The greater the ionic misfit (difference in size — Sn is a much larger ion than Ni), the greater the strengthening effect. The strength increase can be quite dramatic — as much as a 20-fold increase with a 1.5% addition to copper. Alternatively, a large addition of a very soluble element such as nickel can give major strengthening — monel, the Cu–70Ni alloy is more than four times stronger than pure copper (Figure 12.1.6). Interstitial solid solution carbon contributes to the strength of iron and is one contributor to strength in steels and cast irons. Solute strengthening can become ineffective in strengthening at elevated temperature relative to the absolute melting point of a metal as a result of rapid diffusion of substitutional and interstitial elements. The addition of more than one solute element can lead to synergistic strengthening effects, as this and other strengthening mechanisms can all contribute to the resistance of a metal to deformation.

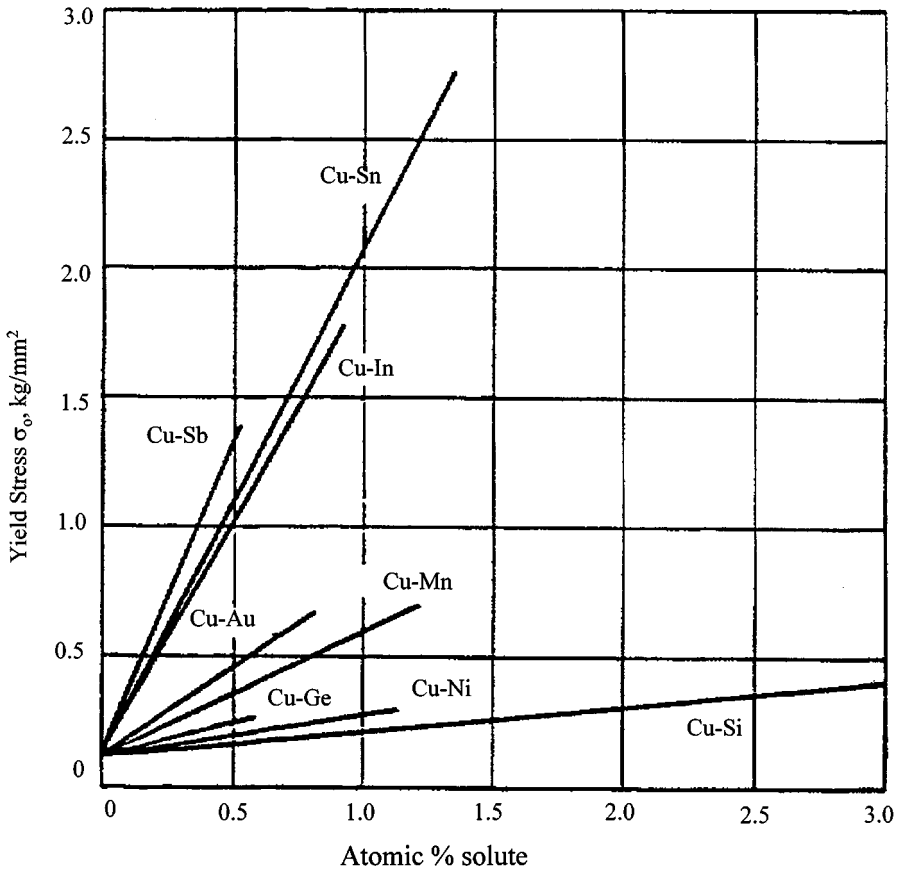


FIGURE 12.1.5 Effect of various substitutional atoms on the strength of copper. Note that as the ionic size of the substitutional atom becomes larger the strengthening effect becomes greater.

Ultrafine particles can also provide strengthening. A second phase is introduced at submicroscopic levels within each crystal grain of the metal. This may be done by a variety of phase-diagram reactions, the most common being precipitation. In this case the solid alloy is heated to a temperature at which the secondary elements used to produce fine second-phase particles dissolve in the solid metal — this is termed *solution heat treatment*. Then the metal is usually quenched (cooled rapidly) to an appropriate

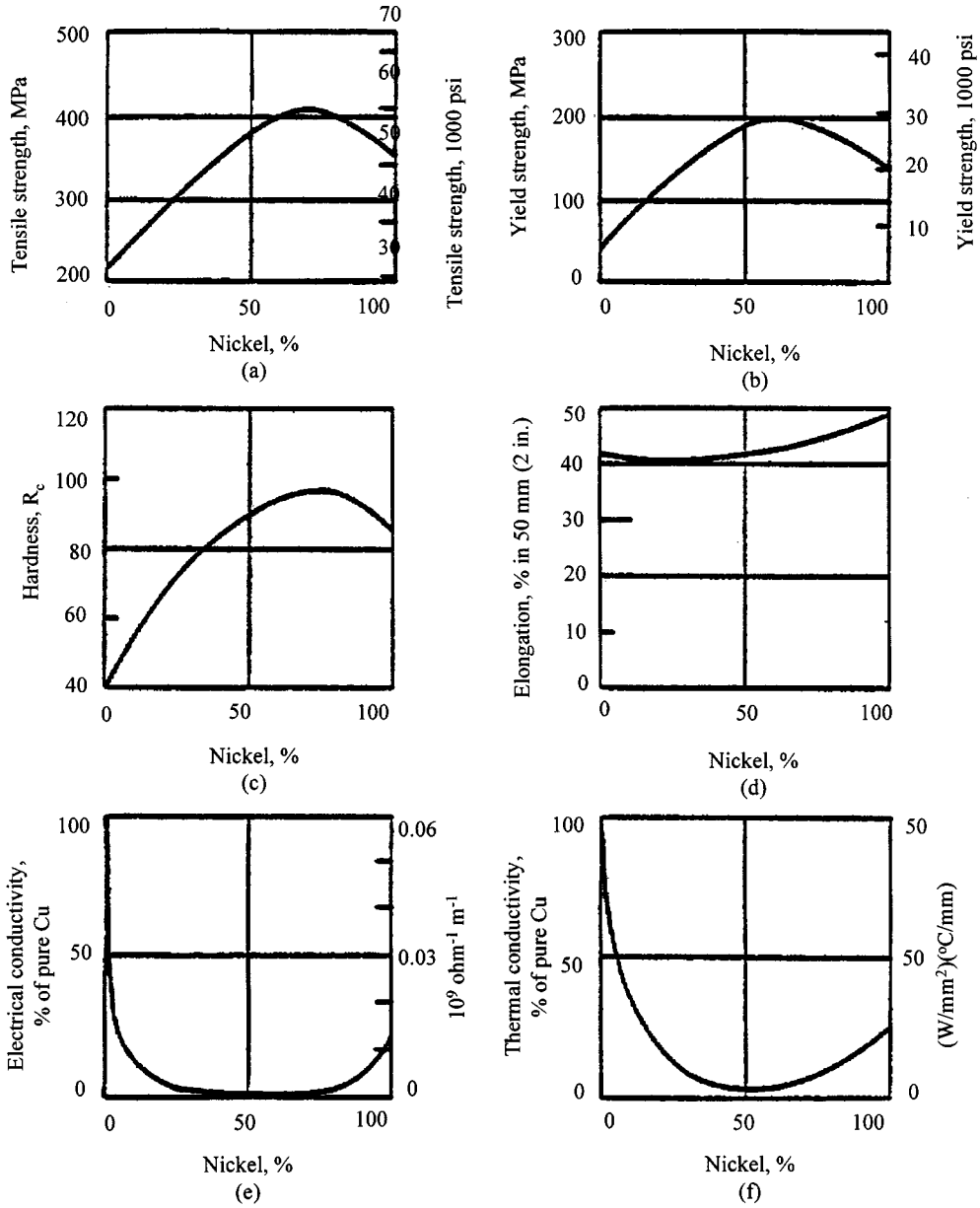


FIGURE 12.1.6 Variation in properties for copper–nickel random solid solution alloys. Note that over time at low temperature alloys (monel) may become nonrandom with significant strength increases.

temperature (e.g., room temperature or ice brine temperature) and subsequently held at an elevated temperature for a specified "aging" time during which particles precipitate and grow in size at near atomic levels throughout the solid metal. Temperature, time, alloy composition, and prior cold work affect the size and distribution of second-phase particles. The combination of treatments can be quite complex, and recently "thermomechanical treatments" combining temperature, time, and dynamic working have resulted in substantial property improvements. Heat treatment can be performed by the user, but it is difficult to achieve the optimum properties obtained by a sophisticated metallurgical mill. The heat treatment can manipulate structure and properties to obtain maximum strength or impact resistance. When metal is to be cold worked, a "softening treatment" can be employed which provides low yield

stress and high ductility. The difference between the "dead soft" and maximum strength conditions can be over an order of magnitude — a useful engineering property change.

Alternative surface diffusion methods such as nitriding and carburizing, which introduce particles for fracture and wear resistance, are presented in the subsection on metal surface treatments.

In the case of dispersion strengthening (hardening), the fine strengthening particles are a discontinuous second phase without atomic continuity with the matrix. The behavior of such particles is shown schematically in Figure 12.1.7a as a function of increasing aging time or aging temperature (fixed time) which result in larger, more widely spaced dispersed-phase particles. Under stress, dislocations must move around (bypass) such particles, so that yield strength decreases with increased aging. Long aging times may be used to decrease yield strength ("soften") of the metal for fabrication. A short aging time, would be used for maximum strength. The dispersed phase can also provide some enhancement of ductility. A dispersion-strengthened metal for which the dispersed phase is stable at elevated temperatures can provide both high-temperature strength and creep resistance (subsection on high-temperature effects). Surface diffusion treatments usually produce dispersion hardening.

Precipitation strengthening (hardening) employs particles which have at least some atomic continuity with the matrix metal. Thus, when the metal is deformed, dislocations can either bypass or pass through (cut) the particles. The resulting behavior is shown in Figure 12.1.7b. As aging time or temperature increases (particles grow larger and more widely spaced), the yield stress increases to a maximum and then decreases. The maximum is termed *critically aged*, and when this designation is part of an alloy treatment, precipitation strengthening may be assumed. For fabrication by cold working, the lower-strength, higher-ductility *underaged* condition is usually employed. There are different possible combinations of thermal and mechanical treatment which will provide a maximum critical aging treatment. Usually the best optimum for strength is given in handbooks and data sheets. However, improved treatments may be available, particularly of the combined thermomechanical type.

Strengthening of Steels and Steel Heat Treatment

Steels, perhaps the most important of all engineering metals, are alloys of iron and carbon usually containing about 0.02 to 1.0 $^{w}/_{o}$ carbon. The binary Fe–C phase diagram is important in describing this behavior and is shown in Figure 12.1.8. This diagram shows what phases and structures will occur in quasi equilibrium at various carbon contents and temperatures (under atmospheric pressure). Steel forming and heat treatment center on the transformation from austenite, γ phase, at elevated temperature to ferrite (α phase) plus cementite (iron carbide, Fe_3C) below 727°C (1340°F), the Ac_1 temperature, a eutectoid transformation. If there are no other intentional alloying elements, the steel is a "plain carbon" steel and has an AISI (American Iron and Steel Institute) designation 1002 to 10100. The first two characters indicate that it is a plain carbon steel, while the latter characters indicate the "points" of carbon.* Alloy steels, containing intentional alloying additions, also indicate the points of carbon by the last digits and together with the first digits provide a unique designation of alloy content. In the phase diagram (Figure 12.1.8) iron carbide (Fe_3C, cementite) is shown as the phase on the right. This is for all practical purposes correct, but the true thermodynamically stable phase is graphite (C) — relevant when the eutectic at 1148°C (2048°F) is used to produce cast irons (alloys greater than 2 $^{w}/_{o}$ C).

The solid-state eutectoid transformation is promoted by the magnetic effect in iron as nonmagnetic austenite transforms below the eutectoid (Ac_1) temperature to the two magnetic solid phases ferrite (iron with solid solution carbon) and cementite solid phase.** At the eutectoid composition, 0.77 $^{w}/_{o}$ carbon,

* Plain carbon steels contain about 0.2 $^{w}/_{o}$ Si, 0.5 $^{w}/_{o}$ Mn, 0.02 $^{w}/_{o}$ P, and 0.02 $^{w}/_{o}$ S.

** It should be noted that austenitic stainless steels (300 and precipitation hardening, PH, series designations), nonmagnetic alloys with considerable chromium and nickel content to provide corrosion resistance, do not ordinarily transform from austenite to the lower-temperature phases. They are not intentionally alloyed with carbon, are not magnetic, and do not show the phase transformation strengthening mechanisms of steels. The term *steel* is something of a misnomer for these alloys.

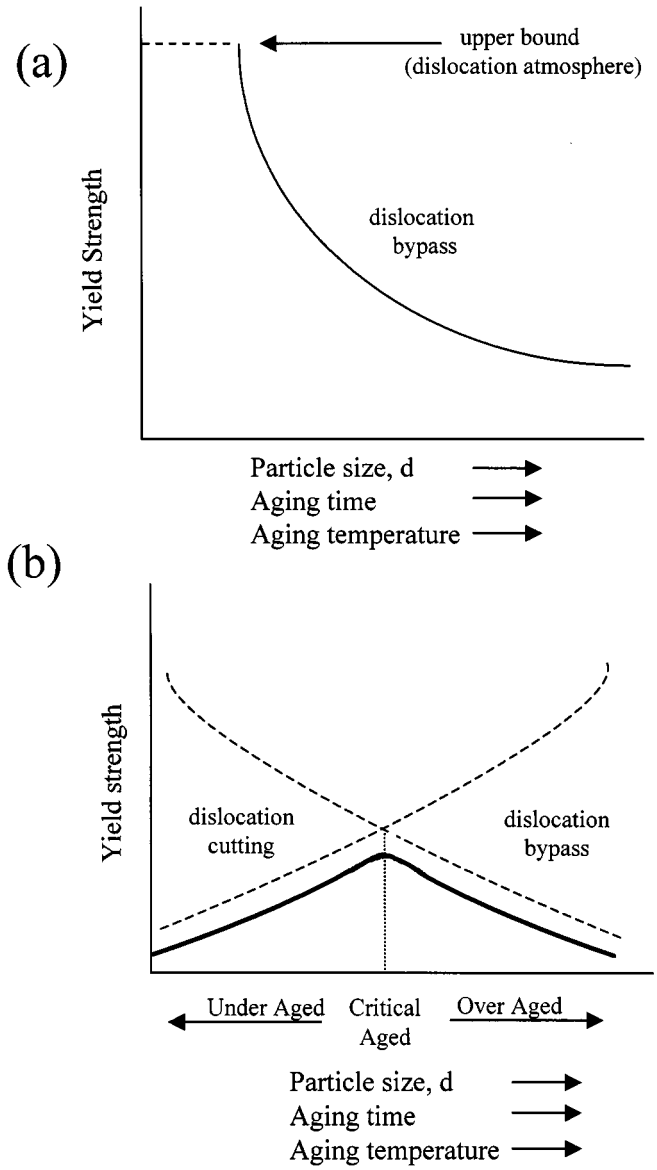


FIGURE 12.1.7 Effect of aging on dispersion- and precipitation-strengthened alloys for a fixed second–phase addition. (a) Dispersion strengthening: as aging time or temperature increases (dispersed phase particles larger and more separated), yield strength increases. A lower bound exists for near atomic size particles. (b) Precipitation hardening: two behaviors can occur giving a composite curve with a maximum at the critical aging time or temperature (optimum size and spacing of particles).

the two phases form as a fine alternating set of plates (lamellae) termed *pearlite* because of their pearllike appearance in a metallographic microscope. This two-phase structure of metal (ferrite) and carbide (cementite) provides strength (very slowly cooled — about 65 ksi, 14% tensile elongation), which increases as a more-rapid quenching yields a finer pearlite microstructure (to about 120 ksi). As strength increases, ductility and fracture toughness decrease. With yet more rapid quenching and more local atomic diffusion, the austenite transforms to bainite, a phase of alternating carbon and iron-rich atomic planes. This has yet higher strength (to about 140 ksi) and lower ductility. When the metal is quenched so rapidly that carbon diffusion is prevented, the austenite becomes unstable. Below a critical temperature,

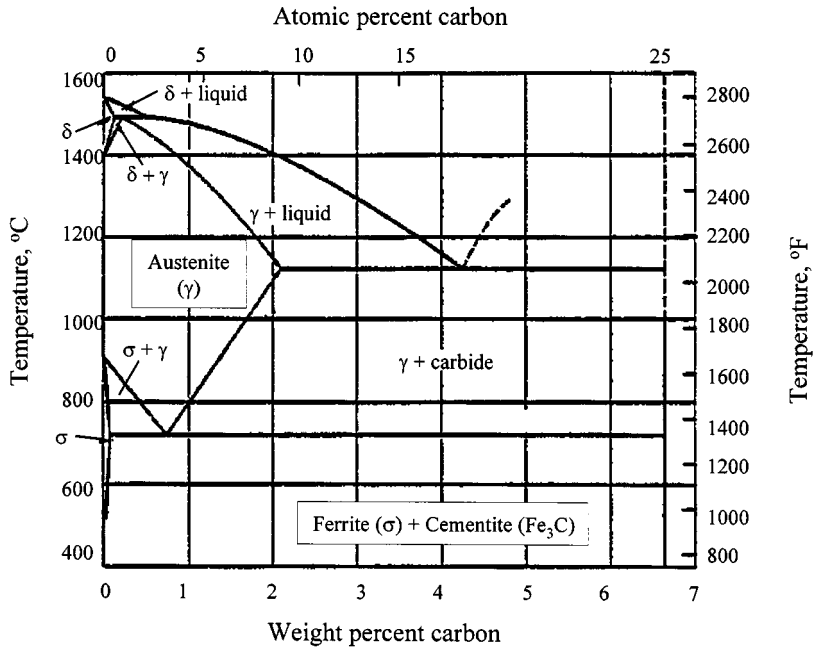


FIGURE 12.1.8 Iron–carbon phase diagram relevant for steel. The steel composition range is from about 0.02 to 1.00 $^{w}/_{o}$ carbon. Steel strengthening treatments require heating into the austenite region (above the Ac_3) and then quenching.

the martensitic start temperature (M_s), the metal transforms spontaneously by shear to martensite. Full transformation occurs below the martensite finish temperature (M_f). The formation of this hard phase introduces enormous microscopic deformation and residual stress. The strength is very much higher (about 300 ksi) but there is almost no ductility. This rapidly cooled material can spontaneously fail from "quench cracking," which results from residual stresses and the martensite acting as an internal flaw. To relieve stresses and provide some fracture toughness, martensitic steel is "tempered" at an intermediate temperature such as 500°C for about an hour to provide some ductility (about 7%) while sacrificing some strength (about 140 ksi). Tempering for shorter times or at lower temperatures can give intermediate properties. High-carbon steels are often used for cutting tools and forming dies because of their surface hardness and wear resistance. When a high-carbon steel (>0.7 $^{w}/_{o}$ C) requires fabrication at lower temperature, it may be held at a temperature just under the eutectoid for an extended time (e.g., for 1080 steel: 700°C, 1300°F — 100 hr) either after or without quenching to provide a soft condition (<60 ksi, 20% extension). A variety of different quenching temperatures, media, and procedures can be used to vary required combinations of the microstructures above and mechanical properties. The discussion above centered on eutectoid steel and holds for other high-carbon steels. Increasing carbon content favors the formation of martensite in steels, thereby providing increased strength, hardness, and wear resistance. However, such steels can be quite brittle.

At lower carbon content, "primary ferrite" forms (Figure 12.1.8) as the steel is quenched from above the boundary of austenitic (γ) region, the Ac_3. Subsequently, pearlite (ferrite and cementite), bainite, and/or martensite can form. Lower carbon content increases the amount of primary ferrite, a weaker/ductile phase, and decreases the tendency to form martensite, a stronger/brittle phase. The result is a more fracture tough, ductile ("safer") steel, but strength is lower. Such steels are also mechanically more forgiving if welded. Thus, a 1010 steel (0.10 $^{w}/_{o}$ C) might be used for applications where extreme "formability" and "weldability" are required, such as for car bodies and cans, while a 1020 steel (0.20 $^{w}/_{o}$ C) might be used for construction materials for which some increased strength is desired while

maintaining safety. A medium-carbon steel such as a 1040 would be used when a balance of strength and toughness (and ease of welding) is needed.

In order to quench (and temper) steels continuous cooling transformation (CCT) diagrams are used such as that in Figure 12.1.9 for a 1040, medium-carbon steel. The steel is quenched from above the Ac_3. Vertical lines indicate quenching rates as shown in the lower left inset of the diagram. The cooling rate for the center of a round bar of given diameter quenched in air, oil, and water is given below the diagram. Solid lines on the diagram indicate percent transformation (start, 10%, 50%, 90%, finish of transformation), and a dotted line separates the region where primary ferrite forms from that of transformation to pearlite. The lower diagram shows the indentation hardness to be expected, as this and other mechanical behavior can be predicted from the CCT curve. As an example, the center of a 15 mm ($\sim^1/_2$ in.) diameter bar quenched in air would be about 25% primary ferrite and 75% fine pearlite, with a Rockwell C hardness (HRC) of about 15. If quenched in oil, the bar center would be chiefly bainite with a small amount of martensite (HRC $\cong$ 25). The same bar quenched in water would be all martensite (HRC $\cong$ 55) before tempering. The transformation at other positions in the bar and for other engineering shapes (sheet, pipe, square rod, etc.) can be obtained from conversion curves.

It should be noted that layers in the steel closer to the quenched surface cool more quickly and are therefore displaced toward the left of the CCT diagram. There is a variation in structure and mechanical properties from the quenched surface to the center. One result is that the surface of the steel tends to be stronger, harder, and more wear resistant than the center. A steel beam undergoing bending has maximum strength at the near surface which undergoes the greatest stresses, while the center provides safety because of its relative fracture toughness and ductility. A thin knife edge cools quickly and can be resistant to deformation and wear, while the thick back prevents the blade from snapping in half when bent.

Many elements may be incorporated in steels to promote specific properties. Almost all common additions (other than cobalt) tend to promote strengthening by the formation of martensite (or bainite) instead of pearlite. When alloying additions tend to promote martensite throughout a thick section independent of cooling rate, the alloy is said to have *high hardenability.* Some elements such as chromium, molybdenum, and nickel also may help to provide high-temperature strength and environmental resistance. One class of alloys with relatively small alloying additions are termed *HSLA* steels (high strength, low alloy) and usually show somewhat superior mechanical properties to their plain carbon steel equivalents. As steel is alloyed, the relative cost increases substantially. A useful strategy in steel selection (and steels are usually the first engineering candidate on a cost basis) is to start by determining if a medium-carbon steel will do. If greater safety and formability are needed, a lower carbon content may be used (with slightly lower cost); alternatively, a higher-carbon steel would be chosen for greater strength and wear resistance. Heat treatment would be used to manipulate the properties. If plain carbon steels prove unsatisfactory, the HSLA steels would be the next candidate. For very demanding applications, environments, and long-term operations, specialty alloy steels would be selected insofar as they are cost-effective.

Fatigue

Fatigue is the repeated loading and unloading of metal due to direct load variation, eccentricity in a rotating shaft, or differential thermal expansion of a structure. Even substantially below the yield point (elastic limit) of a metal or alloy this repeated loading can lead to failure, usually measured in terms of the number of cycles (repeated load applications) to failure. Some studies have suggested that well over 80% of all mechanical failures of metal are attributable to fatigue.

High-stress, low-cycle fatigue usually occurs at stresses above the yield point and lifetimes are tens or hundreds of cycles (to about a thousand cycles). Failure occurs as a result of the accumulation of plastic deformation, that is, the area (energy) under the stress–strain curve (Figure 12.1.3). A simple lifetime predictive equation can be used to predict lifetime:

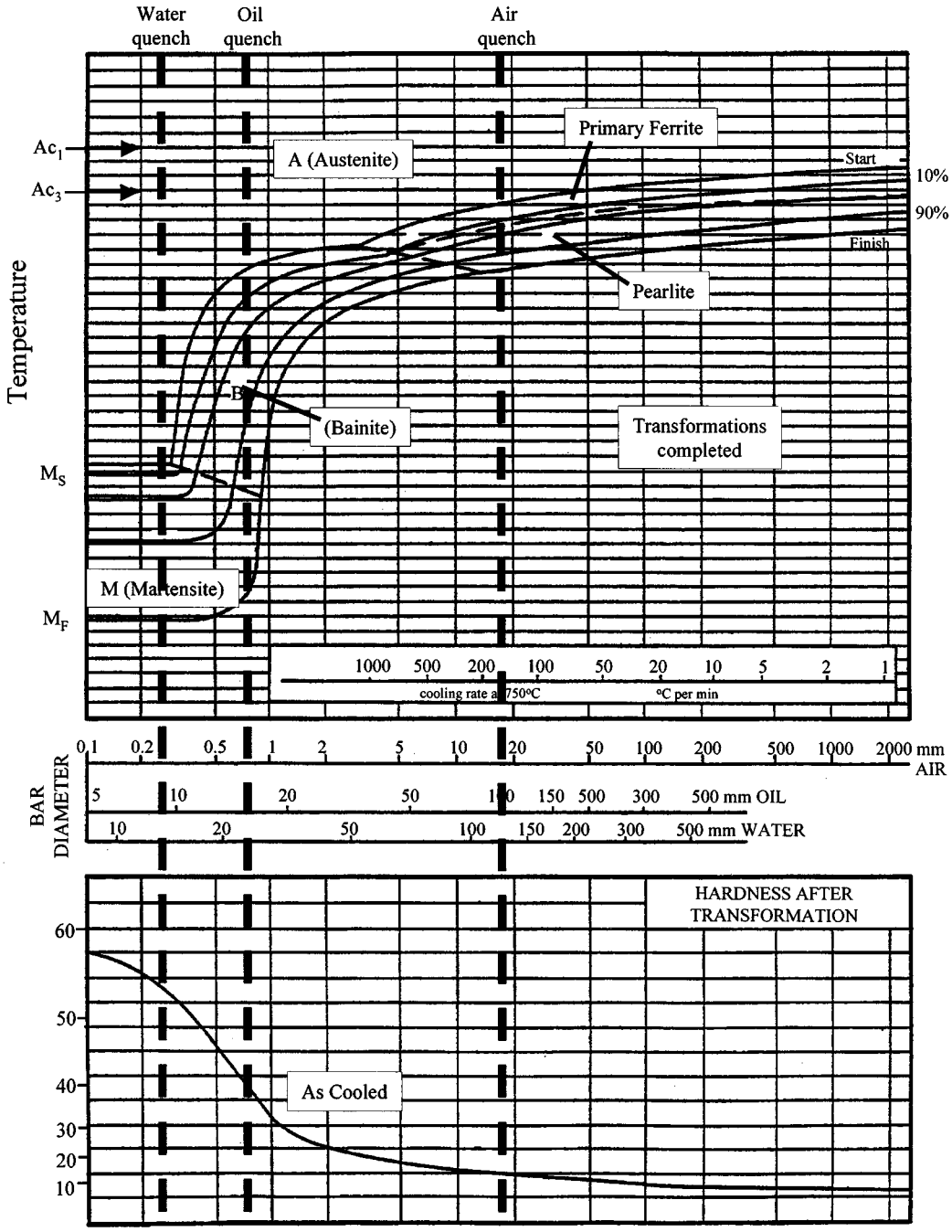


FIGURE 12.1.9 CCT diagram for 1040 steel. A 15-mm round bar air quenched from above the Ac_3 will be chiefly fine pearlite (ferrite and iron carbide, cementite) with about a quarter of the structure primary ferrite at the center. Quenching in oil will yield bainite with about 10% martensite. Water quenching will produce wholly martensite at the center. (After Atkins, M., *Atlas of Continuous Cooling Transformation Diagrams for Engineering Steels*, ASM International, Materials Park, OH, 1980.)

$$N \approx \left[\varepsilon_u / 2\varepsilon_{pf}\right]^2 \tag{12.1.3}$$

where N is the number of cycles to failure, ε_u is total strain from the stress–strain curve, and ε_{pf} is the plastic strain amplitude in each fatigue cycle.

More commonly, metals are used well below their yield point and fail after many, many cycles of repeated loading in low-stress, high-cycle fatigue. There is microscopic, local plastic deformation (cold working) and vacancy generation (recovery effects) during such cyclic loading which result in "fatigue hardening" (strengthening) of unworked metal and "fatigue softening" of unworked metal. Some have even found success relieving residual stressed with a vibratory anneal. Early in the fatigue process surface flaws or in some cases severe internal flaws begin to propagate. The fatigue crack propagates in areas of high stress a small, usually submicroscopic, distance with each tensile loading. The propagation on each cycle frequently leaves identifiable marking on the failure surface termed *fatigue striations* which mark the progress of the subcritical crack. When the crack becomes so large that the fracture toughness criterion is exceeded (Equation 12.1.2), catastrophic overload failure occurs. When the future fatigue loading can be predicted and the cyclic crack propagation rate is known, fatigue cracks can be inspected or monitored in different applications such as aircraft structures and pressure vessels to decommission or replace parts before fatigue failure. This must be done cautiously because a change to a more aggressive (corrosive, oxidative, elevated temperature) environment can increase the crack propagation rate. If a harmonic resonance occurs in the metal part, vibratory maxima can cause premature fatigue failure. Harmonics can change as fatigue cracks propagate. Harmonic vibration can be prevented with vibratory (dynamic) design concepts and/or direct monitoring.

Figure 12.1.10 shows typical metal *S–N* curves (stress vs. number of cycles to failure) for a high-strength aluminum and for a titanium alloy. Note that the convention is to make stress the vertical axis and to plot the number of cycles to failure on a logarithmic scale. For high-stress, low-cycle fatigue ($<10^3$ cycles) the curve is flat and linear, consistent with the model of Equation12.1.3. For high cycle fatigue the lifetime is a rapidly varying function of stress until very low stresses (long lifetimes occur). The actual fatigue life varies statistically about the mean value shown in approximate proportion to the number of cycles to failure. These curves are for testing in ambient air. Fatigue life would be longer in an inert environment and may be shortened drastically in an aggressive environment. Iron- and titanium-based alloys, such as the example shown, usually have an "endurance limit," a stress below which lifetime is ostensibly infinite. In air, at room temperature, the endurance limit is about half the tensile strength for most iron and titanium alloys. Other metals appear to show no stress below which they last indefinitely. Therefore, a "fatigue limit" is designated — usually the stress at which the fatigue life is 10^8 cycles. This may be a long lifetime or not depending on the frequency of loading and engineering lifetime. The fatigue limit is generally about 0.3 times the tensile strength for metals with strengths below about 100 ksi (700 MPa). The factor is somewhat poorer for higher-strength metals. It is apparent that on a relative strength basis, iron- and titanium-based alloys are fatigue-resistant metals when compared with others. Dispersion-strengthened alloys have been seen to provide some lifetime advantage in fatigue.

A number of mathematical relationships have been proposed to predict fatigue life, but none works with complete success and all require experimental data. Perhaps the most successful of the so-called fatigue "laws" are the "cumulative damage" laws. The simplest is Miner's law:

$$\sum i\left[n_i / N_i\right] = 1 \tag{12.1.4}$$

where n_i is the number of cycles applied and N_i is the number of cycles for failure at a particular stress level, σ_i. The conceptual basis is that the number of fatigue cycles at a stress level uses up its relative fraction of total fatigue lifetime and may be correlated to fatigue crack propagation (striation spacing). Modifications of this model account for the order and relative magnitude of loads. Several techniques

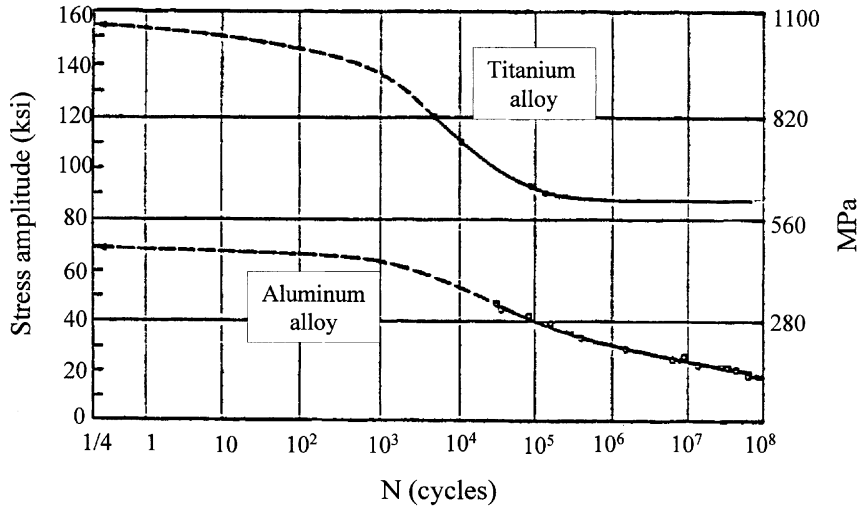


FIGURE 12.1.10 *S–N* (fatigue) curves for high-strength titanium (upper curve) and aluminum alloys. Note that titanium (and iron) alloys show an endurance limit, a stress below which the metal lasts indefinitely.

have met with partial success in increasing fatigue lifetime beyond that predicted by the cumulative damage models:

1. Coaxing — intermittent or continuous, superimposed vibration at very low stress,
2. Overstressing — intermittent or superimposed compressive loading,
3. Surface compression — intermittent shot peening or surface rolling,
4. Surface removal — chemical or mechanical surface polishing.

All but 1 are thought to close or eliminate fatigue cracks and surface damage. Coaxing has been said to introduce more of the vacancy-related recovery effects associated with fatigue.

High-Temperature Effects — Creep and Stress Rupture

Elevated temperature can cause a significant loss of strength and elastic modulus, so that a metal part may fail as a result of overheat even at loads which appear small at room temperature. High temperature is a relative matter and is usually judged as a fraction of the melting point measured on the absolute temperature scale. Thus, even moderate temperature excursions can be important for low-melting-point metals and alloys. As indicated in several sections above, many alloying and cold work strengthening methods depend upon heat treatment; an alloy can undergo metallurgical change due to overheat or to long-term holds at moderate temperatures and thereby alter properties significantly. Thus, the thermal stability of the microstructure should be determined. For example, metallographic replica techniques have been developed for determining in the field if microstructure has coarsened, making the metal weak.

Creep deformation is the continued deformation of a metal under load at elevated temperature, usually at a design stress well below the yield point. While measurable creep can occur at low temperatures over very long times or at very high (compressive) loads, creep usually becomes of engineering importance above about two thirds the melting point (absolute) of an alloy. Thus, both lead, which creeps at room temperature, and tungsten, which creeps in an incandescent light bulb at white heat, require a mechanical support or creep-resistant alloying additions. Figure 12.1.11 shows a schematic creep curve plotting creep strain* vs. time, *t*, at a particular tensile load and temperature. After the immediate elastic

* If a constant extension is applied to an object, the force it generates will decline over time due to creep. This is called *stress relaxation* and can be treated in a similar way.

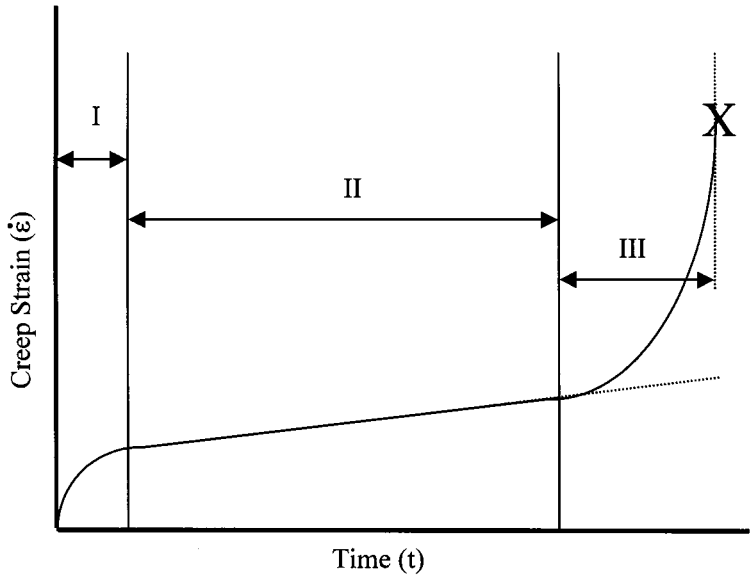


FIGURE 12.1.11 Schematic creep curve showing transient creep (I), steady-state creep (II), and tertiary creep (III). Note that the slope of the straight-line portion is the steady-state creep rate (minimum creep rate).

strain (t = 0), "transient creep" occurs in Region I. At elevated temperature this usually follows a $t^{1/3}$ behavior (Andrade creep). The majority of the curve has a straight-line behavior (Region II) in which the extension with time is constant. The slope of this part of the curve is termed the *steady-state strain rate* or *minimum creep rate*. It is used to calculate the creep extension which could cause functional failure when dimensional tolerances are exceeded. A simplified predictive model is

$$\varepsilon' = A\sigma^m e^{-h/kT} \tag{12.1.5}$$

where ε' is the minimum creep rate, A is a constant, σ is the applied stress, m is the stress dependence exponent (often 4 to 8), h is the activation enthalpy for creep, and T the absolute temperature.

In Region III the creep rate accelerates as the metal necks down severely in a local area, thereby increasing the local stress. The steady-state rate would continue (dotted line) if the load were adjusted to give constant load at the minimum cross section. Since loads do not readjust to compensate for necking in real applications, a final accelerated stage can be experienced. For example, a blowout can occur in late creep of a pressurized high-temperature piping system. The time to failure is termed the *stress rupture lifetime*. Predictive models are developed from Equation (12.1.5) to provide lifetime information.

Lowering use temperature or applied stress decreases susceptibility to creep deformation and increases stress rupture lifetime. Frequently, a moderate temperature decrease is most effective in this regard. Higher-melting-point metals are more creep resistant, so that the refractory metals tungsten and molybdenum can be used but require an inert atmosphere or protective coating to prevent rapid oxidation. In air or other active atmosphere, niobium-, nickel-, titanium-, and iron-based alloys may be used for creep resistance. Dispersion hardening, particularly with a high-temperature stable phase, such as an oxide, nitride, or carbide, can confer a degree of creep resistance.

Corrosion and Environmental Effects

Corrosion usually involves the slow removal of metal due to chemical and/or electrochemical reaction with an environment. Most metallic corrosion involves a galvanic, that is, electrochemical, component and localized attack in the form of pitting, attack in crevices, grain boundary attack, selective leaching

of one phase, and exfoliation (attack parallel to the surface causing layers to peel away). Uniform corrosion is less common and also less of a concern as long as the rate of material removal is gradual and predictable.

Galvanic corrosion requires a medium, usually liquid (water), often containing a specific agent which promotes corrosion, a potential difference, and a complete electrical circuit, as well as an anode and a cathode. The anode will be the attacked, corroded, portion of the system. One cause for potential corrosion is dissimilar metals. For example, the galvanic series in seawater shows magnesium alloys as most anodic followed by zinc, aluminum, iron, nickel, brass, and copper alloys which are progressively more cathodic. When coupled electrically, the more-anodic metal would be corroded and the more cathodic would promote corrosion. One way of preventing corrosion is to isolate with insulator material the metal parts electrically so that a circuit is interrupted. The surfaces can be isolated from the corroding medium with paint, protective metal, conversion coatings, or a corrosion inhibitor additive to the liquid. In this situation it is more important to coat the cathode (noncorroding metal) since it impresses corrosion on the anodic metal dependent on exposed area. When in doubt, all parts should be painted repeatedly to prevent pinholes. Other possibilities include cathodic protection with an impressed electrical counterpotential or use of a sacrificial anode which is attacked instead of the metal components. Some metals such as stainless steels and aluminum alloys provide protection via an oxide coating, a *passive film*, which will form under specific, controlled electrochemical conditions providing *anodic protection.*

Another source for an anode and cathode can be the solution itself. Differences in temperature, ion concentration, oxygen content, and pH can all lead to a potential difference which results in corrosion. Often corrosion occurs far more rapidly for a specific range of solution concentration — indeed, sometimes dilution may accelerate corrosive attack. An oxygen deficiency cell under dirt or in a crevice frequently causes attack at the resulting anode. Potential differences can also exist in the metal itself: between different phases, inclusions, or grain chemistries; between grains and grain boundaries; between surface films and metal; and between different grain orientations. Selective leaching such as dezincification of high-zinc brasses (Cu–Zn) is an example for composition differences, while exfoliation corrosion of aluminum alloys and sensitized stainless steel are examples of grain boundary attack. Cold-worked metal tends to be anodic to annealed material so that a heavily formed part of a metal part may corrode preferentially.

Another important environmental cause of failure is stress corrosion cracking (SCC). A combination of applied or residual tensile stress and environmental attack results in progressive slow crack propagation over a period of time. Eventually, the crack becomes critical in size (Equation 12.1.2) and catastrophic failure occurs. There need be no evidence of corrosion for SCC to occur. When the loading is of a cyclic nature the effect is termed *corrosion fatigue*. To prevent these long-term crack propagation effects the environment and/or the source of tensile load may be removed. Considerable effort has been made to identify an ion concentration below which stress corrosion cracking will not occur (e.g., Cl for austenitic stainless steels), but there may be no level for complete immunity, merely a practical maximum permissible level.

Hydrogen embrittlement and hydrogen cracking can occur in the presence of stress and a hydrogen source. Embrittlement results when hydrogen diffuses into the metal and/or acts on the crack tip altering the fracture toughness. Hydrogen cracking may be regarded as a special case of stress corrosion cracking. Either environmental effect can lead to catastrophic failure. The source of hydrogen can be an acid solution, hydrogen evolved during corrosion, electrochemical treatment (plating, electropolishing), or hydrocarbons. Often isolation from the hydrogen source is difficult because hydrogen diffuses quickly through most materials and barrier coatings. Glass coating has met with some success.

Metal Surface Treatments

A number of treatments are employed to strengthen the surface of steels and make them more resistant to failure or wear. Some of the techniques may also be applied to selected nonferrous alloys. Flame, induction, and laser hardening provide intense heat to the outer surface of a medium- or high-carbon

(hardenable) steel bringing it into the austenitic region, above Ac_3 (see subsection on fatigue). The bulk of the metal is not heated so that the surface can then be quenched rapidly forming hardening phases and a compressive surface stress. This provides strength and wear resistance. Another surface-hardening technique is carburizing. This can also be used on steels with lower carbon content. The metal is exposed to a controlled balance of carbon monoxide and carbon dioxide or is packed in graphite. At elevated temperature (usually above the Ac_3) carbon diffuses into the surface, converting it to a high-carbon steel. The steel is then either quenched directly from the carburizing temperature or reaustenitized and quenched. The result is similar to flame hardening, but higher hardness and surface compression can be accomplished compared with flame hardening. The center of the piece, with much lower carbon content, can provide fracture toughness, ductility, and safety. Nitriding exposes steel containing appropriate alloying elements (chromium, aluminum, vanadium, …) to monatomic nitrogen in the form of cracked ammonia, cyanide, or high-energy dissociated nitrogen gas. This is done below the eutectoid (lower transformation) temperature. Dispersed-phase nitrides are formed as nitrogen diffuses into the surface which harden the surface without a need for further heat treatment. The effects of both carburizing and nitriding can be introduced by carbonitriding above the transformation temperature and quenching or nitrocarburizing below the transformation to austenite.

Suggested Reading

ASM International, 1985, *Metals Handbook Desk Edition*, ASM International, Materials Park, OH.
ASM International, *Metals Handbook, 8th–10th ed.*, ASM International, Materials Park, OH.

12.2 Polymers

James D. Idol and Richard L. Lehman

Introduction

Polymers constitute a wide range of materials which are derived at least in part from organic, usually petroleum-based, raw materials; they consist of repeating molecular units and have special properties obtained by engineering the form of the molecular structures. The term *polymer* is derived from Greek roots and means "having many parts," a term which aptly describes the infinite number of compounds which can be synthesized from a relatively limited number of monomer units. The term *plastic* is often used in describing polymers, although this term is not in current usage since it is a general descriptive which refers to the forming rheology of many polymers but is too general to accurately describe this group of materials.

Polymers are used as engineering materials in the neat form, i.e., as the pure material, or in combination with a large diversity of additives, both organic and inorganic. These additives may be, among others, plasticizers which reduce the rigidity or brittleness of the material, fillers which increase strength and load deflection behavior under load, or stabilizers which protect the polymer against ultraviolet radiation.

The following discussion will separate polymers into two groups, thermoplastic and thermosetting, based on the distinctly different thermal processing behavior of these two broad classes of polymers. Thermoplastic polymers soften when heated and can be reshaped, the new shape being retained on cooling. The process can be repeated many times by alternate heating and cooling with minimal degradation of the polymer structure. Thermosetting polymers (or thermosets) cannot be softened and reshaped by heating. They are plastic and moldable at some state of processing, but finally set to a rigid solid and cannot be resoftened. Thermosets are generally stronger and stiffer than thermoplastic.

Table 12.2.1 of this section gives an overview of the physical properties of the most commonly used industrial polymers. Table 12.2.2 provides an overview of properties such as chemical resistance, ease of machining, and compressive strength for thermoplastic and thermosetting plastics, while Table 12.2.3 is a selection guide for polymers by application. A detailed summary of polymer properties, including electrical properties, thermal properties, optical properties, and fabrication, is presented in Table 12.2.4.

Thermoplastic Polymers

Acetal and Polyacetal

These combine very high strength, good temperature and abrasion resistance, exceptional dimensional stability, and low coefficient of thermal expansion. They compete with nylon (but with many better properties) and with metal die castings (but are lighter). Chemical resistance is good except for strong acids. Typical applications are water-pump parts, pipe fittings, washing machines, car instrument housings, bearings, and gears.

Acrylics (Methylmethacrylate, PMMA)

These are noted for their optical clarity and are available as sheet, rod, tubings, etc., as Perspex (U.K.) and Plexiglas (U.S., Germany, etc.). They are hard and brittle and quite resistant to discoloring and, especially, weathering. Applications include outdoor display signs, optical lenses and prisms, transparent coverings, drafting instruments, reflectors, control knobs, baths, and washbasins. They are available in a wide range of transparent and opaque colors.

Acrylonitrile-Butadiene-Styrene (ABS)

This combination of three monomers gives a family of materials which are strong, stiff, and abrasion resistant with notable impact-resistance properties and ease of processing. The many applications include

TABLE 12.2.1 Physical Properties of Polymers

Properties of Plastics	ρ (kg m^{-3})	Tensile Strength (N mm^{-2})	Elongation (%)	*E* (GN m^{-2})	BHN	Machinability
Thermoplastics						
PVC rigid	1330	48	200	3.4	20	Excellent
Polystyrene	1300	48	3	3.4	25	Fair
PTFE	2100	13	100	0.3	—	Excellent
Polypropylene	1200	27	200–700	1.3	10	Excellent
Nylon	1160	60	90	2.4	10	Excellent
Cellulose nitrate	1350	48	40	1.4	10	Excellent
Cellulose acetate	1300	40	10–60	1.4	12	Excellent
Acrylic (methacrylate)	1190	74	6	3.0	34	Excellent
Polyethylene (high density)	1450	20–30	20–100	0.7	2	Excellent
Thermosetting plastics						
Epoxy resin (glass filled)	1600–2000	68–200	4	20	38	Good
Melamine formaldehyde (fabric filled)	1800–2000	60–90	—	7	38	Fair
Urea formaldehyde (cellulose filled)	1500	38–90	1	7–10	51	Fair
Phenol formaldehyde (mica filled)	1600–1900	38–50	0.5	17–35	36	Good
Acetals (glass filled)	1600	58–75	2–7	7	27	Good

Note: BHN = Brinell hardiness number, ρ = density, *E* = Young's modulus.

TABLE 12.2.2 Relative Properties of Polymers

Material	Tensile Strength[a]	Compressive Strength[b]	Machining Properties	Chemical Resistance
Thermoplastics				
Nylon	E	G	E	G
PTFE	F	G	E	O
Polypropylene	F	F	E	E
Polystyrene	E	G	F	F
Rigid PVC	E	G	E	G
Flexible PVC	F	P	P	G
Thermosetting plastics				
Epoxy resin (glass-fiber filled)	O	E	G	E
Formaldehyde (asbestos filled)	G	G	F	G
Phenol formaldehye (Bakelite)	G	G	F	F
Polyester (glass-fiber filled)	E	G	G	F
Silicone (asbestos filled)	O	G	F	F

Note: O = outstanding, E = excellent, G = good, F = fair, P = poor.

[a] Tensile strength (typical): E = 55 Nmm^{-2}; P = 21 Nmm^{-2}.

[b] Compressive strength (typical): E = 210 Nmm^{-2}; P = Nmm^{-2}.

pipes, refrigerator liners, car-instrument surrounds, radiator grills, telephones, boat shells, and radio and television parts. Available in medium, high, and very high impact grades.

Cellulosics

"Cellulose nitrate" is inflammable and has poor performance in heat and sunlight. Its uses are therefore limited. Cellulose acetate has good strength, stiffness, and hardness and can be made self-extinguishing. Glass-filled grades are made. Cellulose acetate-butyrate (CAB) has superior impact strength, dimensional stability, and service temperature range and can be weather stabilized. Cellulose proprionate (CP) is similar to CAB, but has better dimensional stability and can have higher strength and stiffness. Ethyl cellulose has better low-temperature strength and lower density than the others. Processing of cellulose

TABLE 12.2.3 Selection Guide for Polymers by Application

Application or Service	Properties Required	Suitable Plastics
Chemical and thermal equipment	Resistance to temperature extremes and to wide range of chemicals; minimum moisture absorption; fair to good strength	Fluorocarbons, chlorinated polyether, polyvinylidene fluoride, polypropylene, high-density polyethylene, and epoxy glass
Heavily stressed mechanical components	High-tensile plus high-impact strength; good fatigue resistance and stability at elevated temperatures; machinable or moldable to close tolerance	Nylons, TFE-filled acetals, polycarbonates, and fabric-filled phenolics
Electrostructural parts	Excellent electrical resistance in low to medium frequencies; high-strength and -impact properties; good fatigue and heat resistance; good dimensional stability at elevated temperatures	Allylics, alkyds, aminos, epoxies, phenolics, polycarbonates, polyesters, polyphenylene oxides, and silicones
Low-friction applications	Low coefficient of friction, even when nonlubricated; high resistance to abrasion, fair to good form stability and heat and corrosion resistance	Fluorocarbons (TFE and FEP), filled fluorocarbons (TFE), TFE fabrics, nylons, acetals, TFE-filled acetals, and high-density polyethylenes
Light-transmission components, glazing	Good light transmission in transparent or translucent colors; good to excellent formability and moldability; shatter resistance; fair to good tensile strength	Acrylics, polystyrenes, cellulose acetates, cellulose butyrates, ionomers, rigid vinyls, polycarbonates, and medium-impact styrenes
Housings, containers, ducts	Good to excellent impact strength and stiffness; good formability and moldability; moderate cost; good environmental resistance; fair to good tensile strength and dimensional stability	ABS, high-impact styrene, polypropylene, high-density polyethylene, cellulose acetate butyrate, modified acrylics, polyester–glass and epoxy–glass combinations

plastics is by injection molding and vacuum forming. Applications include all types of moldings, electrical insulation, and toys.

Ethylene-Vinyl Acetate (EVA)

This material gives tough flexible moldings and extrusions suitable for a wide temperature range. The material may be stiffened by the use of fillers and is also specially formulated for adhesives. Applications include all types of moldings, disposable liners, shower curtains, gloves, inflatables, gaskets, and medical tubing. The material is competitive with polyvinyl chloride (PVC), polyethene, and synthetic rubbers, and is also used for adhesives and wax blends.

Fluorocarbons

This class of polymers, characterized by fluorine substitution, has outstanding chemical, thermal, and electrical properties and is characterized by the following four main classes of structures.

Polytetrafluoroethylene (PTFE), known commercially as Teflon or Fluon, is the best-known material and resists all known chemicals, weather, and heat, has an extremely low coefficient of friction, and is "non-stick." These materials are inert with good electrical properties. They are nontoxic, nonflammable, and have a working temperature range of –270 to 260°C. They may be glass filled for increased strength and rigidity. They do not melt and they must be formed by sintering of powders. Applications include chemical, mechanical, and electrical components, bearings (plain or filled with glass and/or bronze), tubing, and vessels for "aggressive" chemicals.

Fluoroethylenepropylene (FEP), unlike PTFE, can be processed on conventional molding machines and extruded, but thermal and chemical resistance properties are not quite as good.

Ethylenetetrafluoroethylene (ETFE) possess properties similar to but not as good as those of PTFE. However, the material exhibits a thermoplastic character similar to that of polyethylene which gives it a very desirable molding behavior.

TABLE 12.2.4 Properties of Polymers

Chemical class	Cellulose acetate	Cellulose acetate	Cellulose acetate butyrate	Cellulose acetate butyrate	Nylon	Polycarbonates	Polyethylene	Polyethylene
	Thermoplastic	Thermoplastic	Thermoplastic	Thermoplastic	Thermoplastic	Thermoplastic	Thermoplastic	Thermoplastic
	Soft	Hard	Soft	Hard	6/6	Unfilled	Low Density	Medium Density
ELECTRICAL PROPERTIES								
D.C. resistivity, ohm-cm	10^{10}-10^{13}	10^{10}-10^{13}	10^{10}-10^{12}	10^{10}-10^{12}		2×10^{16}	$>10^{15}$	$>10^{15}$
Dielectric constant, 60 cps	3.5-7.5	3.5-7.5	3.5-6.4	3.5-6.4	4.0-4.6	3.17	2.3-2.35	2.3
Dielectric constant, 10^6 cps	3.2-7.0	3.2-7.0	3.2-6.2	3.2-6.2	3.4-3.6	2.96	2.3-2.35	2.3
Dissipation factor, 60 cps	0.01-0.06	0.01-0.06	0.01-0.04	0.01-0.04	0.014-0.04	0.0009	<0.0005	<0.0005
Dissipation factor, 10^6 cps	0.01-0.10	0.01-0.10	0.01-0.04	0.01-0.04	0.04	0.01	<0.0005	<0.0005
MECHANICAL PROPERTIES								
Modulus of elasticity, 10^3 psi	86-250	190-400	74-126	150-200		290-325	14-38	35-90
Tensile strength, psi	1,900-4,700	4,600-8,500	1,900-3,800	5,6800	9,000-12,000	8,000-9,500	1,000-1,400	1,200-3,500
Ultimate elongation, %	32-50	6-40	60-74	38-54	60-300	20-100	400-700	50-600
Yield stress, psi	2,200-4,200	4,100-7,600	1,200-2,600	3,600-6,100		8,000-10,000	1,100-1,700	1,500-2,600
Yield strain, %							20-40	10-20
Rockwell hardness	R 49-R 103	R 101-R 123	R 59-R 95	R 108-R 117	R 108-R 120	M 70-M 180		
Notched Izod impact strength, ft lb/in.	2.0-5.2	0.4-2.7	2.5-5.4	0.7-2.4	1.0-2.0	8-16	No break	0.5->16
Specific gravity	1.27-1.34	1.27-1.34	1.15-1.22	1.19-1.25	1.13-1.15	1.2	0.91-0.925	0.926-0.941
THERMAL PROPERTIES								
Burning rate	Medium	Medium	Medium	Medium	Self-extinguishing	Self-extinguishing	Very slow	Slow
Heat distortion, 264 psi. C	44-57	60-113	49-58	70-99		135-145		
Specific heat, cal/g	0.3-0.42	0.3-0.42	0.3-0.4	0.3-0.4	0.4	0.3	0.55	0.55
Linear thermal expansion coefficient, 10^{-5}, C	8-16	8-16	11-17	11-17	8.0	6.6	10-20	14-16
Maximum continuous service temperature, C					80-150	138-143	60-77	71-93
CHEMICAL RESISTANCE								
Mineral acids, weak	Fair to good	Fair to good	Good	Good	Very good	Excellent	Good	Excellent
Mineral acids, strong	Poor	Poor	Fair to good	Fair to good	Poor	Fair	Good	Excellent
Oxidizing acids, concentrated	Very poor	Very poor			Poor		Good to poor	Good to poor
Alkalies, weak	Poor	Poor	Good	Good	No effect	Poor	Good	Excellent
Alkalies, strong	Very poor	Very poor	Poor	Poor	No effect	Poor	Good	Excellent
Alcohols	Poor	Poor	Poor	Poor	Good	Poor	Excellent to poor	Excellent to poor
Ketones	Poor	Poor	Poor	Poor	Good	Poor	Excellent to poor	Excellent to poor
Esters	Poor	Poor	Poor	Poor	Good	Poor	Excellent to poor	Excellent to poor
Hydrocarbons, aliphatic	Fair to good	Fair to good	Fair to good	Fair to good	Very good	Poor	Fair	Fair
Hydrocarbons, aromatic	Poor to fair	Poor to fair	Poor	Poor	Fair to good	Poor	Fair	Good
Oils: vegetable, animal, mineral	Fair to good	Fair to good	Good	Good	Good	Poor	Good	Excellent
MISCELLANEOUS PROPERTIES								
Clarity	Excellent	Excellent	Good to excellent	Good to excellent	Clear	Clear	Translucent	Translucent
Color	Pale to colorless	Pale to colorless	Pale to colorless	Pale to colorless	Pale amber to colorless	Colorless	Colorless	Colorless
Refractive index, n_D	1.46-1.50	1.46-1.50	1.46-1.49	1.46-1.49	1.53	1.60	1.50-1.54	1.52-1.54
FABRICATION								
Cl—calendering, Cs—casting, E—extrusion, F—hot forming or drawing, I—impregnation, MC—blow molding, MC—compression molding, MI—injection molding, S—spreading	Cs, E, F, MB, MC, MI, S	Cs, E, F, MB, MC, MI, S	Cs, E, F, MB, MC, MI, S	Cs, E, F, MB, MC, MI, S	E, F, MB, MC, MI	Cs, E, F, MB, MC, MI	Cl, E, F, MB, MC, MI	Cl, E, F, MB, MC, MI

TABLE 12.2.4 Properties of Polymers (continued)

Polyethylene	Polymethyl-methacrylate	Polypropylene	Polypropylene	Polystyrene	Polystyrene-acrylonitrile	Polytetrafluoro-ethylene	Polytrifluorochloro-ethylene	Polyvinylchloride and vinylchloride acetate	Polyvinylchloride and vinylchloride acetate	Epoxy
Thermoplastic	Thermoplastic	Thermoplastic	Thermoplastic	Thermoplastic	Thermoplastic	Thermoplastic	Thermoplastic	Thermoplastic	Thermoplastic	Thermosetting
High Density	Unmodified	Unmodified	Copolymer	Unmodified	Unmodified	Unmodified	Unmodified	Unmodified, rigid	Plasticized, non-rigid	Unfilled
$>10^{15}$	$>10^{14}$	$>10^{15}$	$>10^{17}$	$>10^{16}$	10^{13}-10^{17}	10^{18}	10^{18}	10^{12}-10^{16}	10^{11}-10^{14}	10^{12}-10^{14}
2.3-2.35	3.5-4.5	2.2-2.6	2.3	2.5-2.65	2.6-3.4	2.	2.2-2.8	3.2-4.0	5.0-9.0	3.5-5.0
2.3-2.35	3.0-3.5	2.2-2.6	2.3	2.5-2.65	2.5-3.1	2.	2.3-2.5	3.0-4.0	3.0-4.0	3.4-4.4
<0.0005	0.04-0.06	<0.0005	0.0001-0.0005	0.0001-0.0003	0.006-0.008	0.0002	0.001	0.01-0.02	0.03-0.05	0.001-0.005
<0.0005	0.02-0.03	0.0005-0.002	0.0001-0.002	0.0001-0.0004	0.008-0.01	0.0002	0.005	0.006-0.02	0.06-0.1	0.03-0.05
85-160	350-500	1.4-1.7		400-600	$>10^{16}$	33-65	150	200-600		>300
3,100-5,500	7,000-11,000	4,300-5,500	2,900-4,500	5,000-10,000	9,000-12,000	2,000-4,500	4,500-6,000	5,000-9,000	1,500-3,000	4,000-13,000
15-100	2.0-10	>220	200-700	1.0-2.5	1.0-2.5	200-400	250	2.0-40	200-400	2.0-6.0
2,400-5,000		4,900				1,600-2,000	4,200			
5-10		15				50-75	10	1.0-5.0		
R 30-R 50	M 80-M 105	93	R 50-R 96	M 65-M 85	M 75-M 90	D 50-D 65	J 75-J 95	R 110-R 120		M 75-M 110
1.5-20	0.3-0.6	1.0	1.1-12	0.25-0.60	0.3-0.6	2.5-4.0	2.5-4.0	0.4-2.0		0.2-1.0
0.941-0.965	1.18-1.20	0.90	0.90	1.04-1.08	1.05-1.1	2.1-2.3	2.1-2.3	1.36-1.4	1.15-1.35	1.115
Slow	Slow	Medium	Medium	Medium to slow	Slow	Self-extinguishing	Self-extinguishing	Self-extinguishing	Slow to self-extinguishing	Slow
	66-99				91-104	60		60-80		Up to 120
0.55	0.35	0.5	0.5	0.32-0.35	0.32-0.35	0.25	0.22	0.2-0.28	0.36-0.5	0.25-0.4
11-13	5.0-9.0	5.8-10	8-10	6.0-8.0	3.6-3.8	10	7.0	5.0-18	7.0-25	4.5-9.0
92-200	60-93		190-240	66-82	77-88	260	200	70-74	80-105	80
Excellent	Good	Excellent	Excellent	Excellent	Excellent	Excellent	Excellent	Excellent	Fair to good	Excellent
Excellent	Fair to poor	Excellent	Excellent	Excellent	Good to excellent	Excellent	Excellent	Good to excellent	Fair to good	Fair to good
Good to poor	Attacked	Good to poor	Poor	Poor	Poor	Excellent	Excellent	Fair to good	Poor to fair	Excellent
Excellent	Good	Excellent to good	Excellent	Excellent	Excellent	Excellent	Excellent	Excellent	Fair to good	Excellent
Excellent	Poor	Excellent to good	Good	Excellent	Good to excellent	Excellent	Excellent	Good	Fair to good	Excellent
Excellent to poor		Excellent to good	Good below 80 C	Excellent	Good to excellent	Excellent	Excellent	Excellent	Fair	Poor
Excellent to poor	Dissolves	Excellent to good	Good below 80 C	Dissolves	Dissolves	Excellent	Excellent	Poor	Poor	
Excellent to poor	Dissolves	Excellent to good	Good below 80 C	Poor	Dissolves	Excellent	Excellent	Poor	Poor	Excellent
Fair	Good	Good to fair	Good below 80 C	Poor	Good	Excellent	Excellent	Excellent	Poor	Excellent
Fair	Softens	Good to fair	Good below 80 C	Dissolves	Fair to good	Excellent	Excellent	Poor	Poor	Excellent
Good	Good	Good		Fair to poor	Good to excellent	Excellent	Excellent	Excellent	Poor	
Translucent		Transparent	Transparent	Transparent	Transparent	Translucent	Transparent	Transparent	Transparent	Transparent
Colorless	Colorless	Colorless to sl. yellow	Colorless to sl. yellow	Colorless	Colorless to amber	Colorless to gray	Colorless to pale	Colorless to amber	Colorless to amber	Colorless
1.54	1.48-1.50	1.49		1.59-1.60	1.56-1.57	1.30-1.40	1.43	1.54	1.50-1.55	1.58
Cl, E, F, MB, MC, MI	Cs, E, F, Lq, MB, MC, MI		Cl, E, F, MB, MC, MI	E, F, MB, MC, MI	Cl, E, F, MB, MC, MI	E, F, MC, MI	Cs, E, F, I, MC, MI, S	Cl, Cs, E, F, I, MB, MC, MI, S	Cl, Cs, E, MB, MC, MI, S	Cs, I, S

TABLE 12.2.4 Properties of Polymers (continued)

Properties of Polymers [continued]

Melamine-formaldehyde	Melamine-formaldehyde	Phenol-formaldehyde	Phenol-formaldehyde	Phenol-formaldehyde	Polyester (styrene-alkyd)	Silicones	Urea Formaldehyde	Acrylonitrile-butadiene-styrene (ABS)	Acetal	Alkyd resins
Thermosetting	Thermosetting	Thermosetting	Thermosetting	Thermosetting	Thermosetting	Thermosetting	Thermoplastic	Thermoplastic	Thermoplastic	Thermosetting
-Cellulose filled	Mineral filled (electrical	Cord filled	Cellulose filled	Unfilled cast phenolic, mechanical and chemical grade	Glassfiber mat reinforced	Mineral filled	-Cellulose filled	High-heat resistant	Homopolymer	Synthetic-fiber filled
10^{12}-10^{14}	10^{13}-10^{14}	10^{11}-10^{12}	10^{11}-10^{13}	$1.0\text{-}7.0 \times 10^{12}$	10^{11}	$>10^{12}$	0.5-5.0	2.4-5.0		3.8-5.0
7.9-9.4	10.2	7.0-10.0	5.0-9.0	6.5-7.5	4.0-5.5	3.5-3.6	7.7-9.5	2.4-3.8	3.7	3.6-4.7
7.2-8.4	6.1	5.0-6.0	4.0-7.0	4.0-5.5	4.0-5.5	3.4-3.6	6.7-8.0	0.003-0.008		0.012-0.026
0.03-0.08	0.10	0.1-0.3	0.04-0.3	0.10-0.15	0.01-0.04	0.004	0.036-0.043	0.007-0.015	0.004	0.01-0.016
0.03-0.043	0.051	0.04-0.09	0.03-0.07	0.04-0.05	0.01-0.06	0.005-0.007	0.025-0.035			
1,300	1,950	900-1,300	800-1,200	4.0-5.0	500-1,500		1,300-1,400	7,000-8,000	10,000-12,000	4,500-6,500
7,000-13,000	5,500-6,500	6,000-9,000	6,500-8,500	6,000-9,000	30,000-50,000	3,000-4,000	5,500-13,000	1.0-20	15-75	
0.6-0.9		0.5-1.0	0.6-1.0	1.5-2.0	0.5-1.5		0.6	4,000-9,000		10,000-13,000
M 110-M 124	E 90		M 110-M 120	M 93-M 120	M 80-M 120	M 85-M 95	E 94-E 97	R 110-R 115	M 94, R 120	E 76
0.24-0.35	0.3-0.4	4.0-8.0	0.24-0.34	0.25-0.4	7.0-30	0.25-0.35	0.24-0.40	2.0-4.0	1.4-2.3	0.50-4.5
1.47-1.52	1.78	1.36-1.43	1.32-1.55	1.307-1.318	1.5-2.1	1.8-2.8	1.47-1.52	1.06-1.08	1.43	1.24-2.6
Self-extinguishing	Self-extinguishing	Self-extinguishing	Self-extinguishing	Self-extinguishing	Self-extinguishing	Self-extinguishing	Self-extinguishing	Slow	Slow	Self-extinguishing
204	130	121-127	143-171	74-80	93-288	>260	130	115-118		
0.4			0.35-0.40		0.2-0.4	0.2-0.3	0.6	0.3-0.4	0.35	
2.0-5.7	2.1-4.3		3.0-4.5	6.0-8.0	1.8-3.0	2.0-4.0	2.2-3.6	6.0-6.5	8.1	4.0-5.5
99.0	149	121	149-177		121-204	288	77	88-110	84	149-220
Good	Fair	Variable	Variable	Fair to good	Good	Fair to good	Poor	Good	Fair	Good
Poor	Poor	Poor	Poor	Poor to good	Poor	Poor to good	Poor	Good	Poor	Fair
Poor	Poor	Poor	Poor	Poor	Poor		Poor	Poor	Poor	
Good	Fair	Variable	Variable	Poor to good	Good	Fair	Fair	Good	Poor	Good
Poor	Poor	Poor	Poor	Poor	Poor	Poor	Poor	Good	Poor	Fair
Good	Good	Good	Good to excellent	Good to excellent	Good	Poor	Good	Good	Good	Fair to good
Good	Good	Poor to fair	Fair	Fair	Poor	Poor	Good	Poor	Good	Fair to good
Good	Good	Fair to good	Fair to good	Fair to good	Good		Good	Poor	Good	Fair to good
Good	Good	Fair to good	Excellent	Good to excellent	Good	Fair to good	Good	Fair	Good	Fair to good
Good	Good	Fair to good	Excellent	Good	Poor to fair	Poor	Good	Fair	Good	Fair to good
Good	Good	Good	Excellent	Excellent	Good	Good	Good	Good	Good	
Translucent	Opaque	Opaque	Opaque	Clear	Translucent	Opaque	Translucent	Translucent to opaque	Translucent to opaque	Opaque
Colorless	Dark			Colorless to amber	Colorless	Pale to dark	Colorless	Colorless	Colorless	Colorless
							1.54-1.56		1.48	
MC	MC	MC	MC	Cs, F	I	MC	MC	CI, E, MB, MI	MI, E	Cs MC, MI

TABLE 12.2.4 Properties of Polymers (continued)

Properties of Polymers [continued]

Chemical class	Polyketone							
	aliphatic thermoplastic							
	neat resin							
ELECTRICAL PROPERTIES								
D.C. resistivity, ohm-cm	103							
Dielectric constant, 60 cps	5-6							
Dielectric constant, 10^6 cps								
Dissipation factor, 60 cps								
Dissipation factor, 10^6 cps	0.04							
MECHANICAL PROPERTIES								
Modulus of elasticity, 10^3 psi	250							
Tensile strength, psi	>9000							
Ultimate elongation, %	300							
Yield stress, psi	9000							
Yield strain, %	2.2							
Rockwell hardness								
Notched Izod impact strength, ft lb/in.	4							
Specific gravity	1.24							
THERMAL PROPERTIES								
Burning rate								
Heat distortion, 264 psi. C	105-110							
Specific heat, cal/g								
Linear thermal expansion coefficient, 10^{-5}, C	9-11							
Maximum continuous service temperature, C								
CHEMICAL RESISTANCE								
Mineral acids, weak	Good							
Mineral acids, strong	Fair							
Oxidizing acids, concentrated	Fair							
Alkalies, weak	Good							
Alkalies, strong	Good							
Alcohols	Excellent							
Ketones	Excellent							
Esters	Excellent							
Hydrocarbons, aliphatic	Excellent							
Hydrocarbons, aromatic	Excellent							
Oils: vegetable, animal, mineral								
MISCELLANEOUS PROPERTIES								
Clarity	translucent							
Color	white crystals							
Refractive index, n_D								
FABRICATION								
Cl—calendering, Cs—casting, E—extrusion, F—hot forming or drawing, I—impregnation, MC—blow molding, MC—compression molding, MI—injection molding, S—spreading	E, MI							

Perfluoroalkoxy (PFA) is the fourth group of fluorinated polymers. These materials have the same excellent properties as PTFE, but the compound is melt processible and, therefore, suitable for linings for pumps, valves, pipes, and pipe fittings.

Ionomers

These thermoplastics are based on ethylene and have high melt strength, which makes them suitable for deep forming, blowing, and other similar forming processes. They are used for packaging, bottles, moldings for small components, tool handles, and trim. They have a high acceptance of fillers.

Polymethylpentene

Polymethylpentene (TPX) is a high-clarity resin with excellent chemical and electrical properties and the lowest density of all thermoplastics. It has the best resistance of all transparent plastics to distortion at high temperature — it compares well with acrylic for optical use, but has only 70% of its density. It is used for light covers, medical and chemical ware, high-frequency electrical insulation, cables, microwave oven parts, and radar components. It can withstand soft soldering temperatures.

Polyethylene Terephthalate

Polyethylene terephthalate (PETP) and modified versions thereof have high strength, rigidity, chemical and abrasion resistance, impact resistance in oriented form, and a low coefficient of friction. It is attacked by acetic acid and concentrated nitric and sulfuric acids. It is used for bearings, tire reinforcement, bottles, automotive parts, gears, and cams.

Polyamides (Nylons)

The polyamides are a family of thermoplastics, e.g., Nylon 6, Nylon 66, and Nylon 610, which are among the toughest engineering plastics with high vibration-damping capacity, abrasion resistance, inherent lubricity, and high load capacity for high-speed bearings. They have a low coefficient of friction and good flexibility. Pigment-stabilized types are not affected by ultraviolet radiation and chemical resistance is good. Unfilled nylon is prone to swelling due to moisture absorption. Nylon bearings may be filled with powdered molybdenum disulfide or graphite. Applications include bearings, electrical insulators, gears, wheels, screw fasteners, cams, latches, fuel lines, and rotary seals.

Polyethylene

Low-density polyethylene (orinally called *polythene*) is used for films, coatings, pipes, domestic moldings, cable sheathing, and electrical insulation. High-density polyethylene is used for larger moldings and is available in the form of sheet, tube, etc. Polyethylene is limited as an engineering material because of its low strength and hardness. It is attacked by many oxidizing chemical agents and some hydrocarbon solvents.

Polyketone, Aliphatic

Aliphatic polyketones are relatively strong, tough, ductile polymeric resins derived from equal proportions of ethylene and carbon monoxide with an additional few percent of higher olefin for property and processibility adjustment. Their physical, thermal, and mechanical properties are similar to polyamides and polyacetals. Mechanical properties are characterized by preservation of high levels of stiffness, toughness, and strength over a broad temperature range. Resistance to hydrolysis, swelling, and permeation provides broad chemical resistance. Relatively new in commercial supply, they find application in gears, machine components, and similar engineering applications. Tribological performance is very good, and in particular they have a low coefficient of friction and a low wear factor against steel. The electrical properties of the neat polyketone are typical of those of polar, semicrystalline thermoplastics.

Polyethersulfone

Polyethersulfone is a high-temperature engineering plastic — useful up to 180°C in general and some grades have continuous operating ratings as high as 200°C. It is resistant to most chemicals and may

be extruded or injection molded to close tolerances. The properties are similar to those of nylons. Applications are as a replacement for glass for medical needs and food handling, circuit boards, general electrical components, and car parts requiring good mechanical properties and dimensional stability.

Polystyrene

This polymer is not very useful as an engineering material because of brittleness in unmodified forms, but it is well known for its use in toys, electrical insulation, refrigerator linings, packaging, and numerous commercial articles. It is available in unmodified form as a clear transparent resin and also in clear and opaque colors. High-impact forms are achieved by compounding with butadiene or other rubbery resins and heat-resistant forms are achieved by the use of fillers. Polystyrene can be stabilized against ultraviolet radiation and also can be made in expanded form for thermal insulation and filler products. It is attacked by many chemicals, notably aromatic hydrocarbon solvents, and by ultraviolet light.

Polysulfone

Polysulfone has properties similar to nylon, but these properties are retained up to 180°C compared with 120°C for nylon, which greatly expands the range of applications. Its optical clarity is good and its moisture absorption lower than that of nylon. Applications are as a replacement for glass for medical needs and chemistry equipment, circuit boards, and many electrical components.

Polyvinyl Chloride

This is one of the most widely used of all plastics. With the resin mixed with stabilizers, lubricants, fillers, pigments, and plasticizers, a wide range of properties is possible from flexible to hard types, in transparent, opaque, and colored forms. It is tough, strong, with good resistance to chemicals, good low-temperature characteristics and flame-retardant properties. PVC does not retain good mechanical performance above 80°C. It is used for electrical conduit and trunking, junction boxes, rainwater pipes and gutters, decorative profile extrusions, tanks, guards, ducts, etc.

Polycarbonate

Polycarbonate is an extremely tough thermoplastic with outstanding strength, dimensional stability, and electrical properties, high heat distortion temperature and low-temperature resistance (down to –100°C). It is available in transparent optical, translucent, and opaque grades (many colors). Polycarbonates have only fair resistance to chemicals as evidenced by the stress cracking caused by many solvents. The weathering tendencies can be stabilized against ultraviolet radiation by the use of proper additives. Polycarbonate compounds are used for injection moldings and extrusions for glazing panels, helmets, face shields, dashboards, window cranks, and gears. Polycarbonate is an important engineering plastic.

Polypropylene

Polypropylene is a low-density, hard, stiff, creep-resistant plastic with good resistance to chemicals, good wear resistance, low water absorption, and is relatively low cost. Polypropylene can be spun into filaments, converted into weaves, injection molded, and is commonly produced in a large variety of forms. Glass-filled polypropylene is widely used for its enhanced mechanical properties. It is used for food and chemical containers, domestic appliances, furniture, car parts, twine, toys, tubing, cable sheath, and bristles.

Polyphenylene Sulfide

Polyphenylene sulfide is a high-temperature plastic useful up to 260°C. Ambient temperature properties are similar or superior to those of nylon. It has good chemical resistance and is suitable for structural components subject to heat. Glass filler improves strength and enables very high heat resistance to 300°C. Uses are similar to those of nylon, but for higher temperatures.

Polyphenylene Oxide

This is a rigid engineering plastic similar to polysulfone in uses. It can be injection molded and has mechanical properties similar to those for nylon. It is used for automotive parts, domestic appliances, and parts requiring good dimensional stability. Frequently, the commercially available product is blended (or "alloyed") with polystyrene which acts as a cost-effective extender.

Thermosetting Polymers

Alkyds

There are two main groups of alkyds: diallyphthalate (DAP) and diallylisopthalate (DIAP). These have good dimensional stability and heat resistance (service temperature 170°C; intermittent use 260°C), excellent electrical properties, good resistance to oils, fats, and most solvents, but limited resistance to strong acids and alkalis. The mechanical properties are improved by filling with glass or minerals. The main uses are for electrical components and encapsulation. A wide range of colors and fast-curing grades are available.

Amino Resins

These are based on formaldehyde reacted with urea or melamine and are formulated as coatings and adhesives for laminates, impregnated paper textiles, and molding powders. The resins are usually compounded with fillers of cellulose, wood flour, and/or other extenders. As composites with open-weave fabric, they are used for building panels. Uses also include domestic electrical appliances and electric light fittings; the melamine type is used for tableware. The strength is high enough for use in stressed components, but the material is brittle. Electrical, thermal, and self-extinguishing properties are good.

Epoxies

Epoxy resins are used extensively across industry as engineering polymers and as adhesives. They can be cold cured without pressure using a "hardener" or may be heat cured. Inert fillers, plasticizers, flexibilizers, and extenders give a wide range of properties from soft flexible to rigid solid materials. Bonding to nearly all materials, e.g., wood, metal, glass, is excellent as are the mechanical, electrical, and chemical properties. Epoxies are used in all branches of engineering, including large castings, electrical parts, circuit boards, potting, glass and carbon fiber structures, flooring, protective coatings, and adhesives. Importantly, they exhibit little or no shrinkage on cure.

Phenolics (Phenol Formaldehyde, PF)

PF, the original "Bakelite," is usually filled with 50 to 70% wood flour for molded nonstressed or lightly stressed parts. Other fillers are mica for electrical parts, asbestos for heat resistance, glass fiber for strength and electrical properties, nylon, and graphite. Phenolics represent one of the best polymers for low-creep applications. Moldings have good strength, good gloss, and good temperature range (150°C wood filled; intermittent use 220°C), but are rather brittle. Applications include electrical circuit board, gears, cams, and car brake linings (when filled with asbestos, glass, metal powder, etc.). The cost is low and the compressive strength very high.

Polyester

Polyester resins can be cured at room temperature with a hardener or alone at 70 to 150°C. It is used unfilled as a coating, for potting, encapsulation, linings, thread locking, castings, and industrial moldings. It is used mostly for glass-reinforced-plastic (GRP) moldings.

Polyimides

Polyimides are noted for their unusually high resistance to oxidation and service temperatures up to 250°C (400°C for intermittent use). The low coefficient of friction and high resistance to abrasion make them ideal for nonlubricated hearings. Graphite or molybdenum disulfide filling improves these properties. They are used for high-density insulating tape. Polyimides have high strength, low moisture absorption, and resist most chemicals, except strong alkalis and ammonia solutions.

Silicones

These may be cold or heat cured and are used for high-temperature laminates and electrical parts resistant to heat (heat distortion temperature 450°C). Unfilled and filled types are used for special-duty moldings. Organosilicones are used for surface coatings and as a superior adhesive between organic and nonorganic materials.

Laminated Polymer Structures

A wide range of composite structures are prepared from polymer resins combined with fibers. The reader is referred to Section 12.6 for a more extensive discussion of polymer composites. Laminated polymer structures consist of layers of fibrous material impregnated with and bonded together usually by a thermosetting resin to produce sheets, bars, rods, tubes, etc. The laminate may be "decorative" or "industrial," the latter being of load-bearing mechanical or electrical grade.

Phenolics

Phenolic plastics can be reinforced with paper, cotton fabric, asbestos paper fabric or felt, synthetic fabric, or wood flour. They are used for general-purpose mechanical and electrical parts. They have good mechanical and electrical properties.

Epoxies

These are used for high-performance mechanical and electrical duties. Fillers used are paper, cotton fabric, and glass fiber.

Tufnol

"Tufnol" is the trade name for a large range of sheet, rod, and tube materials using phenolic resin with paper and asbestos fabric and epoxy resin with glass or fabric.

Polyester

This is normally used with glass fabric (the cheapest) filler. The mechanical and electrical properties are inferior to those of epoxy. It can be rendered in self-colors.

Melamine

Fillers used for melamine are paper, cotton fabric, asbestos paper fabric, and glass fabric. Melamines have a hard, nonscratch surface, superior electrical properties, and can be rendered in self-colors. They are used for insulators, especially in wet and dirty conditions, and for decorative and industrial laminates.

Silicone

Silicone is used with asbestos paper and fabric and glass fabric fillers for high-temperature applications (250°C; intermittent use 300°C). It has excellent electrical but inferior mechanical properties.

Polyimide

Polyimide is most often used with glass fabric as filler. Polyimides have superior thermal and electrical properties with a service temperature similar to that for silicones but with two to three times the strength and flexibility.

Foam and Cellular Polymers

Thermoplastics

Polyurethane Foams. The "flexible" type is the one most used. It is "open cell" and used for upholstery, underlays, thermal and vibration insulation, and buoyancy. It can be generated *in situ*. The rigid type has "closed cells" and is used for sandwich construction, insulation, etc. Molded components may be made from rigid and semirigid types.

Expanded Polystyrene. This material is produced only in rigid form with closed cells. It can be formed *in situ*. The density is extremely low, as is the cost. Chemical resistance is low and the service temperature is only 70°C. It is used for packaging, thermal and acoustic insulation, and buoyancy applications.

Cellular Polyvinyl Chlorides. The low-density type is closed cell and flexible. It is used for sandwich structures, thermal insulation, gaskets, trim, to provide buoyancy, and for insulating clothing. The moderate- to high-density open-cell type is similar to latex rubber and is used as synthetic leather cloth. The rigid closed-cell type is used for structural parts, sandwich construction, thermal insulation, and buoyancy. Rigid open-cell PVC (microporous PVC) is used for filters and battery separators. In general, cellular PVC has high strength, good flame resistance, and is easy to work.

Polyethylene Foams. The flexible type is closed cell and has low density with good chemical resistance and color availability, but is a poor heat insulator and is costly. The flexible foams are used for vibration damping, packaging, and gaskets. The rigid type has high density and is used for filters and cable insulation. A structural type has a solid skin and a foam core.

Ethylene Vinyl Acetates. These are microcellular foams similar to microcellular rubber foam, but are much lighter with better chemical resistance and color possibilities.

Other Types. Other types of thermoplastics include cellular acetate, which is used as a core material in constructions; expanded acrylics, which have good physical properties, thermal insulation, and chemical resistance; expanded nylon (and expanded ABS), which are low-density, solid-skin constructions; expanded PVA, which has similar properties to expanded polystyrene; and expanded polypropylene, which gives high-density foams.

Thermosets

Phenolics. These can be formed *in situ*. They have good rigidity, thermal insulation, and high service temperature, but are brittle.

Urea Formaldehyde (UF) Foam. This is readily formed *in situ* and has good thermal insulation. It has open pores and is used for cavity-wall filling.

Expanded Epoxies. These have limited use because of their high cost. They give a uniform texture and good dimensional stability and are used for composite forms, e.g., with polystyrene beads.

Silicon Foams. These are rigid and brittle with a high service temperature (300°C; 400°C intermittent use). Their use is limited to high-temperature-resistant sandwich constructions. The flexible closed-cell type is costly but will operate up to 200°C and is used for high-temperature seals and gaskets.

Elastomers

Cellular Rubbers. There are three types: *sponge*, solid rubber blown to give an open-cell structure, *foam*, a liquid rubber expanded to form open or closed cells which is stiffer than sponge; and *expanded*, a solid rubber blown with mainly closed cells, also it is stiffer than sponge. Uses include gaskets, seals, thermal insulation, cushioning, shock absorption, sound and vibration damping, buoyancy, and sandwich constructions.

Elastomers

Elastomers, or rubbers, are essentially amorphous polymers with linear chain molecules with some cross-linking, which ensures elasticity and the return of the material to its original shape when a load is removed. They are characterized by large strains (typically 100%) under stress. The synthetic rubber styrene butadiene is the most-used elastomer, with natural rubber a close second. The following describes the commonly used elastomers and gives some applications and properties.

Natural Rubbers (Polyisoprene, NR)

These elastomers have high strength, flexibility, and resilience, but have poor resistance to fuels, oils, flame, and sunlight aging. They are more costly than synthetic rubbers, which often replace them. "Soft-rubber" contains 1 to 4% sulfur as a vulcanizer. Wear resistance is increased by inclusion of fillers such as carbon black, silicon dioxide, clay, and wood flour. "Hard rubber" may contain up to 25% sulfur. Applications include vehicle tires and tubes, seals, antivibration mountings, hoses, and belts. Full vulcanization of 45% produces ebonite. Shore hardness: 30 to 90. Temperature range: –55 to 82°C.

Synthetic Rubbers

Styrene Butadiene Rubbers (SBR, GRS, BUNAS). These are similar to natural rubbers in application, but are usually inferior in mechanical properties, although cheaper. They are used in car brake hydraulic systems and for hoses, belts, gaskets, and antivibration mountings. Shore hardness: 40 to 80. Temperature range: –50 to 82°C.

Butadiene Rubbers (Polynutadiene, BR). These are used as substitutes for natural rubber, but are generally inferior. They have similar applications as natural rubber. Shore hardness: 40 to 90. Temperature range: –100 to 93°C.

Butyl Rubbers (Isobutylene Isoprene, GR 1). These are extremely resistant to water, silicon fluids and grease, and gas permeation. They are used for puncture-proof tires, inner tubes, and vacuum seals. Shore hardness: 40 to 90. Temperature range: –45 to 150°C.

Nitrile Rubbers (Butadiene Acrylonitrile, BUNA, N.NBR). These have good physical properties and good resistance to fuels, oils, solvents, water, silicon fluids, and abrasion. They are used for O rings and other seals, petrol hoses, fuel-pump diaphragms, gaskets, and oil-resistant shoe soles. Shore hardness: 40 to 95. Temperature range: –55 to 82°C.

Neoprene Rubbers (Polychloroprene, Chloroprene). These are some of the best general-purpose synthetic rubbers. They have excellent resistance to weather aging, moderate resistance to oils, and good resistance to refrigerants and mild acids. Shore hardness: 40 to 95. Temperature range: –40 to 115°C.

Chlorosulfonated Polyethylene Rubbers (CSM). These have poor mechanical properties but good resistance to acids and heat with complete resistance to ozone. They are used in chemical plants, tank linings, and high-voltage insulation. Shore hardness: 45 to 100. Temperature range: –100 to 93°C.

Ethylene Propylene Rubbers (EP, FPM). These specialized rubbers are especially resistant to weather aging heat, many solvents, steam, hot water, dilute acids and alkalis, and ketones, but not petrol or mineral oils. They are used for conveyor belts, limited automotive applications, silicone fluid systems, and electrical insulation. Shore hardness: 40 to 90. Temperature hardness: –50 to 177°C.

Fluorocarbon Rubbers. These comprise a wide range of rubbers with excellent resistance to chemical attack, heat, acids, fuels, oils, aromatic compounds, etc. They have a high service temperature. They are particularly suitable for vacuum applications. Shore hardness: 60 to 90. Temperature hardness: –23 to 260°C.

Isoprenes (Polyisoprene, IR). These are chemically the same as natural rubber but are more costly. The properties and applications are similar to those of natural rubber. Shore hardness: 40 to 80. Temperature hardness: –50 to 82°C.

Polyacrylic Rubbers (ACM, ABR). This is a group of rubbers with properties midway between nitrile and fluorocarbon rubbers with excellent resistance to mineral oils, hypoid oils, and greases and good resistance to hot air and aging. The mechanical strength is low. They are often used for spark plug seals and transmission seals. Shore hardness: 40 to 90. Temperature hardness: –30 to 177°C.

Polysulfide Rubbers. These have poor physical properties and heat resistance, but good resistance to oils, solvents, and weathering and are impermeable to gases and moisture. They are used for caulking and sealing compounds and as a casting material. Shore hardness: 40 to 85. Temperature hardness: –50 to 121°C.

Polyurethane Rubbers. These have exceptional strength and tear and abrasion resistance (the best of all rubbers), low-temperature flexibility, and good resistance to fuels, hydrocarbons, ozone, and weather. Resistance to solutions of acids and alkalis, hot water, steam, glycol, and ketones is poor. They are used for wear-resistant applications such as floor coverings. Shore hardness: 35 to 100. Temperature hardness: –53 to 115°C.

Silicone Rubbers (SI). These have exceptionally high service-temperature ranges, but the mechanical properties and chemical resistance are poor. They cannot be used in applications which expose them to fuels, light mineral oils, or high-pressure steam. They are used for high- and low-temperature seals, high-temperature rotary seals, cable insulation, hydraulic seals, and aircraft door and canopy seals. Shore hardness: 30 to 90. Temperature hardness: –116 to 315°C (380°C for intermittent use).

Fluorosilicone Rubbers. These are similar to silicone rubbers but have better oil resistance and a lower temperature range. Shore hardness: 40 to 80. Temperature hardness: –64 to 204°C.

12.3 Adhesives

Richard L. Lehman

Introduction

Adhesives are substances capable of holding materials together in a useful manner by surface attachment. The principal attribute of adhesives is their ability to form strong bonds with surfaces of a wide range of materials and to retain bond strength under expected use conditions. Although most adhesives do not have excellent bulk properties and it is therefore important to keep adhesive films thin, some materials such as epoxies have bulk properties which qualify them as engineering materials and thus can be used in multifunctional applications.

Advantages and Limitations of Use

The principal advantages of adhesives are their ability to bond similar to dissimilar materials of different thickness; to enable the fabrication of complex shapes not feasible by other fastening means; to smooth external joint surfaces; to permit economic and rapid assembly; to distribute stresses uniformly over joined interfaces, to provide weight reduction in critical structures via the elimination of fasteners; to dampen vibrations; to prevent or reduce galvanic corrosion; and to provide thermal and electrical insulation.

The limitations of adhesives depend on the specific adhesive and application and may include the necessity of surface preparation, long curing times, service-temperature limitations, loss of properties during service, toxicity of flammability during assembly or use, and the tendency of many adhesives to creep under sustained load.

Classes of Adhesives

Thermoplastic adhesives are a general class of adhesives based upon long-chained polymeric structure, and are capable of being softened by the application of heat with subsequent hardening upon cooling (hot-melt adhesives). The softening process if reversible for numerous cycles, which facilitates assembly and disassembly of structures. Thermosetting adhesives are a general class of adhesives based upon cross-linked polymeric structures which develop strong bonds that cannot be reversibly broken once they are formed. Thus, the thermoset adhesives are incapable of being softened once solidified.

Thermoplastic and thermosetting adhesives are cured, a process often referred to as setting, by polymerization or solidification, by heat, catalysis, chemical reaction, free-radical activity, radiation, evaporation of solvent, or another process as governed by the chemical nature of the particular adhesive.

Elastomers are a special class of thermoplastic adhesive possessing the common quality of substantial flexibility or elasticity. Refer to Section 12.2 on polymers.

Anaerobic adhesives are a special class of thermoplastic adhesive, the polyacrylates, that set only in the absence of air (oxygen). The two basic types are (1) machinery — possessing shear strength only and (2) structural — possessing both tensile and shear strength.

Pressure-sensitive adhesives are permanently and aggressively tacky solids which form immediate bonds when two parts are brought together under pressure. They are available as films and tapes as well as hot-melt systems.

Performance of Adhesives

To obtain optimum mechanical performance of an adhesive, it is critical to select the proper compound for the target application. Table 12.3.1 illustrates compatibility of adhesives and five broad classes of common materials. Generally, for good adhesive bonds the chemistry of the adhesive must match or be

TABLE 12.3.1 Relative Performance of Adhesive Resins

Adhesive Resin	Adherence To:					Resistance			
	Paper	Wood	Metal	Ceramics	Rubbers	Water	Solvents	Alkali	Acids
Alkyd	6	7	5	6	7	7	2	2	5
Cellulose acetate	4	3	1	3	5	2	3	1	3
Cellulose acetate butyrate	3	1	4	5	2	3	1	3	3
Cellulose nitrate	5	1	5	5	3	2	2	4	4
Ethyl cellulose	3	1	3	5	2	3	3	3	3
Methyl cellulose	1	1	3	3	1	6	3	3	3
Carboxy methyl cellulose	1	2	3	2	1	6	1	4	4
Epoxy resin	10	8	8	8	8	7	8	8	8
Furane resin	7	2	8	7	8	9	7	8	8
Melamine resin	10	5	2	5	4	9	5	5	5
Phenolic resins	8	5	5	7	6	10	7	8	8
Polyeseter, unsaturated	8	4	5	7	7	6	2	5	7
Polyethylacrylate	4	3	5	6	8	4	6	7	7
Polymethylmethacrylate	4	4	3	6	6	5	6	7	7
Polystyrene	3	2	2	5	8	5	5	8	8
Polyvinylacetate	7	7	7	3	3	3	4	6	6
Polyvinyl alcohol	2	2	4	6	1	7	1	3	3
Polyvinyl acetyl	7	8	7	7	8	5	3	5	5
Polyvinyl chloride	7	6	7	6	8	6	10	9	9
Polyvinyl acetate chloride	8	6	7	5	8	5	8	8	5
Polyvinylidene copolymer	7	6	7	7	8	7	10	9	9
Silicone T.S.	6	7	7	8	10	7	6	6	6
Urethane T.S.	10	10	9	10	7	8	4	4	4
Acrylonitrile rubber	6	8	6	9	7	5	8	8	8
Polybutene rubber	3	6	2	8	8	3	10	9	9
Chlorinated rubber	5	7	4	7	6	3	10	9	9
Styrene rubber	7	6	5	8	7	3	8	9	9

Note: 1 = low performance; 10 = high performance.
Source: Adapted from Simonds, H.R. and Church, J.M., *A Concise Guide to Plastics*, 2nd ed., Reinhold, New York, 1963. With permission.

TABLE 12.3.2 High-Performance Engineering and Machine Part Adhesives

Thread locking	• Anaerobic acrylic
Hub mounting	• Anaerobic acrylic — compatible materials or flow migration unimportant
	• Modified acrylic — large gaps or migration must be avoided
	• Epoxy — maximum strength at high temperatures
Bearing mounting	• Anaerobic acrylic — compatible materials necessary and flow into bearing area to be prevented
	• Modified acrylic — for lowest cost
Structural joining	• Epoxies and modified epoxies — for maximum strength (highest cost)
	• Acrylics — anaerobic or modified cyanacrylates
Gasketing	• Silicones — primarily anaerobic

similar to the surface energy, polarity, and/or chemistry of the material being bonded. The elastic modulus of the adhesive should not be greater than the bonded material.

Adhesives are used in two classes of application, those requiring only shear strength and those requiring structural properties, often tensile and shear strength. Table 12.3.2 provides a quick reference for some typical applications. A much more detailed summary and classification of adhesives is given in Table 12.3.3.

Table 12.3.3 presents a sample of a number of adhesives (with practical information) that are available from various sources. The table is adapted from the rather extensive one found in J. Shields, *Adhesives Handbook*, CRC Press, Boca Raton, FL, 1970. For other extensive lists of trade sources, the reader is

TABLE 12.3.3 Properties and Applications of Adhesive Materials

Basic Type	Curing Cycle, Time at Temperature	Service Temperature Range, C	Adherends	Main Uses	Remarks
			Animal		
Animal (hide)	Melted at 70–75°C; sets on cooling	<70	Paper, wood, textiles	Woodworking, carpet materials, paper, bookbinding	May be thinned with water
Animal (hide) + plasticizers	Applied as a melt at 60°C	<60	Paper, cellulosic materials	Bookbinding, stationery applications	Cures to permanent flexible film
Fish glue	1 hr at 20°C	<60	Wood, chipboard, paper	General-purpose for porous materials	Rapid setting; good flexibility, moderate resistance to water; high tack
Casein	Cold setting after 20 min standing period on mixing	<50	Timber with moisture content	Laminated timber arches and beams, plybox beams, and engineering timber work	Full bond strength developed after seasoning period of 48 hr
Casein + 60% latex	Cold setting after 20 min standing period on mixing	<60	Aluminum, wood, phenolic formaldehyde (rigid, leather, rubber)	Bonding of dissimilar materials to give flexible, water-resistant bond	Flexible
			Vegetable		
Dextrine	Air drying		Paper, cardboard, leather, wood, pottery	General-purpose glue for absorbent materials	Medium drying period of 2–3 hr
Dextrine–starch blend	Applied above 15°C air drying	<48	Cellulosic materials, cardboard, paper	Labeling, carton sealing, spiral-tube winding	Fast setting; may be diluted with water
Gum arabic	Cold setting	<50	Paper, cardboard	Stationery uses	Fast drying
			Mineral		
Silicate	8 hr at 20°C	10–430	Asbestos, magnesia	Lagging asbestos cloth on high-temperature insulation	Unsuitable where moisture; not recommended for glass or painted surfaces

TABLE 12.3.3 Properties and Applications of Adhesive Materials (continued)

Basic Type	Curing Cycle, Time at Temperature	Service Temperature Range, C	Adherends	Main Uses	Remarks
Silicate with china-clay filler	Dried at 80°C before exposure to heat	–180–1500	Asbestos, ceramics, brickwork, glass, silver, aluminum, steel (mild)–steel	General purpose cement for bonding refractory materials and metals; furnace repairs and gastight jointing of pipe work; heat-insulating materials	Resistant to oil, gasoline, and weak acids
Sodium silicate	Dried at 20–80°C before exposure to heat	0–850	Aluminum (foil), paper, wood–wood	Fabrication of corrugated fiber board; wood bonding, metal foil to paper lamination	Suitable for glass-to-stone bonding
Aluminum phosphate + silica filler	Dried $^1/_2$ hr at 20°C, then $^1/_2$ hr at 70°C + $^1/_2$ hr at 100°C + 1 hr at 200°C + 1 hr at 250°C; repeat for two overcoatings and finally cure at 1 hr at 350°C	<750	Steels (low-alloy), iron, brass, titanium, copper, aluminum	Strain-gauge attachment to heat-resistant metals; heater-element bonding	Particularly suited to heat-resistant steels where surface oxidation of metal at high temperatures is less detrimental to adhesion
Bitumen/latex emulsion	Dried in air to a tacky state	0–66	Cork, polystyrene (foam), polyvinyl chloride, concrete, asbestos	Lightweight thermal-insulation boards, and preformed sections to porous and nonporous surfaces; building applications	Not recommended for constructions operated below 0°C
			Elastomers		
Natural rubber	Air dried 20 min at 20°C and heat-cured 5 min at 140°C	<60	Rubber (styrene butadiene), rubber (latex), aluminum, cardboard, leather, cotton	Vulcanizing cement for rubber bonding to textiles and rubbers	May be thinned with toluene
Natural rubber in hydrocarbon solvent	Air dried 10 min at 20°C and heat-cured for 20 min at 150°C	<100	Hair (keratin), bristle, polyamide fiber	Brush-setting cement for natural- and synthetic-fiber materials	Resistant to solvents employed in oil, paint and varnish industries, can be nailed without splitting
Rubber latex	Air drying within 15 min	<60	Canvas, paper, fabrics, cellulosic materials	Bonding textiles, papers packaging materials; carpet bonding	Resistant to heat; should be protected from frosts, oils

TABLE 12.3.3 Properties and Applications of Adhesive Materials (continued)

Basic Type	Curing Cycle, Time at Temperature	Service Temperature Range, C	Adherends	Main Uses	Remarks
Chlorinated rubber in hydrocarbon solvents	Air dried 10 min at 20°C and contact bonded	–20–60	Polyvinyl chloride, acrylonitrile butadiene styrene, polystyrene, rubber, wood	General-purpose contact adhesive	Resistant to aging, water, oils, petroleum
Styrene–butadene rubber lattices	Air drying	–20–60	Polystyrene (foam), wood, hardboard, asbestos, brickwood	Bonding polystyrene foams to porous surface	—
Neoprene/nitrile rubbers in (?)	Dried 30 min in air and bonded under pressure; tacky	–20–60	Wood, linoleum, leather, paper, metals, nitrile rubbers, glass, fabrics	Cement for bonding synthetic rubbers to metals, woods, fabrics	May be thinned with ketones
Acrylonitrile rubber + phenolic resin	Primer air dried 60 min at 20°C, film cured 60 min at 175°C under pressure; pressure released on cooling at 50°C	–40–130	Aluminum (alloy)–aluminum to DTD 746	Metal bonding for structural applications at elevated temperatures	Subject to creep at 150°C for sustained loading
Polysulfide rubber in ketone solvent and catalyst	3 days at 25°C	–50–130, withstands higher temperatures for short periods	Metals	Sealant for fuel tanks and pressurized cabins in aircraft, where good weatherproof and waterproof properties are required	Resistant to gasoline, oil, hydraulic fluids, ester lubricants; moderate resistance to acids and alkalis
Silicone rubber	24 hr at 20°C (20% R.H.); full cure in 5 days	–65–260	Aluminum, titanium, steel (stainless), glass, cork, silicone rubber, cured rubber–aluminum, cured rubber–titanium, cured rubber–steel (stainless), aluminum–aluminum (2024 Alclad), cork–cork (phenolic bonded)	General-purpose bonding and scaling applications; adhesive/sealant for situations where material is expected to support considerable suspended weight; high-pressure exposure conditions	Resistant to weathering and moisture
Reclaim rubber	Contact bonded when tacky	<50	Fabric, leather, wood, glass, metals (primed)	General industrial adhesive for rubber, fabric, leather, porous materials	May be thinned with toluene
Polychloroprene	Air dried 10–20 min at 20°C	<60	Rubber, steel, wood, concrete	Bonding all types of rubber flooring to metals, woods, and masonry	Good heat resistance

TABLE 12.3.3 Properties and Applications of Adhesive Materials (continued)

Basic Type	Curing Cycle, Time at Temperature	Service Temperature Range, C	Adherends	Main Uses	Remarks
Modified polyurethane	3 hr at 18°C to 16 hr at –15°C	–80–110	Concrete, plaster, ceramics, glass, hardboards, wood, polyurethane (foam), phenol formaldehyde (foam), polystyrene (foam), copper, lead, steel, aluminum	Bonding to rigid and semirigid panels to irregular wall surfaces, wall cladding and floor laying; building industry applications	Foam remains flexible on aging even at elevated temperatures; will withstand a 12% movement
			Thermoplastic		
Nitrocellulose in ester solvent	Heat set 1 hr at 60°C after wet bonding	60	Paper, leather, textiles, silicon carbide, metals	Labeling, general bonding of inorganic materials including metals	Good resistance to mineral oils
Modified methyl cellulose	Dries in air	<50	Vinyl-coated paper, polystyrene foam	Heavy-duty adhesive; decorating paper and plastics	Contains fungicide to prevent biodeterioration
Ethylene vinyl acetate copolymer + resins	Film transfer at 70–80°C followed by bonding at 150–160°C	60, or 1 hr at 90	Cotton (duck)–cotton, resin rubber–leather, melamine laminate–plywood, steel (mild)–steel, acrylic (sheet) acrylic	Metals, laminated plastics, and textiles; fabrication of leather goods; lamination work	Good electrical insulation
Polyvinyl acetate	Rapid setting	<60	Paper, cardboard	Carton sealing in packaging industry	Resistant to water
Synthetic polymer blend	Applied as a melt at 177°C	<70	Paper, cardboard, polythene (coated materials)	Carton and paperbag sealing; packaging	—
Polychloroprene/resin blend in solvent	Air dried 10 min at 20°C and cured 4 days at 20°C to 7 hr at 75°C	<70	Chlorosulfonated polythene, polychloroprene fabrics, polyamide fabrics, leather, wood, textiles	Bonding synthetic rubbers and porous materials; primer for polyamide-coated fabrics such as nylon, terylene	—
Polychloroprene	Air dried 10–20 min at 20°C	—	Rubber, steel, wood, concrete	Bonding all types of rubber flooring to metals, woods, and masonry	Good heat resistance
Saturated polyester + isocyanate catalyst in ethyl acetate	Solvent evaporation and press cured at 40–80°C when tacky	—	Cellulose, cellulose acetate, polyolefins (treated film), polyvinyl chloride (rigid), paper, aluminum (foil), copper (foil)	Lamination of plastic films to themselves and metal foils for packaging industry, printed circuits	Resistant to heat, moisture, and many solvents

TABLE 12.3.3 Properties and Applications of Adhesive Materials (continued)

Basic Type	Curing Cycle, Time at Temperature	Service Temperature Range, C	Adherends	Main Uses	Remarks
Cyanoacrylate (anaerobic)	15 sec to 10 min at 20°C substrate dependent	Melts at 165	Steel–steel, steel–aluminum, aluminum–aluminum, butyl rubber–phenolic	Rapid assembly of metal, glass, plastics, rubber components	Anaerobic adhesive. Curing action is based on the rapid polymerization of the monomer under the influence of basic catalysts; absorbed outer layer on most surfaces suffices to initiate polymerization and brings about bonding
Polyacrylate resin (anaerobic)	3 min at 120°C to 45 min at 65°C or 7 days at 20°C	–55–95	Aluminum–aluminum	Assembly requirements requiring high resistance to impact or shock loading; metals, glass and thermosetting plastics	Anaerobic adhesive
			Thermosetting		
Urea formaldehyde	9 hr at 10°C to 1 hr at 21°C after mixing powder with water (22%)	<90	Wood, phenolic laminate	Wood gluing and bonding on plastic laminates to wood; plywood, chipboard manufacture; boat building and timber engineering	Excess glue may be removed with soapy water
Phenolic formaldehyde + catalyst PX-12	Cold acting	<100	Wood	Timber and similar porous materials for outdoor-exposure conditions; shop fascia panels	Good resistance to weathering and biodeterioration
Resorcinol formaldehyde + catalyst RXS-8	Cured at 16–80°C under pressure	<100	Wood, asbestos, aluminum, phenolic laminate, polystyrene (foam), polyvinyl chloride, polyamide (rigid)	Constructional laminates for marine craft; building and timber applications; aluminum–plywood bonding; laminated plastics	Recommended for severe outdoor-exposure conditions
Epoxy resin + catalyst	24–48 hr at 20°C to 20 min at 120°C	100	Steel, glass, polyester–glass fiber composite, aluminum–aluminum	General-purpose structural adhesive	—

TABLE 12.3.3 Properties and Applications of Adhesive Materials (continued)

Basic Type	Curing Cycle, Time at Temperature	Service Temperature Range, C	Adherends	Main Uses	Remarks
Epoxy resin + catalyst	8 hr at 24°C to 2 hr at 66°C to 45 min at 121°C	65	Steel, copper, zinc, silicon carbide, wood, masonry, polyester–glass fiber composite, aluminum–aluminum	Bonding of metals, glass, ceramics, and plastic composites	Cures to strong, durable bond
Epoxy + steel filler (80% w/w)	1–2 hr at 21°C	120	Iron, steel, aluminum, wood, concrete, ceramics, aluminum–aluminum	Industrial maintenance repairs; metallic tanks, pipes, valves, engine castings, castings	Good resistance to chemicals, oils, water
Epoxy + amine catalyst (ancamine LT)	2–7 days at 20°C for 33% w/w catalyst content	–5–60	Concrete stonework	Repair of concrete roads and stone surfaces	Excellent pigment-wetting properties; effective underwater and suited to applications under adverse wet or cold conditions
Epoxy resin (modified)	4–5 hr at 149°C to 20 min at 230°C to 7 min at 280°C	150	Aluminum, steel, ceramics	One-part structural adhesive for high-temperature applications	Good gap-filling properties for poorly fitting joints; resistant to weather, galvanic action
Epoxy	45 sec at 20°C	—	Gem stones, glass, steel, aluminum–aluminum	Rapid assembly of electronic components, instrument parts, printed circuits; stone setting in jewelry and as an alternative to soldering	—
Epoxy resin in solvent + catalyst	8 hr at 52°C to $^1/_2$ hr at 121°C	–270–371	Aluminum and magnesium alloys for elevated-temperature service	Strain gauges for cryogenic and elevated-temperature use; micromeasurement strain gauges	Cured material resists outgassing in high vacuum
Epoxy polyamide	8 hr at 20°C to 15 min at 100°C	100	Copper, lead, concrete, glass, wood, fiberglass, steel–steel aluminum–aluminum	Metals, ceramics, and plastics bonding; building and civil engineering applications	Resists water, acids, oils, greases

TABLE 12.3.3 Properties and Applications of Adhesive Materials (continued)

Basic Type	Curing Cycle, Time at Temperature	Service Temperature Range, C	Adherends	Main Uses	Remarks
Epoxy/polysulfide	24 hr at 20°C to 3 hr at 60°C to 20 min at 100°C	<120	Asbestos (rigid), ceramics, glass–fiber composites, carbon, polytetrafluoroethylene (treated), polyester (film), polystyrene (treated), rubber (treated), copper (treated), tungsten carbide, magnesium alloys, aluminum–aluminum, steel (stainless)–steel	Cold-setting adhesive especially suitable for bonding materials with differing expansion properties	Cures to flexible material; resistant to water, petroleum, alkalis, and mild acids
Phenol furfural + acid catalyst	2 days at 21°C	90–140	Alumina, carbon (graphite)	Formulation of chemically resistant cements; bedding and joining chemically resistant ceramic tiles	Extremely resistant to abrasion and heat
			Pressure-sensitive		
	Heated by air drying for several hours or 15–30 min at 210°F	—	Teflon–Teflon, Teflon–metal	—	Good resistance to acids and alkalis; excellent electrical properties
			Miscellaneous		
Ceramic-based	Dried for 1/2 hr at 77°C and cured 1/2 hr at 200°C + 1 hr at 250°C; postcured, 1 hr at 350°C	816	Metals	Strain gauges, temperature sensors for elevated-temperature work	—

referred to Charles V. Cagle, Ed., *Handbook of Adhesive Bonding*, McGraw-Hill, New York, 1972, and Lerner et al., *Adhesives Red Book*, Palmerton Publishing Co., New York, 1968.

References

ANSI/ASTM Standards D896-D3808, 1996 Annual Book of Standards, ASTM, Conshohocken, PA.

Avallone, Ed., *Mark's Standard Handbook for Mechanical Engineers*, 9th ed., McGraw-Hill, New York, 6-141–6-148.

Bikerman, 1968. *The Science of Adhesive Joints*, Academic Press, New York.

Cagle, C.V., Ed., 1972. *Handbook of Adhesive Bonding*, McGraw-Hill, New York.

Cook, 1970. *Construction Sealants and Adhesives*, John Wiley & Sons, New York.

Lerner, Kotsher, and Sheckman, 1968. *Adhesive Red Book*, Palmerton Publishing Co., New York.

NASA SP-5961 (01) Technology Utilization, *Chemistry Technology: Adhesives and Plastics*, National Technical Information Services, VA.

Patrick, 1973. *Treatise on Adhesives*, Marcel Dekker, New York.

Shields, J. 1970. *Adhesives Handbook*, CRC Press, Boca Raton, FL.

Simonds, H.R. and Church, J.M. 1963. *A Concise Guide to Plastics*, Reinhold, New York.

12.4 Wood

Daniel J. Strange

Definition

Wood is the structural component of a tree. It is composed of dead cells which were originally formed near the cambium (just beneath the bark). As the tree grows, the cambium moves outward, leaving the dead cells behind to serve as structure.

The two broad classifications for types of wood and softwoods and hardwoods. Softwoods come from conifers while hardwoods come from deciduous trees, and as the name implies softwoods are generally softer and hence weaker and with lower elastic modulus than hardwoods, although this generalization is not universally true.

The wood closest to the bark is called sapwood, and this layer extends about an inch into the tree. Although most of the cells in this layer are dead, this is the layer which transports moisture to the rest of the tree by capillary action. Beneath the sapwood layer is the heartwood, which is almost totally inactive except to provide structural support.

Composition

Wood is a fibrous cellular material with the cell walls composed primarily of cellulose, hemicellulose, and lignin. Cellulose is a linear polymer of glucose units and is the single most common organic chemical in nature. It accounts for roughly 40 to 50% of the wood by weight. Hemicellulose is a modified form of cellulose comprising approximately 30% of the cell wall whose structure can vary depending upon the exact type of wood. Lignin is a complex three-dimensional phenolic polymer which composes 20 to 30% of the structure of the wood. The rest of the weight of the wood is composed primarily of extraneous extractive substances which reside within the cellular structure and affect properties such as specific gravity, moisture absorption, and durability.

Variations in the compositions, structures, and volumes of these four components can have dramatic effects on the properties of the wood. These properties are relatively constant within a species, although growing conditions can have a significant influence.

Mechanical Properties

The properties of wood can vary substantially depending upon moisture content. “Green” wood, wood taken directly from a living tree with associated high moisture level, is significantly weaker (lower Young’s modulus, tensile strength, and compressive strength) than oven-dried wood. Typical ratios of dry to green wood strength properties vary from 1.2 to 1.8.

The failure modes of wood are more complex than one might expect. Because wood is essentially a composite material, it does not follow Hooke’s law at high stresses but instead exhibits viscoelastic behavior. Figure 12.4.1 shows typical stress–strain curves for wood in tension and compression. At very high stresses, creep can occur. Wood is strongest in tension parallel to the grain, on the order of 100 MPa for a typical softwood. Perpendicular to the grain this value drops to about 4 MPa. In compression the strength values are approximately half the tension values, due to the collapse of the cellular structure and buckling of the wood fibers.

There are three principal shear failure modes in wood, six if one distinguishes between the radial and tangential directions. These modes are shear perpendicular to grain, shear parallel to grain, and rolling shear. Rolling shear occurs when the failure plane is parallel to the grain but the sliding direction is perpendicular to the grain, hence the fibers “roll” over each other. Wood is strongest when the shear is perpendicular to the grain, and weakest in rolling shear.

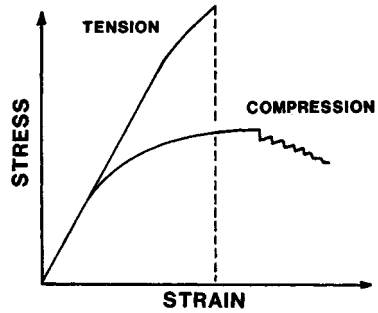


FIGURE 12.4.1 Stress–strain behavior of wood in tension and compression. Stress–strain curves for wood in tension and compression parallel to grain. Signs of stress and strain have been ignored so that both may be plotted together. Note particularly the difference in strength and in the extent of the nonlinear deformation prior to maximum stress.

Decay Resistance

Totally dry wood does not decay. Furthermore, wood kept completely submerged will not decay significantly. Wood decays most rapidly in warm, humid, low-altitude areas. Some species of wood are more decay resistant than others as a result of the presence of extractives. White oak, walnut, cherry, cedar, and yew are examples of highly decay-resistant woods. Pines, willows, elms, beeches, and spruces are examples of low decay-resistant woods.

Composites

By gluing wood plies, chips, fibers, or pulp together a composite material can be formed which has more isotropic and homogenous properties than regular timber. In addition, this technique creates a strong and durable product out of wood unsuitable for timber. In general, wood composites homogenize the extreme anisotropy of timber into a nearly isotropic material whose properties are an average of the properties in each direction of the original timber. Table 12.4.1 classifies the various wood composites based upon the constitutive particle and the binder.

The most common wood composite is paper. Paper is made from pulped and chemically treated cellulose fibers, which is then rolled into sheets, pressed, and dried. There is no glue involved, as the microfibrils of the cellulose interlock and form hydrogen bonds. Fiberboard is similar to paper, only thicker (by an arbitrary value, typically 0.012 in.) and with larger fiber bundles. *Hardboard* simply refers to a high-density fiberboard.

Wood chips and/or sawdust pressed and glued together is referred to as *particleboard*. Typically, the chips range in size from 10 to 300 mm long and are in the form of flakes or fibers. Particleboard is effectively isotropic, easily machineable, and inexpensive. It is often used in furniture and for floor underlayment. When larger wood chips are used the product is referred to as *flakeboard*.

Plywood is created by layering plies of radial-cut wood. Typically, the plies are oriented at 90° to each other, which results in a strong material when the stress in parallel or perpendicular to the grain of the face plies, but a lesser strength at any other angle. Some design strength specifications of plywood are given in Table 12.4.2. Plywood makes efficient use of timber due to its radial cut, minimizes the effects of imperfections, resists warping, and can be formed into large sheets. Plywood is also less expensive than clear lumber.

The reader is referred to Table C.14 in the Appendix for an overview of the physical properties and uses of American woods. The table also contains recommendations for appropriate applications. Table E.1 in the Appendix gives nominal sizes for lumber and timber, as well as allowable stresses in tension and compression and moduli of elasticity of various kinds of woods.

TABLE 12.4.1 Allowable Stresses for Plywood (Stresses in MPa)

Type of Stress	Species Group of Face Ply	Grade Stress Level				
		S-1		S-2		S-3
		Wet	Dry	Wet	Dry	Dry Only
Tension in plane of plies	1	9.86	13.79	8.20	11.38	11.38
(at 45° to face use 1/6	2,3	6.76	9.65	5.65	8.27	8.27
value)	4	6.48	9.17	5.38	7.65	7.65
Compression in plane of	1	6.69	11.31	6.21	10.62	10.62
plies (at 45° to face use	2	5.03	8.27	4.69	7.58	7.58
1/3 value)	3	4.21	7.31	4.00	6.83	6.83
	4	4.21	6.89	4.00	6.55	6.55
Shear in plane	1	1.41	1.72	1.41	1.72	1.45
perpendicular to plies	2,3	1.10	1.28	1.10	1.28	1.10
(45° use 2 × value)	4	1.00	1.21	1.00	1.21	1.07
Shear, rolling, in the	Marine and	0.43	0.52	0.43	0.52	—
plane of plies (at 45° to	Structural I					
face grain use $1^1/_3$	Structural II	0.34	0.39	0.34	0.39	0.38
value)	and 2.4.1					
	All Other	0.30	0.37	0.30	0.37	0.33
Modulus of rigidity	1	480	620	480	620	570
(shear in plane	2	410	520	410	520	470
perpendicular to plies)	3	350	410	350	410	380
	4	310	350	310	350	310
Bearing (on face)	1	1.45	2.34	1.45	2.34	2.34
(perpendicular to plane	2,3	0.93	1.45	0.93	1.45	1.45
of plies)	4	0.72	1.10	0.72	1.10	1.10
Modulus of elasticity in	1	10,300	12,400	10,300	12,400	12,400
bending in plane of	2	9,000	10,300	9,000	10,300	10,300
plies (face grain	3	7,600	8,300	7,600	8,300	8,300
parallel or ⊥ to span)	4	6,200	6,900	6,200	6,900	6,900

Adapted from American Plywood Association, Plywood Design Specifications, 1976.

TABLE 12.4.2 Classification of Wood Composites

Material	Constitutive Particle	Binder
Wood flour molding	Wood flour	Plastic; synthetic resin
Fiber-reinforced plastic	Fiber	Plastic
Paper	Fiber segment; fiber	Cellulose; hemicellulose; synthetic resin
Fiberboard	Fiber segment; fiber; fiber bundle	Lignin; synthetic resin
Particleboard	Splinter; chip; flake; planer shaving	Synthetic resin
Plywood	Veneer	Synthetic resin; natural glue
Laminated wood	Lumber	Synthetic resin; natural glue; mechanical connector
Solid wood	Single fiber or earlywood–latewood and wood ray, etc.	Lignin; hemicellulose

Selected Reference and Bibliography

Bodig, J. and Jayne, B. 1982. *Mechanics of Wood and Wood Composites*, Van Nostrand Reinhold, New York.

Forest Products Laboratory. 1974. *Wood Handbook: Wood as an Engineering Material*, U.S. Government Printing Office, Washington, D.C.

Perkins, R., Ed. 1990. *Mechanics of Wood and Paper Materials*, American Society of Mechanical Engineers, New York.

Shirasishi, N., Hiromu, K., and Norimoto, M., Eds. 1993. *Recent Research on Wood and Wood-Based Materials*, Elsevier Science Publishers, Barking, Essex, U.K.

Wangaard, F.F., Ed. 1981. *Wood: Its Structure and Properties*, The Pennsylvania State University, University Park.

12.5 Portland Cement Concrete

Steven H. Kosmatka

Introduction

Portland cement concrete is a simple material in appearance with a very complex internal nature. In contrast to its internal complexity, versatility, durability, and economy of concrete have made it the most-used construction material in the world. This can be seen in the variety of structures it is used in, from highways and bridges to buildings and dams.

Concrete is a mixture of portland cement, water, and aggregates, with or without admixtures. The portland cement and water form a paste that hardens as a result of a chemical reaction between the cement and water. The paste acts as a glue, binding the aggregates (sand and gravel or crushed stone) into a solid rocklike mass. The quality of the paste and the aggregates dictate the engineering properties of this construction material. Paste qualities are directly related to the amount of water used in relation to the amount of cement. The less water that is used, the better the quality of the concrete. Reduced water content results in improved strength and durability and in reduced permeability and shrinkage. As the fine and coarse aggregates make up 60 to 75% of the total volume of the concrete (Figure 12.5.1), their selection is important. The aggregates must have adequate strength and resistance to exposure conditions and must be durable.

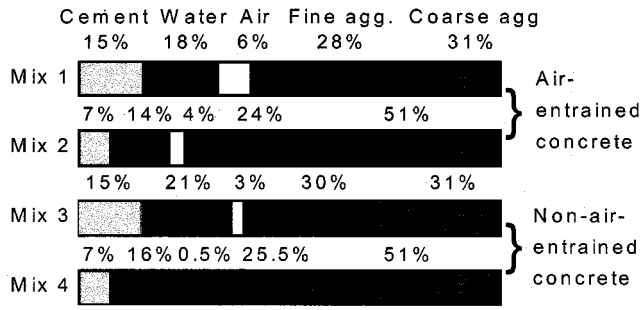


FIGURE 12.5.1 Range of proportions of materials used in concrete. (From *Design and Control of Concrete Mixtures*, EB001, Portland Cement Association, Skokie, IL, 1992. With permission.)

Fresh Concrete Properties

Freshly mixed concrete should be in a semifluid state capable of being molded by hand or mechanical means. All the particles of sand and coarse aggregate are encased and held in suspension. The ingredients should not segregate or separate during transport or handling. After the concrete hardens, it becomes a homogeneous mixture of all the components. Concrete of plastic consistency should not crumble, but flow sluggishly without segregation.

Concrete must be consolidated to form a homogeneous mass without the presence of large voids to achieve the desired strength and durability of the construction material. Internal and external vibration of concrete using vibrators allows stiff, slow-slump mixtures to be properly densified. The use of mechanical vibration provides an economical, practical method to quickly consolidate concrete without detrimentally affecting its properties.

Hardened Concrete Properties

Strength

Concrete gains strength by the reaction between cement and water — called hydration. Portland cement is primarily a calcium silicate cement. The calcium silicate combines with water and forms calcium

silicate hydrate, which is responsible for the primary engineering properties of concrete, such as setting, hardening, strength, and dimensional stability. The compressive strength of concrete increases with age as long as an appropriate moisture content and temperature are available. This is illustrated in Figure 12.5.2. Compressive strength is usually specified at the age of 28 days; however, depending on the project, ages of 3 and 7 days can also be specified. For general-use concrete, a 28-day compressive strength between 20 and 40 MPa (3000 and 6000 psi) is used. 28 MPa (4000 psi) is most common. Higher-strength concrete, 50 to 140 MPa (7000 to 20,000 psi), is used in special applications to minimize structural dimensions, increase abrasion resistance and durability, and minimize creep (long-term deformation).

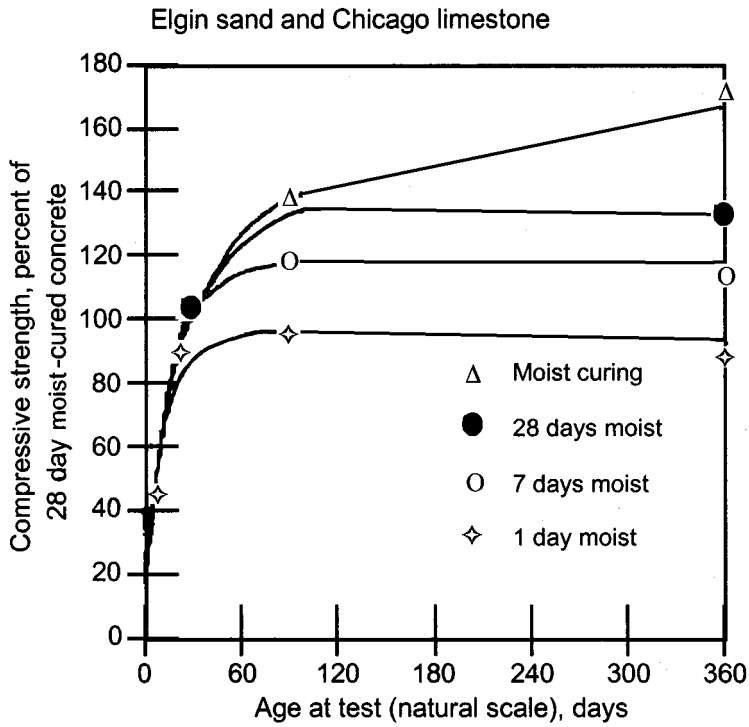


FIGURE 12.5.2 Concrete strength increases with age as long as moisture and a favorable temperature are present.

Increase in strength with age continues as long as any unhydrated cement is still present, the relative humidity in the concrete is approximately 80% or higher, and the concrete temperature is favorable. In order to maintain this increase in strength, concrete must be properly cured. Curing means that not only will a favorable temperature be present, but also moisture loss will not be permitted or extra water will be provided at the surface.

The compressive strength of concrete is directly related to the water/cement ratio. A decrease in water/cement ratio results in higher strength. Concrete achieves about 70 to 75% of its 28-day strength in 7 days. Although concrete is very strong in compression, it is weak in tensile strength. Tensile strength is about 8 to 12% of the compressive strength. Flexural strength is 0.7 to 0.8 times (for English units 5 to 7.5 times) the square root of the compressive strength. Shear strength is about 20% of the compressive strength. Modulus of elasticity ranges from 14,000 to 41,000 MPa and can be estimated as 5000 times the square root of the compressive strength (2 to 6 million psi or 57,000 times the square root of the compressive strength in English units). Refer to Tables 12.5.1, 12.5.2, and 12.5.3.

TABLE 12.5.1 Typical Properties of Normal-Strength Portland Cement Concrete

Compressive strength	20–40 MPa (3000–6000 psi)
Flexural strength	3–5 MPa (400–700 psi)
Tensile strength	2–5 MPa (300–700 psi)
Modulus of elasticity	14,000–41,000 MPa (2–6 million psi)
Permeability	1×10^{-10} cm/sec
Coefficient of thermal expansion	10^{-5}/°C (5.5×10^{-6}/°F)
Drying shrinkage	$4–8 \times 10^{-4}$
Drying shrinkage of reinforced concrete	$2–3 \times 10^{-4}$
Poisson's ratio	0.20–0.21
Shear strain	6000–17,000 MPa (1–3 million psi)
Density	2240–2400 kg/m^3 (140–150 lb/ft^3)

Density

Normal-weight concrete has a density of 2240 to 2400 kg/m^3 (140 to 150 lb/ft^3). The density of concrete varies with the relative density of the aggregate, the amount of air present in the paste, and the amount of water and cement in the mixture.

Permeability

Concrete permeability is a function of the permeability of the paste and aggregate and the interface between them. Decreased permeability improves the resistance of concrete to saturation, sulfate attack, chemical attack, and chloride penetration. Paste permeability has the greatest influence on concrete permeability. Paste permeability is directly related to the water/cement ratio and the degree of hydration or length of moist curing. A low water cement ratio and an adequate moist-curing period result in concrete with low permeability (Figure 12.5.3). The water permeability of mature, good-quality concrete, is approximately 1×10^{-10} cm/sec.

Abrasion Resistance

Abrasion resistance is directly related to the compressive strength of the concrete. The type of aggregate and the surface finish also have a strong influence on abrasion resistance. A hard aggregate, such as a granite, would provide more abrasion resistance than a soft limestone aggregate.

Volume Change and Crack Control

Concrete changes slightly in volume for various reasons. Understanding the nature of these changes is useful in planning concrete work and preventing cracks from forming. If concrete is free to move, normal volume changes would have very little consequence; but since concrete in service is usually restrained by foundations, subgrades, reinforcement, or connecting elements, significant stresses can develop. As the concrete shrinks, tensile stresses develop that can exceed the tensile strength of the concrete, resulting in crack formation.

The primary factors affecting volume change are temperature and moisture changes. Concrete expands slightly as temperature rises and contracts as temperature falls. The average value for the coefficient of thermal expansion of concrete is about 1.0×10^{-5}/°C (5.5×10^{-6}/°F). This amounts to a length change of 5 mm for a 10-m length (0.66 in. for 100 ft) of concrete subjected to a rise or fall of 50°C (100°F). The thermal coefficient of expansion for steel is about 1.2×10^{-5} per degree Celsius (6.5×10^{-6}/°F), comparable to that of concrete. The coefficient for reinforced concrete can be assumed as 1.1×10^{-5}/°C (6×10^{-6}/°F).

Concrete expands slightly with a gain in moisture and contracts with a loss in moisture. The drying-shrinkage of concrete specimens ranges from 4 to 8×10^{-4} when exposed to air at a 50% relative humidity. Concrete with a unit shrinkage of 5.5×10^{-4} shortens about the same amount as a thermal contraction caused by a decrease in temperature of 55°C (100°F). The shrinkage of reinforced concrete is less than that for plain concrete because of restraint offered by the reinforcement. Reinforced concrete structures with normal amounts of reinforcement have a drying-shrinkage in the range of 2 to 3×10^{-4}.

TABLE 12.5.2 Compressive Strength of Concrete Made from Type I Cement, psi[a]

Series	Mix Id.	w/c by Weight	1 Day	3 Days	7 Days	28 Days	3 Months	1 Year	3 Years	5 Years	10 Years	20+ Years
						Moist Curing						
308	1	0.37	2160	4430	5930	7080	8260	8410		10400		
308	2	0.51	1040	2690	4200	5890	6410		8520			
308	3	0.65	610	1770	2780	4320	5030	5020		6050		
308	4	0.82	330	990	1580	2700	3180	3290		3680		
308	5	0.36	2060	4300	5820	7010	7750	8930		10330		
308	6	0.50	990	2780	4110	5950	6440	7280		8180		
308	7	0.64	550	1710	2700	4420	5190	5390		6180		
308	8	0.83	500	1000	1690	2850	3330	3490		3790		
308	9	0.36	2010	4330	5770	6940	7940	8550		10170		
308	10	0.50	900	2600	4250	6210	6490	7280		8670		
308	11	0.64	530	1720	2810	4510	5130	5770		6470		
308	12	0.82	250	920	1670	2880	3570	3540		3750		
356	AV1	0.40		5650	7140	9020	9460	8870	10760	10070	11020	12700
356	AV2	0.53		3390	4760	6510	7480	6890	7780	7720	8900	9680
356	AV3	0.71		1770	2540	4160	4540	4540	4960	5030	5840	6130
356	DV1	0.40		5000	6710	7980	8530	8660	10240	10340	10540	12300
356	DV2	0.53		3070	4810	6720	7400	7840	8780	8720	9330	9980
356	DV3	0.71		1580	2610	4380	5060	5360	5680	5700	6470	6710
356	EV1	0.40		4670	6650	8530	9790	10090	9940	11330	10900	12460
356	EV2	0.53		2700	4580	6730	7690	7860	8240	8960	8580	10290
356	EV3	0.71		1470	2500	4290	5010	5070	5110	5770	6200	6420
356	XL1	0.40		4780	6300	8090		9470		10780	10740	11730
356	XL2	0.53		3170	4470	6040		7740		8070	8370	8060
356	XL3	0.71		1800	2650	4060		4990		5220	5280	5260
356	XV1	0.40		5220	6950	8520		9720		10550	10850	13100
356	XV2	0.53		3350	4870	6700		7950		8710	8760	9800
356	XV3	0.71		1680	2810	4120		5060		5610	6280	6590
356	XW1	0.40		4680	6200	7780		9840		10100	10310	11480
356	XW2	0.53		3100	4520	6270		7420		8180	8150	8760
356	XW3	0.71		1770	2680	4140		5290		5560	5630	5460
374	11	0.41	1550		5680	7390	7610	9160	9810		10070	10460
374	11	0.56	780		4210	5870	6390	7020	6910		6740	7410
374	12	0.41	1120		5920	7490	8770	9170	9710		9710	10900
374	12	0.55	580		3800	5710	6650	7010	7380		7010	7300
374	13	0.42	1520		4320	6280	7560	8280	8620		9100	9940
374	13	0.57	890		2740	4730	5760	6400	6750		6340	7360
374	14	0.41	1490		5020	6460	7160	8730	9280		9400	10140
374	14	0.55	800		3480	5190	5700	6540	7030		6560	7410
374	15	0.45	2230		6080	7180	7820	8530	9290		9520	10770
374	15	0.59	1260		4730	5830	6280	6440	6360		5980	
374	16	0.41	1820		6040	7230	8080	9540	10160		10400	
374	16	0.56	1020		4000	5820	6520	7060	7640		7230	6840
374	17	0.46	1340		5220	7040	7560	8700	9310		9280	
374	17	0.61	740		3480	5770	6410	6560	6890		6700	7190
374	18	0.49	1220		5290	7000	7600	8360	9110		10290	10630
374	18	0.58	670		3600	5600	6230	6540	6810		6660	
374	19A	0.45	770		3090	4810	6600	7350	7930		8610	9240
374	19A	0.54	410		1710	3100	4280	5060	5940		5950	6480
374	19B	0.45	1110		4350	6560	7450	7740	8850		9550	10550
374	19B	0.60	560		2640	4260	5140	5420	6000		6360	6730
374	19C	0.45	1540		5370	6860	7390	7960	9000		9520	
374	19C	0.59	890		3910	5520	6200	6580	6970		6640	7220
436	1	0.36	2640	4780	6460	8070	8890	9670	9840		11030	10710

TABLE 12.5.2 Compressive Strength of Concrete Made from Type I Cement, psi[a] (continued)

Series	Mix Id.	w/c by Weight	1 Day	3 Days	7 Days	28 Days	3 Months	1 Year	3 Years	5 Years	10 Years	20+ Years
436	2	0.49	1320	2780	4200	6110	7210	7920	7990		8990	8940
436	3	0.62	700	1620	2550	4170	5210	5840	5850		6010	5390
436	4	0.42	1760	3500	5080	7200	8100	8720	9640		10540	10820

[a] To convert to MPa, multiply by 0.00689476.

Source: From Wood, S.L., *Research and Development Bulletin RD102T,* Portland Cement Association, Skokie, IL, 1992. With permission.

The amount of shrinkage is directly related to the amount of water in the concrete. Higher water content results in higher shrinkage. Specimen size also has an effect. Larger specimens shrink less than small specimens.

Drying-shrinkage is an inherent and unavoidable property of concrete; therefore, properly positioned reinforcing steel is used to reduce crack widths or joints are used to predetermine or control the location of cracks. Shrinkage control joints should be spaced about 25 to 30 times the thickness of a concrete slab on ground.

Deformation and Creep

Concrete will deform a small amount when a load is placed on it. When concrete is loaded, the deformation caused by the load can be divided into two parts: a deformation that occurs immediately, such as elastic strain, and a time-dependent deformation that begins immediately but continues at a decreasing rate for as long as the concrete is loaded (Figure 12.5.4). This latter deformation is called creep. The amount of creep is dependent on the magnitude of the stress, the age and strength of the concrete when the stress is applied, and the length of time the concrete is stressed. Creep is of little concern for normal concrete pavements, bridges, and small buildings; however, creep should be considered in design for very tall buildings or very long bridges.

Concrete Ingredients

Portland Cements

Portland cements are hydraulic cements; that is, they set and harden by reacting with water. This reaction, called hydration, combines water and cement to form a stonelike mass. Portland cement was invented in 1824 by an English mason, Joseph Aspdin, who named his product portland cement because it produced a concrete that was of the same color as natural stone on the Isle of Portland in the English Channel.

Portland cement is produced by combining appropriate proportions of lime, iron, silica, and alumina and heating them. These raw ingredients are fed into a kiln that heats the ingredients to temperatures of 1450 to 1650°C (2600 to 3000°F) and changes the raw materials chemically into cement clinker. The clinker is cooled and then pulverized. During this operation a small amount of gypsum is added to control the setting of the cement. The finished pulverized product is portland cement. Portland cement is essentially a calcium silicate cement.

The American Society for Testing and Materials (ASTM) Standard C 150, Specification for Portland Cement, provides for the following types of portland cement:

Type I	General portland cement
Type II	Moderate-sulfate-resistant cement
Type III	High-early-strength cement
Type IV	Low–heat-of-hydration cement
Type V	High-sulfate-resistant cement

Types I, II, and III may also be designated as being air entraining. Type I portland cement is a general cement suitable for all uses where special properties of other cements are not required. It is commonly used in pavements, building, bridges, and precast concrete products.

TABLE 12.5.3 Flexural Strength of Concrete Made from Type I Cement, psi[a] (Third-Point Loading)

Series	Mix Id.	w/c by Weight	1 Day	3 Days	7 Days	28 Days	3 Months	1 Year	3 Years	5 Years	10 Years	20 Years
						Moist Curing						
308	1	0.37	295	540	625	855	975	925		960		
308	2	0.51	160	415	570	765	805	845		780		
308	3	0.65	80	290	425	595	675	680		645		
380	4	0.82	40	155	285	450	505	485		480		
308	5	0.36	285	510	680	825	890	940		930		
308	6	0.50	165	445	545	720	810	760		815		
308	7	0.64	95	290	465	605	695	655		690		
308	8	0.83	45	180	310	450	530	525		470		
308	9	0.36	310	535	655	820	905	970		915		
308	10	0.50	175	410	570	710	860	815		830		
308	11	0.64	90	320	440	675	715	715		690		
308	12	0.82	40	180	330	490	575	555		495		
356	XL1	0.40			705	880						
356	XL2	0.53			555	720						
356	XL3	0.71			420	555						
356	XV1	0.40			625	725						
356	XV2	0.53			555	655						
356	XV3	0.71			385	490						
356	XW1	0.40			620	750						
356	XW2	0.53			515	665						
356	XW3	0.71			395	515						
374	11	0.41	240		640	765	840	905	855		945	1070
374	11	0.56	135		520	625	710	690	730		730	830
374	12	0.41	180		640	790	910	925	965		970	1030
374	12	0.55	100		530	705	785	755	770		795	850
374	13	0.42	260		525	690	845	885	915		995	1140
374	13	0.57	165		420	595	705	765	765		800	880
374	14	0.41	250		620	725	800	915	890		965	
374	14	0.55	150		490	630	740	765	745		780	890
374	15	0.45	320		660	755	865	865	870		945	970
374	15	0.59	215		580	650	710	710	665		735	830
374	16	0.41	265		675	755	890	920	935		1035	1110
374	16	0.56	165		535	655	760	770	775		820	930
374	17	0.46	210		600	685	855	870	935		905	
374	17	0.61	120		490	650	735	760	720		695	
374	18	0.49	220		585	720	830	915	895		965	1090
374	18	0.58	130		510	650	750	740	730		725	
374	19A	0.45	145		450	605	745	840	870		950	1000
374	19A	0.54	75		300	470	610	655	700		735	850
374	19B	0.45	195		555	670	810	835	875		955	1060
374	19B	0.60	105		415	580	680	725	715		785	840
374	19C	0.45	275		605	750	890	920	890		950	1050
374	19C	0.59	170		510	665	775	810	765		790	890
436	1	0.36	370	555	655	770	925	l980	955		960	1030
436	2	0.49	225	435	565	745	825	860	905		900	960
436	3	0.62	135	300	420	620	690	720	730		730	800
436	4	0.42	320	505	655	755	875	890	1005		1010	1060

[a] To convert to MPa, multiply by 0.00689476.

Source: From Wood, S.L., *Research and Development Bulletin RD102T,* Portland Cement Association, Skokie, IL, 1992. With permission.

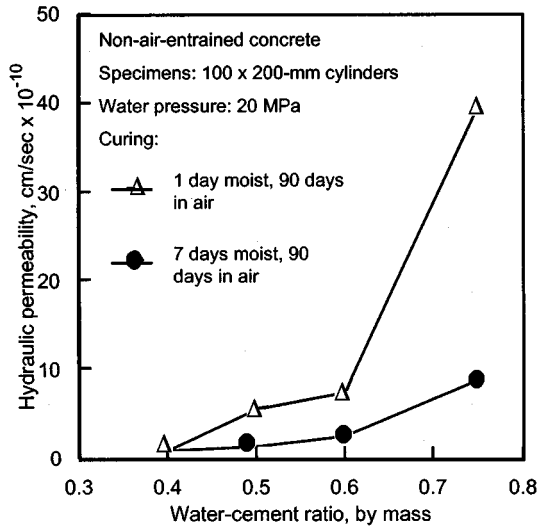


FIGURE 12.5.3 Water permeability of concrete as affected by water/cement ratio and curing. (From *Design and Control of Concrete Mixtures*, EB001, Portland Cement Association, Skokie, IL, 1992. With permission.)

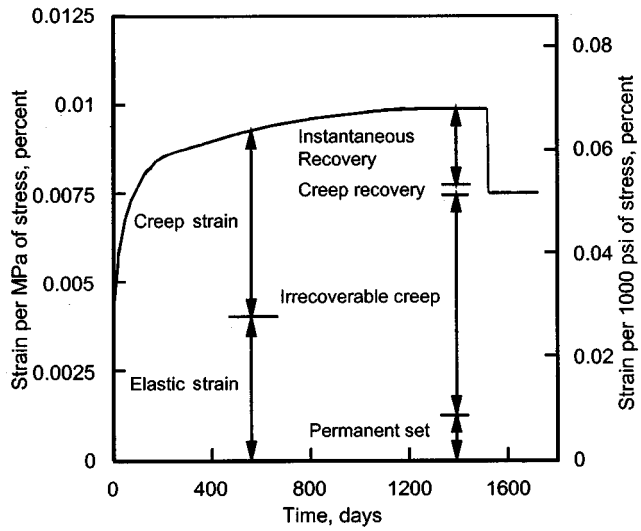


FIGURE 12.5.4 Combined curve of elastic and creep strains showing amount of recovery. (From *Design and Control of Concrete Mixtures*, EB001, Portland Cement Association, Skokie, IL, 1992. With permission.)

Type II portland cement is used where precaution against moderate sulfate attack is important where sulfate concentrations in groundwater or soil are higher than normal, but not severe. Type II cement can also be specified to generate less heat than Type I cement. This moderate heat of hydration requirement is helpful when placing massive structures, such as piers, heavy abutments, and retaining walls. Type II cement may be specified when water-soluble sulfate in soil is between 0.1 and 0.2%, or when the sulfate content in water is between 150 and 1500 ppm. Types I and II are the most common cements available.

Type III portland cement provides strength at an early age. It is chemically similar to Type I cement except that the particles have been ground finer to increase the rate of hydration. It is commonly used in fast-track paving or when the concrete structure must be put into service as soon as possible, such as in bridge deck repair.

Type IV portland cement is used where the rate and amount of heat generated from hydration must be minimized. This low heat of hydration cement is intended for large, massive structures, such as gravity dams. Type IV cement is rarely available.

Type V portland cement is used in concrete exposed to very severe sulfate exposures. Type V cement would be used when concrete is exposed to soil with a water-soluble sulfate content of 0.2% and higher or to water with over 1500 ppm of sulfate. The high sulfate resistance of Type V cement is attributed to its low tricalcium aluminate content.

Blended Hydraulic Cements

Blended hydraulic cements are produced by intimately blending two or more types of cementitious material. Primary blending materials are portland cement, ground granulated blast-furnace slag, fly ash, natural pozzolans, and silica fume. These cements are commonly used in the same manner as portland cements. Blended hydraulic cements conform to the requirements of ASTM C 595 or C 1157. ASTM C 5195 cements are as follows: Type IS — portland blast-furnace slag cement, Type IP and Type P — portland–pozzolan cement, Type S — slag cement, Type I (PM) — pozzolan-modified portland cement, and Type I (SM) — slag-modified portland cement. The most common blended cements available are Types IP and IS.

ASTM C 1157 blended hydraulic cements include the following: Type GU — blended hydraulic cement for general construction, Type HE — high-early-strength cement, Type MS — moderate-sulfate-resistant cement, Type HS — high-sulfate-resistant cement, Type MH — moderate-heat-of-hydration cement, and Type LH — low-heat-of-hydration cement.

Supplementary Cementing Materials (Mineral Admixtures)

Supplementary cementing materials, also called mineral admixtures, are sometimes added to concrete mixtures. They contribute to the properties of hardened concrete through hydraulic or pozzolanic activity. Typical examples are natural pozzolans, fly ash, ground granulated blast-furnace slag, and silica fume. These materials react chemically with calcium hydroxide released from the hydration of portland cement to form cement compounds.

Below is a summary of the specifications and classes of supplementary cementing materials:

1. Ground granulated iron blast-furnace slag—ASTM C 989

   Grade 80 — Slags with a low activity index
   Grade 100 — Slags with a moderate activity index
   Grade 120 — Slags with a high activity index
2. Fly ash and natural pozzolans — ASTM C 618
3. Class N — Raw or calcined natural pozzolans including diatomaceous earth, opaline cherts, shales, tuffs, volcanic ashes, and some calcined clays and shales
4. Class F — Fly ash with pozzolanic properties
5. Class C — Fly ash with pozzolanic and cementitious properties
6. Silica fume — ASTM C 1240

Mixing Water for Concrete

Almost any natural water that is drinkable can be used as mixing water for making concrete. However, some waters that are not fit for drinking may be suitable for concrete. Reference 1 provides guidance concerning the use of waters containing alkali carbonates, chlorides, sulfates, acids, oils, and other materials, and provides guidance as to allowable levels of contamination.

Aggregates for Concrete

The importance of using the right type and quality of aggregates cannot be overemphasized since the fine coarse aggregates occupy between 60 to 75% of the concrete volume and strongly influence the freshly mixed and hardened properties, mix proportions, and economy of the concrete. Fine aggregates consist of natural sand or crushed rock with particles smaller than 5 mm (0.2 in). Coarse aggregates

consist of a combination of gravel or crushed aggregate with particles predominately larger than 5 mm (0.2 in.) and generally between 10 and 13 mm ($^3/_8$ and $^1/_2$ in.). The most common coarse aggregate size is 19 and 25 mm ($^3/_4$ and 1 in.) aggregate.

Normal weight aggregates should meet the requirements of ASTM C 33. This specification limits the amounts of harmful substances and states the requirements for aggregate characteristics, such as grading. The grading and maximum size of the aggregate affect the relative aggregate proportions as well as cement and water requirements, workability, pumpability, economy, shrinkage, and durability of the concrete.

Chemical Admixtures for Concrete

Admixtures are ingredients in concrete other than portland cement, water, and aggregates that are added to the mixture immediately before or during mixing. Common chemical admixtures include air-entraining, water-reducing, retarding, accelerating, and superplasticizing admixtures. The major reasons for using admixtures are to reduce the cost of concrete construction, achieve certain properties in concrete more effectively than by other means, or to ensure the quality of concrete during the states of mixing, transporting, placing, or curing in adverse weather conditions. Refer to Table 12.5.4.

Air-entraining admixtures are used purposely to entrain microscopic air bubbles in concrete. Air entrainment will dramatically improve the durability of concrete exposed to moisture during freezing and thawing. Air-entraining admixtures are commonly used to provide between 5 and 8% air content in concrete.

Water-reducing admixtures are used to reduce the quantity of mixing water required to produce concrete of a certain slump, reduce water/cement ratio, reduce cement content, or increase slump. Typical water-reducing admixtures reduce the water content by approximately 5 to 10%. High-range water reducers (superplasticizers) reduce the water content by approximately 12 to 30% and they can produce a highly fluid concrete.

Retarding admixtures are used to retard the rate of setting of concrete. An accelerating admixture is used to accelerate strength development of concrete at an early age.

Proportioning Normal Concrete Mixtures

The objective in proportioning concrete mixtures is to determine the most economical and practical combination of readily available materials to produce a concrete that will satisfy the performance requirements under particular conditions of use. To fulfill these objectives a properly proportioned concrete mix should possess these qualities: (1) acceptable workability of freshly mixed concrete, (2) durability, strength, and uniform appearance of hardened concrete, and (3) economy. Only with proper selection of materials and mixture characteristics can the above qualities be obtained in concrete production.

The key to designing a concrete mixture is to be fully aware of the relationship between the water/cement ratio and its effect on strength and durability. The specified compressive strength at 28 days and durability concerns dictate the water/cement ratio established for a concrete mixture. The water/cement ratio is simply the weight of water divided by the weight of cement. If pozzolans or slags are used, it would include their weights and would be referred to as the water/cementitious material ratio. The water/cement ratio can be established by a known relationship to strength or by durability requirements. For example, a concrete structure may require only 20 MPa (3000 psi) compressive strength, which would relate to a water/cement ratio of about 0.6; however, if the concrete is exposed to deicers, the maximum water/cement ratio should be 0.45 (Table 12.5.5). For corrosion protection or reinforced concrete exposed to deicers, the maximum water/cement ratio should be 0.40. When designing concrete mixtures, remember that where durability is concerned, water/cement ratio should be as low as practical. Entrained air must be used in all concrete that will be exposed to freezing and thawing and

TABLE 12.5.4 Concrete Admixtures by Classification

Type of Admixture	Desired Effect	Material
Accelerators (ASTM C 494, Type C)	Accelerate setting and early-strength development	Calcium chloride (ASTM D 98) Triethanolamine, sodium thiocyanate, calcium formate, calcium nitrite, calcium nitrate
Air detrainers	Decrease air content	Tributyl phosphate, dibutyl phthalate, octyl alcohol, water-insoluble esters of carbonic and boric acid, silicones
Air-entrained admixtures (ASTM C 260)	Improve durability in environments of freeze-thaw, deicers, sulfate, and alkali reactivity Improve workability	Salts of wood resins (Vinsol resin) Some synethic detergents Salts of sulfonated lignin Salts of petroleum acids Salts of proteinaceous material Fatty and resinous acids and their salts Alkylbenzene sulfonates Salts of sulfonated hydrocarbons
Alkali-reactivity reducers	Reduce alkali-reactivity expansion	Natural pozzolans, fly ash, silica fume, blast-furnace slag, salts of lithium and barium
Bonding admixtures	Increase bond strength	Rubber, polyvinyl chloride, polyvinyl acetate, acrylics, butadiene-styrene copolymers
Coloring agents	Colored concrete	Modified carbon black, iron oxide, phthalocyanine, umber, chromium oxide, titanium oxide, cobalt blue (ASTM C 979)
Corrosion inhibitors	Reduce steel corrosion activity in a chloride environment	Calcium nitrite, sodium nitrite, sodium benzoate, certain phosphates or fluorosilicates, fluoroaluminates
Dampproofing admixtures	Retard moisture penetration into dry concrete	Soaps of calcium or ammonium stearate or eleate Butyl stearate Petroleum products
Finely divided mineral admixtures		
Cementitious	Hydraulic properties Partial cement replacement	Ground granulated blast-furnace slag (ASTM C 989) Natural cement Hydraulic hydrated lime (ASTM C 141)
Pozzolans	Pozzolanic activity Improve workability, plasticity, sulfate resistance; reduce alkali reactivity, permeability, heat of hydration Partial cement replacement Filler	Diatomaceous earth, opaline cherts, clays, shales, volcanic tufts, pumicites (ASTM C 618, Class N); fly ash (ASTM C 618, Classes F and C), silica fume
Pozzolanic and cementitious	Same as cementitious and pozzolan categories	High calcium fly ash (ASTM C 618, Class C) Ground granulated blast-furnace slag (ASTM C 989)
Nominally inert	Improve workability Filler	Marble, dolomite, quartz, granite
Fungicides, germicides, and insecticides	Inhibit or control bacterial and fungal growth	Polyhalogenated phenols Dieldrin emulsions Copper compounds
Gas formers	Cause expansion before setting	Aluminum powder Resin soap and vegetables or animal glue Saponin Hydrolized protein
Grouting agents	Adjust grout properties for specific applications	See Air-entraining admixtures, Accelerators, Retarders, Workability agents
Permeability reducers	Decrease permeability	Silica fume (ASTM C 1240) Fly ash (ASTM C 618) Ground slag (ASTM C 989) Natural pozzolans (ASTM C 618)

TABLE 12.5.4 Concrete Admixtures by Classification (continued)

Type of Admixture	Desired Effect	Material
		Water reducers
		Latex
Pumping aids	Imnprove pumpability	Organic and synthetic polymers
		Organic flocculents
		Organic emulsions of paraffin, coal tar, asphalt, acrylics
		Bentonite and pyrogenic silicas
		Natural pozzolans (ASTM C 618, Class N)
		Fly ash (ASTM C 618, Classes F and C)
		Hydrated lime (ASTM C 141)
Retarders (ASTM C 494, Type B)	Retard setting time	Lignin
		Borax
		Sugars
		Tartaric acid and salts
Superplasticizers[a] (ASTM C 1017, Type 1)	Flowing concrete	Sulfonated melamine formaldehyde condensates
	Reduce water/cement ratio	Sulfonated napthalene formaldehyde condensates
		Ligonsulfonates
Superplasticizer[a] and retarder (ASTM C 1017, Type 2)	Flowing concrete with retarded set Reduce water	See Superplasticizers and also Water reducers
Water reducer (ASTM C 494, Type A)	Reduce water demand at least 5%	Lignosulfonates
		Hydroxylated carboxylic acids
		Carbohydrates
		(Also tend to retard set so accelerator is often added)
Water reducer and accelerator (ASTM C 494, Type E)	Reduce water (minimum 5%) and accelerate set	See Water reducer, Type A (Accelerator is added)
Water reducer and retarder (ASTM C 494, Type D)	Reduce water (minimum 5%) and retard	See Water reducer, Type A
Water reducer — high range (ASTM C 494, Type F)	Reduce water demand (miniumum 12%)	See Superplasticizers
Water reducer — high range — and retarder (ASTM C 494, Type G)	Reduce water demand (miniumum 12%) and retard set	See Superplasticizers and also Water reducers
Workability agents	Improve workability	Air-entraining admixtures
		Finely divided admixtures, except silica fume
		Water reducers

[a] Superplasticizers are also referred to as high-range water reducers or plasticizers. These admixtures often meet both ASTM C 494 and C 1017 specifications simultaneously.

Source: From *Design and Control of Concrete Mixtures*, EB001, Portland Cement Association, Skokie, IL, 1992. With permission.

the presence of deicing chemicals. A typical air content for concrete would range from 5 to 8%. Reference 1 provides step-by-step procedures for proportioning concrete. Refer to Tables 12.5.6, 12.5.7, and 12.5.8.

Mixing, Transporting, and Placing Concrete

All concrete should be mixed thoroughly until it is uniform in appearance with all ingredients evenly distributed. If concrete has been adequately mixed, samples taken from different portions of a batch will have essentially the same unit weight, air content, slump, and strength. Concrete is sometimes mixed at a job site at a stationary mixer or paving mixer, and other times it is mixed in central mixers at ready-mix plants (ASTM C 94). Once concrete is transported to a job site it is then conveyed by a variety of methods including belt conveyors, buckets, shoots, cranes, pumps, wheelbarrows, and other equipment.

TABLE 12.5.5 Relationships Between Water Cement Ratio and Compressive Strength of Concrete

Compressive Strength at 28 days, MPa (psi)	Water-Portland Cement Ratio, by Mass	
	Non-Air-Entrained Concrete	Air-Entrained Concrete
40 (5800)	0.42	—
35 (5100)	0.47	0.39
30 (4400)	0.54	0.45
25 (3600)	0.61	0.52
20 (2900)	0.69	0.60
15 (2200)	0.79	0.70

Note: Strength is based on 150 × 300-mm cylinders moist cured 28 days at 23 ± 2°C. Relationship assumes maximum size of aggregate about 25 mm.
Adapted from *Design and Control of Concrete Mixtures,* EBDD1, Portland Cement Association, Skokie, IL, 1992. With permission.

Concrete should be conveyed in a manner in which the concrete is not allowed to dry out, be delayed, or allowed to segregate before it is placed.

Curing

All concrete must be properly cured. Curing is the maintenance of a satisfactory moisture content and temperature in concrete during some definite time period immediately following placing and finishing so that the desired properties of strength and durability may develop. Concrete should be moist cured for 7 days at a temperature between 10 and 27°C (50 and 80°F). Common methods of curing include ponding, spraying, or fogging; use of wet covers, impervious paper, plastic sheets, and membrane-forming curing compounds; or a combination of these.

Durability

Freeze-Thaw and Deicer Scaling Resistance

As water freezes in wet concrete, it expands 9%, producing hydraulic pressures in the cement paste and aggregate. Accumulated effects of successive freeze-thaw cycles and disruption of the paste and aggregate eventually cause significant expansion and extensive deterioration of the concrete. The deterioration is visible in the form of cracking, scaling, and crumbling.

The resistance of hardened concrete to freezing and thawing in a moist condition, with or without the presence of deicers, is significantly improved by the use of entrained air. Air entrainment prevents frost damage and scaling and is required for all concretes exposed to freezing and thawing or deicer chemicals. An air content of between 5 and 8% should be specified. Air-entrained concrete should be composed of durable materials and have a low water/cement ratio (maximum 0.45), a minimum cement content of 335 kg/m^3 (564 lb/yd^3) or move, proper finishing after bleed water has evaporated from the surface, adequate drainage, a minimum of 7-days moist curing at or above 10°C (50°F), a minimum compressive strength of 28 MPa (4000 psi), and a minimum 30-day drying period after moist curing. Sealers may also be applied to provide additional protection against the effects of freezing and thawing and deicers. However, a sealer should not be necessary for properly proportioned and placed concrete.

Sulfate-Resistant Concrete

Excessive amounts of sulfates in soil or water can, over 5 to 30 years, attack and destroy concrete that is not properly designed. Sulfates attack concrete by reacting with hydrated compounds in the hardened cement paste. Due to crystallization growth, these expansive reactions can induce sufficient pressure to

TABLE 12.5.6a (Metric Units) Example Trial Mixtures for Air-Entrained Concrete of Medium Consistency, 80 to 100 mm Slump

					With Fine Sand, Fineness Modulus = 2.50			With Coarse Sand, Fineness Modulus = 2.90		
Water/ Cement Ratio, kg/kg	Nominal Maximum Size of Aggregate, mm	Air Content, %	Water, kg/m³ of Concrete	Cement kg/m³ of Concrete	Fine Aggregate % of Total Aggregate	Fine Aggregate, kg/m³ of Concrete	Coarse Aggregate, kg/m³ of Concrete	Fine Aggregate % of Total Aggregate	Fine Aggregate, kg/m³ of Concrete	Coarse Aggregate, kg/m³ of Concrete
0.40	10	7.5	202	505	50	744	750	54	809	684
	14	7.5	194	485	41	630	904	46	702	833
	20	6	178	446	35	577	1071	39	648	1000
	40	5	158	395	29	518	1255	33	589	1184
0.45	10	7.5	202	450	51	791	750	56	858	684
	14	7.5	194	428	43	678	904	47	750	833
	20	6	178	395	37	619	1071	41	690	1000
	40	5	158	351	31	553	1255	35	625	1184
0.50	10	7.5	202	406	53	833	750	57	898	684
	14	7.5	194	387	44	714	904	49	785	833
	20	6	178	357	38	654	1071	42	726	1000
	40	5	158	315	32	583	1225	36	654	1184
0.55	10	7.5	202	369	54	862	750	58	928	684
	14	7.5	194	351	45	744	904	49	815	833
	20	6	178	324	39	678	1071	43	750	1000
	40	5	158	286	33	613	1225	37	684	1184
0.60	10	7.5	202	336	54	886	750	58	952	684
	14	7.5	194	321	46	768	904	50	839	833
	20	6	178	298	40	702	1071	44	773	1000
	40	5	158	262	33	631	1225	37	702	1184
0.65	10	7.5	202	312	55	910	750	59	976	684
	14	7.5	194	298	47	791	904	51	863	823
	20	6	178	274	40	720	1071	44	791	1000
	40	5	158	244	34	649	1225	38	720	1184
0.70	10	7.5	202	288	55	928	750	59	994	684
	14	7.5	194	277	47	809	904	51	880	833
	20	6	178	256	41	738	1071	45	809	1000
	40	5	158	226	34	660	1225	38	732	1184

From *Design and Control of Concrete Mixtures*, EB001, Portland Cement Association, Skokie, IL, 1992. With permission.

TABLE 12.5.6b (English Units) Example Trial Mixtures for Air-Entrained Concrete of Medium Consistency, 3- to 4-in. Slump

					With Fine Sand, Fineness Modulus = 2.50			With Coarse Sand, Fineness Modulus = 2.90		
Water/ Cement Ratio, lb per lb	Maximum Size of Aggregate, in,	Air Content, %	Water, lb per cu yd of Concrete	Cement lb per cu yd of Concrete	Fine Aggregate % of Total Aggregate	Fine Aggregate, lb per cu yd of Concrete	Coarse Aggregate, lb per cu yd of Concrete	Fine Aggregate % of Total Aggregate	Fine Aggregate, lb per cu yd of Concrete	Coarse Aggregate, lb per cu yd of Concrete
0.40	3/8	7.5	340	850	50	1250	1260	54	1360	1150
	1/2	7.5	325	815	41	1060	1520	46	1180	1400
	3/4	6	300	750	35	970	1800	39	1090	1680
	1	6	285	715	32	900	1940	38	1010	1830
	1 1/2	5	265	665	29	870	2110	33	990	1990
0.45	3/8	7.5	340	755	51	1330	1260	56	1440	1150
	1/2	7.5	325	720	43	1140	1520	47	1260	1400
	3/4	6	300	665	37	1040	1800	41	1160	1680
	1	6	285	635	33	970	1940	37	1080	1830
	1 1/2	5	265	590	31	930	2110	35	1050	1990
0.50	3/8	7.5	340	680	53	1400	1260	57	1510	1150
	1/2	7.5	325	650	44	1200	1520	49	1320	1400
	3/4	6	300	600	38	1100	1800	42	1220	1680
	1	6	285	570	34	1020	1940	38	1130	1830
	1 1/2	5	265	530	32	980	2110	36	1100	1990
0.55	3/8	7.5	340	620	54	1450	1260	58	1550	1150
	1/2	7.5	325	590	45	1250	1520	49	1370	1400
	3/4	6	300	545	39	1140	1800	43	1260	1680
	1	6	285	520	35	1060	1940	39	1170	1830
	1 1/2	5	265	480	33	1030	2110	37	1150	1990
0.60	3/8	7.5	340	565	54	1490	1260	58	1600	1150
	1/2	7.5	325	540	46	1290	1520	50	1410	1400
	3/4	6	300	500	40	1180	1800	44	1300	1680
	1	6	285	475	36	1100	1940	40	1210	1830
	1 1/2	5	265	440	33	1060	2110	37	1180	1990
0.65	3/8	7.5	340	525	55	1530	1260	59	1640	1150
	1/2	7.5	325	500	47	1330	1520	51	1450	1400
	3/4	6	300	460	40	1210	1800	44	1330	1680
	1	6	285	440	37	1130	1940	40	1240	1830
	1 1/2	5	265	410	34	1090	2110	38	1210	1990

TABLE 12.5.6b (English Units) Example Trial Mixtures for Air-Entrained Concrete of Medium Consistency, 3- to 4-in. Slump (continued)

					With Fine Sand, Fineness Modulus = 2.50			With Coarse Sand, Fineness Modulus = 2.90		
Water/ Cement Ratio, lb per lb	Maximum Size of Aggregate, in,	Air Content, %	Water, lb per cu yd of Concrete	Cement lb per cu yd of Concrete	Fine Aggregate % of Total Aggregate	Fine Aggregate, lb per cu yd of Concrete	Coarse Aggregate, lb per cu yd of Concrete	Fine Aggregate % of Total Aggregate	Fine Aggregate, lb per cu yd of Concrete	Coarse Aggregate, lb per cu yd of Concrete
0.70	3/8	7.5	340	485	55	1560	1260	59	1670	1150
	1/2	7.5	325	465	47	1360	1520	51	1480	1400
	3/4	6	300	430	41	1240	1800	45	1360	1680
	1	6	285	405	37	1160	1940	41	1270	1830
	1 1/2	5	265	380	34	1110	2110	38	1230	1990

TABLE 12.5.7 Proportions by Mass to Make $^1/_{10}$ m^3 of Concrete for Small Jobs

	Air-Entrained Concrete				Non-Air-Entrained Concrete			
Maximum-Size Coarse Aggregate, mm	Cement, kg	Wet Fine Aggregate, kg	Wet Coarse Aggregate, kg	Water, kg	Cement, kg	Wet Fine Aggregate, kg	Wet Coarse Aggregate, kg[a]	Water, kg
10	46	85	74	16	46	94	74	18
14	43	74	88	16	43	85	88	18
20	40	67	104	16	40	75	104	16
40	37	61	120	14	37	69	120	14

[a] If crushed stone is used, decrease coarse aggregate by 5 kg and increase fine aggregate by 5 kg.

Source: From *Design and Control of Concrete Mixtures*, EB001, Portland Cement Associate, Skokie, IL, 1992. With permission.

TABLE 12.5.8 Proportions by Volume[a] of Concrete for Small Jobs

	Air-Entrained Concrete				Non-Air-Entrained Concrete			
Maximum-Size Coarse Aggregate, mm	Cement	Wet Fine Aggregate	Wet Coarse Aggregate	Water	Cement	Wet Fine Aggregate	Wet Coarse Aggregate	Water
10	1	$2^1/_4$	$1^1/_2$	$^1/_2$	1	$2^1/_2$	$1^1/_2$	$^1/_2$
14	1	$2^1/_4$	2	$^1/_2$	1	$2^1/_2$	2	$^1/_2$
20	1	$2^1/_4$	$2^1/_2$	$^1/_2$	1	$2^1/_2$	$2^1/_2$	$^1/_2$
40	1	$2^1/_4$	3	$^1/_2$	1	$2^1/_2$	3	$^1/_2$

[a] The combined volume is approximately $^2/_3$ of the sum of the original bulk volumes.

Source: From *Design and Control of Concrete Mixtures*, EB001, Portland Cement Associate, Skokie, IL, 1992. With permission.

disrupt the cement paste, resulting in cracking and disintegration of the concrete. The first defense against sulfate attack is to use a low water/cement ratio (0.45 or preferably less), and to select a Type II or V cement (see the section on cement).

Corrosion Protection

Concrete protects embedded steel from corrosion through its highly alkaline nature. The high-pH environment (usually greater than 12.5) causes a passive and noncorroding protective oxide film to form on steel. However, carbonation or the presence of chloride ions from deicers or seawater can destroy or penetrate the film, causing rusting of the reinforcing steel. In addition to using a water/cement ratio of 0.40 or less, the following protective strategies can be used individually or in combination to reduce the risk of corrosion:

1. Cover thickness of 90 mm (3.5 in.) or more of concrete over top, reinforcing steel of compression zones. [*Note:* Excessive cover in tension zones exacerbates surface crack width.]
2. Low-slump dense concrete overlay
3. Latex-modified concrete overlay
4. Interlayer membrane/asphaltic concrete systems
5. Epoxy-coated reinforcing steel
6. Corrosion-inhibiting admixtures in concrete
7. Sealers with or without overlay
8. Silica-fume or other pozzolans that significantly reduce concrete permeability
9. Low water/cement ratio (<0.35) superplasticized concrete
10. Cathodic protection
11. Polymer concrete overlay

12. Galvanized reinforcing steel
13. Polymer impregnation
14. Lateral and longitudinal prestressing for crack control
15. Blended cements containing silica fume or other pozzolans to reduce permeability

Alkali–Silica Reaction

Most aggregates are chemically stable in hydraulic cement concrete, without deleterious interaction with other concrete ingredients. However, this is not the case for aggregates containing certain siliceous substances that react with soluble alkalis in concrete. Alkali–silica reactivity (ASR) is an expansive reaction between reactive forms of silica in aggregate and alkali hydroxides in concrete. Very reactive aggregates can induce cracks within a year, whereas slowly reactive aggregates can take over 20 years to induce noticeable cracks. ASR is best controlled through the use of fly ash, slag, silica fume, natural pozzolans, or blended hydraulic cement. With proper care in analyzing aggregates and selecting appropriate concrete ingredients, ASR can be effectively minimized using available materials. Reference 2 provides guidance on ASR.

Related Standards and Specifications

American Society for Testing and Materials (ASTM)

C 33	Specification for Concrete Aggregates
C 150	Specification for Portland Cement
C 595	Specification for Blended Hydraulic Cements
C 618	Specification for Fly Ash and Raw and Calcined Natural Pozzolans for Use as a Mineral Admixture in Portland Cement Concrete
C 989	Specification for Ground Granulated Blast-Furnace Slag for Use in Concrete and Mortars
C 1157	Performance Specification for Blended Hydraulic Cement
C 1240	Specification for Silica Fume for Use in Hydraulic-Cement Concrete and Mortar

References

1. *Design and Control of Concrete Mixtures*, EB001, Portland Cement Association, Skokie, IL, 1992, 214 pages.
2. *Guide Specification for Concrete Subject to Alkali-Silica Reactions*, IS415, Portland Cement Association, Skokie, IL, 1995, 8 pages.
3. *Specifications for Structural Concrete for Buildings*, ACI 301, American Concrete Institute, Farmington Hills, MI, 1996, 43 pages.
4. *Guide to Durable Concrete*, ACI 201.2R-92, American Concrete Institute, Farmington Hills, MI, 1992, 39 pages.
5. Wood, S.L., Evaluation of the Long-Term Properties of Concrete. Research and Development Bulletin RD102T, Portland Cement Association, Skokie, IL, 1992.

12.6 Composites

Introduction

Victor A. Greenhut

A composite material is a macroscopic, physical combination of two or more materials in which one material usually provides reinforcement. Composites have been developed where no single, quasi-continuous material will provide the required properties. In most composites one phase (material) is continuous and is termed the *matrix*, while the second, usually discontinuous phase, is termed the *reinforcement*, in some cases *filler* is applied when the reinforcement is not a quasi-continuous fiber. Matrix-filler nomenclature is one method of categorization. This yields the categories *metal matrix* (MMC), *polymer (plastic) matrix* (PMC), and *ceramic matrix* (CMC) composites — the major subdivisions of this section. Other categories are given the shape and configuration of the reinforcing phase. The reinforcement is usually a ceramic and/or glass. If it is similar in all dimensions, it is a *particulate-reinforced* composite; if needle-shaped single crystals, it is *whisker-reinforced*; if cut continuous filament, *chopped fiber-reinforced*; and if continuous fiber, *fiber composite*. For fiber composites configuration gives a further category. If fibers are aligned in one direction, it is a *uniaxial fiber composite*; if arranged in layers, it is a *laminar composite*; if a three-dimensional arrangement, it is a *3D weave composite*. Laminates and 3D weaves can be further divided by the weave used for the fiber.

Composites are not new materials. Perhaps the first important engineering structural composite was the Biblical straw-reinforced, sun-dried mud brick — adobe. Laminated structures such as bows have been used since prehistoric times. In the early 1900s doped fabric was employed in early aircraft surfaces. Reinforced phenolics were developed in the 1930s and glass-reinforced plastics in the 1940s. More recently, emphasis turned to reinforcements, with graphitic and boron-based fibers developed in the 1960s. High-performance aramids, such as Kevlar™, were developed in the 1970s. This and the previous decade have seen new developments in both fiber and matrix with lightweight aerospace MMCs and high-temperature CMCs showing major advances.

This section will concentrate on advanced composites in which major structural performance is achieved, but it should be noted that the largest tonnage and dollar value of composites is driven by the cost of materials. Polymers are reinforced with glass particulate or chipped fiber because it lowers the price per pound of a plastic, hence the term *filler* for reinforcement material. The structural properties are also improved, as PMCs show increased stiffness (resistance to bending) and strength (resistance to failure). These and other properties will receive the emphasis in this section. In MMCs high hardness, maintenance of hardness at elevated temperature, and precise tolerance are often lead properties, although stiffness and strength are important. The metal matrix confers some ductility so that the catastrophic failure of a solely ceramic material is avoided. In CMCs the reinforcement is chiefly incorporated to prevent catastrophic failure to introduce a more predictable failure stress (increased reliability), and to a lesser extent to increase strength. Particulate-reinforced plastics and fiberglass (glass fiber–reinforced plastic) have become commonplace engineering materials, as have ceramic-reinforced metal matrices (cemented carbides and cermets). Reinforcement with graphite and other advanced technology fibers has emerged for aerospace applications, biomedical use, and high-performance/high-cost consumer products such as sporting goods. Fiber-reinforced MMCs and ceramic–ceramic composites have developed greatly, but are still quite limited in application to very advanced technologies which can bear relatively high costs.

An important consideration for composite production is the bond between the matrix and the reinforcement. For MMCs and PMCs load must be transferred (Figure 12.6.1) to the relatively high-strength, high-elastic-modulus ceramic or glass in order to maximize the mechanical performance. For best transfer there must be no relative sliding or interface failure, so the strain at the matrix-to-reinforcement area can be maximum and equal ($\varepsilon_{matrix} = \varepsilon_{reinforcement} = \varepsilon_{m,c}$). This means that the reinforcement bears the major portion of the load, σ_c, while the weaker matrix bears a lower load, σ_m, because it has lower modulus.

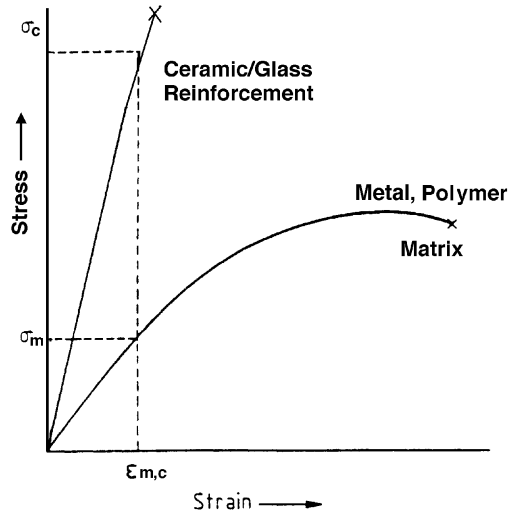


FIGURE 12.6.1 In a composite with good bonding between high-modulus ceramic and low-modulus metal or polymer matrix the interfacial strain must be the same. Thus, the high-strength ceramic fiber bears most of the load (stress).

To accomplish load transfer the bond interface must be engineered because metal and polymers do not show intrinsic bonding to ceramic reinforcement. A *coupling agent*, often a xylene, is applied to the ceramic or glass to promote bonding to a polymer. For metals, a metal matrix often has to be selected because it will wet and bond to the selected ceramic when molten or sintering during fabrication. Commonly, a controlled "forming gas" atmosphere must be used to produce MMCs composites so that a strong metal-to-ceramic bond is made.

The situation is somewhat different for ceramic–ceramic composites. These materials usually bond strongly to each other during fabrication. However, unlike the PMCs and MMc a less perfect bond is required. This is because local failure is required between matrix and reinforcement to introduce toughness or noncatastrophic fracture. This will be discussed further in the section on CMCs.

Polymer Matrix Composites

Weiping Wang, R. Allan Ridilla, and Mathew B. Buczek

Introduction

PMCs are used extensively as commodity and specialty engineering materials and constitute the largest class of composite materials on a dollar basis. The enormous success of PMCs arises from the wide range of properties which can be obtained and the low cost of low-end materials, such as fiberglass-reinforced polyester. Furthermore, the combination of high-strength/high-stiffness fibers in commodity polymer matrices offers an outstanding example of the composite principle — taking the best properties of both materials. The result is a strong, tough, stiff material which, depending on the matrix and reinforcement type, provides value in applications as diverse as consumer products, construction, and aerospace.

A distinction must be made between reinforced plastics and advanced composites. The term *reinforced plastics* generally refers to plastic materials fabricated with a relatively low percentage of discontinuous, randomly oriented fibers, with rather moderate properties, and used in commodity applications. *Advanced composites*, on the other hand, refers to a class of materials where a high percentage of continuous, highly oriented fibers are combined with a suitable polymeric matrix to produce articles of high specific strength and specific stiffness. At present, composites are widely used in the aerospace and sporting

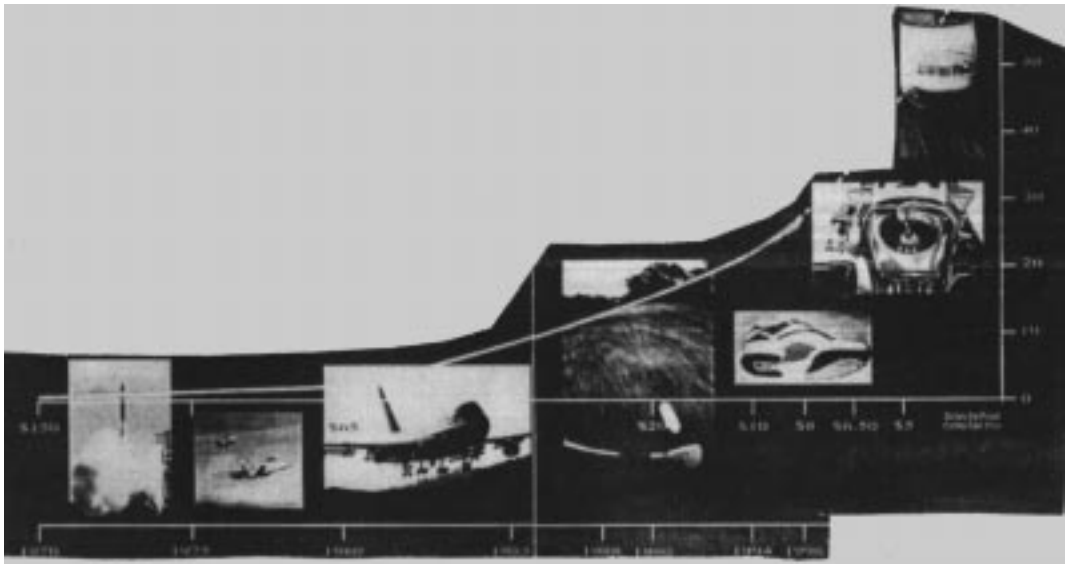


FIGURE 12.6.2 The evolution of carbon fiber composite applications since 1970 with fiber price and industry volume illustrated. (Source: Zoltek Co. Inc.)

goods industries for their superior performance. Figure 12.6.2 illustrates the evolution of applications for carbon fibers.

Polymers are also often "filled" with particulate reinforcements to increase certain properties, such as deflection temperature under load (DTUL). Although these materials are formally composites, they are usually regarded as filled polymers and not composites and thus will not be discussed explicitly in this section. Refer to the polymer section of this chapter.

Architecture

Each of the constituent materials in advanced composites must act synergistically to provide aggregate properties that are superior to the materials individually. The functional effectiveness of composites is principally due to the anisotropy of the materials and the laminate concept, where materials are bonded together in multiple layers. This allows the properties to be tailored to the applied load so that the structure can be theoretically more efficient than if isotropic materials were used. The reinforcements come in a variety of forms. Unidirectional tapes with all fibers along a common axis, woven fabrics constructed with fibers along both axes in the x–y plane, and 3-D architectures with reinforcements in more than two axial directions are just a few of the building blocks of composite structures.

The concept of laminate is illustrated in Figure 12.6.3. On the left is the unidirectional composite where all reinforcements are aligned in one direction. This construction provides excellent properties in the fiber direction but is limited to the properties of the resin in the transverse directions. The cross-plied construction on the right creates a structure that has common properties in the x- and y-direction but is limited to the characteristics of the resin in the z-direction. A large number of variations exist, allowing the designer to tailor the properties of the structure by varying the type of fibers and fiber orientations. Shown in Figure 12.6.4 are typical values of specific tensile strength (strength-to-density ratio) and specific tensile modulus (modulus-to-density ratio) for 65% volume fraction, cross-plied quasi-isotropic composites. Also included in the plot for comparison are properties of mild steel, titanium (Ti-6A1-4V), and aluminum (2024-T6). The actual properties of a composite will vary significantly due to flaws created during processing.

In selecting polymer composite materials for design, one should consider both the potential advantages and concerns typical to composites as illustrated in Table 12.6.1. Clearly, the advantages of materials should be put into use to achieve higher levels of performance in many kinds of products. On the other

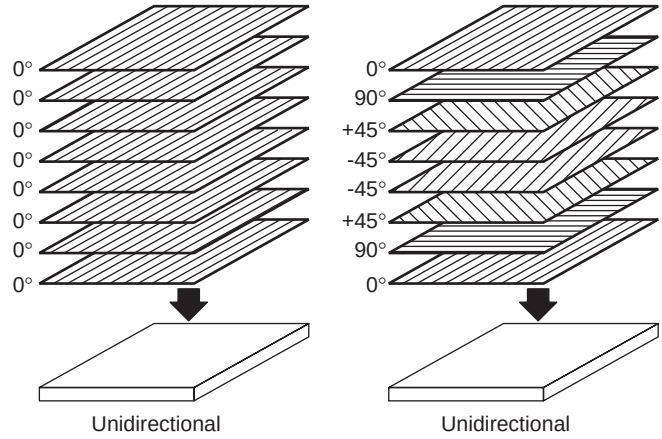


FIGURE 12.6.3 The basic concept of composite laminate.

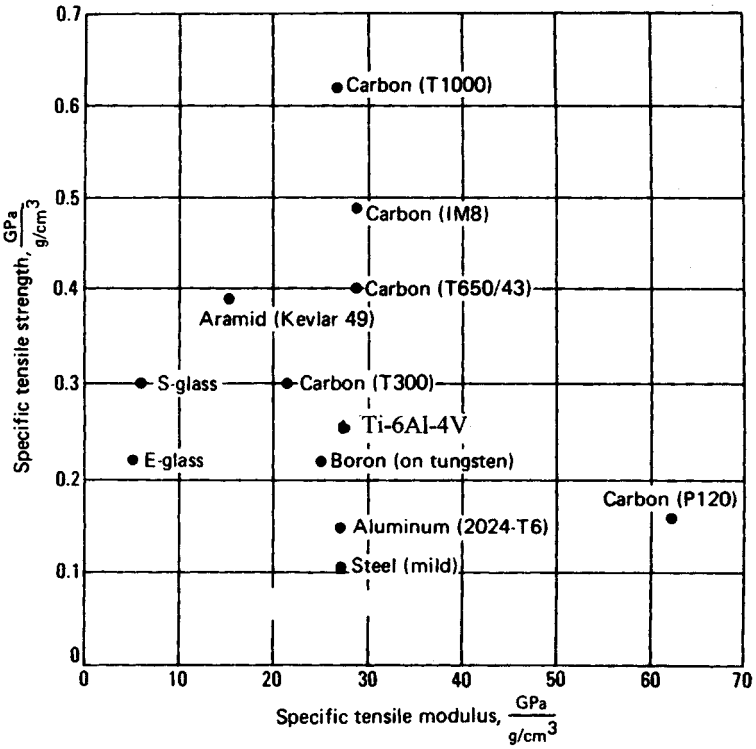


FIGURE 12.6.4 A comparative plot of specific tensile strength (strength-to-density) and specific tensile modulus (modulus-to-density) of composite material.

hand, potential issues in higher raw material cost and lower production volume and yield should also be weighed.

Fiber

Introduction. The most common reinforcement in polymer composites are fibers. Fibrous reinforcements come from compounds of light elements, (e.g., B, C, Si, O). These compounds typically contain

TABLE 12.6.1 Advantages and Limitations of Polymer Matrix Composites

Potential Advantages	Potential Limitations
High strength-to-density ratio	Low–volume production methods
High stiffness-to-density ratio	High raw material cost
Excellent corrosion resistance	Poor impact resistance
Good fatigue resistance	Poor high-temperature performance
Low thermal expansion	Delamination/out-of-plane loading

stable covalent bonds which impart greater strength and stiffness compared with metallic or ionic bonds. The compounds are processed to the final unable form of a fiber or filament with a highly aligned and directional microstructure so that the strength and stiffness properties are optimized along the fiber axis. Glass, graphite/carbon, aramid, and boron are among most notable fibers currently used in polymer composites.

Glass Fiber. Glass fiber reinforcements represent the largest volume used in the composites industry. These fibers are characterized by their low cost, clear to white color, good mechanical and electrical properties, high moisture and chemical resistance, and excellent dimensional stability with operational service to 550°C. Manufacturing of the fibers begins with molten glass which is drawn through a furnace into a fibrous form of final diameter of about 10 μm and then quenched to secure the final amorphous microstructure prior to applying final coatings or sizing. Common types of commercially available glass fiber are E-glass and S-glass, both of which are low alkali boro-alumino-silicate glasses. E-glass fiber, the workhorse of glass fiber applications, is the lower-cost fiber and is used in both structural and electrical applications. S-glass provides higher tensile properties and increased temperature resistance needed for aerospace and aircraft applications with a price premium. Representative properties for the glass fibers are shown in Table 12.6.2.

TABLE 12.6.2 Fibers Used in Polymer Composites — Mechanical Properties

	E-Glass	S-Glass	AS4 PAN-Based Carbon	IM7 PAN-Based Carbon	P120 Pitch-Based Graphite	Kevlar-49	Boron
Tensile strength (ksi)	510	670	578	710	325	530	525
Tensile modulus (MSI)	10.5	12.8	35.5	46	120	18	58
Elongation (%)	4.9	5.5	1.6	1.7	0.27	2.5	1
Density (lb/in^3)	0.095	0.09	0.065	0.063	0.079	0.052	0.093
Axial coefficient of expansion (10^{-6} in./in. F)	2.8	3.1	0 to –0.4	0 to –0.6	0 to –0.7	–1.1	2.5

Carbon Fiber. Graphite/carbon fibers are the reinforcement of industrial choice for advanced composite applications where stiffness and performance are critical. Typical attributes of these fibers are excellent tensile strength and elastic modulus, ease of handling, black color, and a wide range of properties and cost. Graphite fibers are initially formed as from polymer precursor compounds such as polyacrylonitrile (PAN) or rayon and pitch, an amorphous, aromatic by-product of petroleum distillation. PAN-based graphite fiber, the predominant commercial fiber, starts with the liquid PAN polymer, which is spun into a fiber and stretched to align the microstructure. It is then stabilized and the microstructure is aligned at 400 to 500°F in an oxidizing atmosphere under tension. Next, carbonization occurs in an inert atmosphere at 1800 to 2300°F to remove most noncarbon elements. Finally, the graphitization step is performed on the carbon fiber by applying tension to the fiber in an inert atmosphere at 3600 to 6500°F. The result is a highly aligned, highly graphitic fiber with preferred graphite orientation along the fiber axis. The temperature and tension fabrication parameters of the graphitization step, along with the purity of the initial PAN polymer, are the variables which are modified to differentiate “low-end” (low strength,

modulus, cost) from the "high-end" fiber (high strength, modulus, cost) and all of the grades in between. Typical properties of several commercial grades of carbon/graphite fiber are listed in Table 12.6.2.

Aramid Fibers. The aramid fiber derives its properties from long, rodlike aromatic polyamides. These fibers are characterized by low density, white yellow color, extremely high tensile properties, poor compressive strength, and high toughness, and they are difficult to cut mechanically. The manufacturing process for these fibers involves complex polymerization steps as the liquid polymer in acid is extruded or spun into a fiber form, water washed, dried, and finally heat treated under tension. The result is a highly aligned radial system of axially pleated lamellae in the microstructure. Aramid fibers have proved extremely useful in tension-critical application where the intrinsic compressive weaknesses of the fiber cannot be exploited. Commercially known as Kevlar, the most notable applications for aramid fibers are bulletproof vests (without matrix), and various high-pressure composite vessels. Typical properties for Kevlar 49 are listed in Table 12.6.2.

Boron Fibers. Boron reinforcements are referred to as filaments rather than fibers since they are made by the chemical vapor deposition of boron onto a fine tungsten wire. This fabrication process produces a large-diameter, stiff, and expensive reinforcement that is somewhat difficult to handle and produce into subsequent product forms such as fabrics and contoured structures. Although boron filaments possess the combination of high strength and high specific modulus that glass could not achieve, its use has been reduced to a minuscule level. Carbon/graphite fiber varieties have supplanted boron where high specific modulus is a requirement. A small volume of boron filaments still remain in a handful of military aircraft as well as in some recreational products such as golf club shafts. Typical properties of boron filaments are shown in Table 12.6.2.

Polymer Matrix Materials

Introduction. While fibers provide much of the strength and stiffness in advanced composites, equally important is the matrix resin. The matrix holds the fiber network together, to protect the structure from environmental attack and to support the fibers so that loads can be transferred throughout the structure through a shear mechanism. Polymeric matrices can either be thermoset or thermoplastic resins.

Thermosets cannot be reformed or thermally reworked after polymerization (curing). They are subdivided into categories based on their chemical reactions. The first are addition-type polymers, generally considered easier to process, where two (or more) reactants combine to form the final cured product. The second is the condensation type, in which the reactants combine to form products in addition to water and other volatile constituents. These systems are more difficult to process due to the required management of the volatile to minimize process defects such as porosity.

Thermoplastic materials, on the other hand, are polymers which can be softened and melted with the application of heat. Although they can be recycled, thermoplastic matrices have found more-limited applications in the advanced composite applications since they are particularly susceptible to attack by fluids and to creep at high stresses, and are relatively difficult to handle in laminate or structural form prior to final consolidation. The remainder of this discussion will focus on thermoset matrices which account for more than 90% of present-day advanced composite applications. Three principal organic matrix materials will be discussed: polyesters, epoxies, and polyimides.

Polyester Matrix. Low-cost polyester resins constitute the highest volume usage for the general composite industry. There are many resins, formulations, curatives, and other additive constituents that provide a wide array of properties and performance characteristics, such as mechanical strength, toughness, and heat resistance. The vast majority of applications utilize glass since its interfacial adhesion to these resins has been optimized by the development of silane surface treatments. These resins when combined with glass have found wide application in the chemical processing construction, and marine industries where cured properties and low cost are tailor-made for design requirements. Applications with carbon/graphite and aramid fibers are far less frequent since adhesion to these fibers is generally

poor and cure shrinkage on these resins is quite high. Additionally, structures requiring the high performance of carbon/graphite can often justify the use of more-expensive epoxy resins.

Epoxy Matrix. Epoxy resins are the prevalent polymer used with advanced composites. Their extensive use is primarily due to their superior mechanical properties, excellent adhesion, good processibility utilizing addition-type reactions, low cure shrinkage and low cost. When choosing epoxy resins two elements are crucial: the base resin and the curing agent or hardener. Since there are many base resins and curing agents, the following general relationships are given to provide a practical framework when utilizing epoxy compounds:

- *Chemistry*: Aromatic compounds are more thermally stable, stronger, and more resistant to fluids than are aliphatics.
- *Cross-link density*: Higher-temperature cures and longer cure times will increase cross-link density and therefore strength/modulus, service temperature, and chemical resistance.
- *Curing agents*: Amine-cured systems have better chemical resistance and superior thermal stability, but are more brittle than anhydride-cured systems.

Epoxies are categorized by the cure temperature. The "250 F epoxies" are those that cure at 250°F and are suitable for service up to 250°F. The "350 F epoxies" are those with higher processing and service temperatures. One of the design concerns with epoxies is moisture absorption. The effects of moisture often require design stresses to be reduced in applications where moisture is a concern. Modern epoxy formulations include lower cost, higher toughness, and other properties which continue to make epoxies attractive to the end user.

Polyimide resins span the temperature spectrum from 350 to 600°F. There are three general types of polyimides:

1. *Addition-reaction* polyimides such as bismaleimides (BMI)
2. *Condensation-reaction* polyimides such as commercial Monsanto Skybond resin
3. *Combination* condensation/addition-reaction polyimides such as PMR-15

BMIs are similar to epoxies in that they undergo addition reactions, are easy to process, and share many of the handling characteristics that make epoxies desirable. These materials are the composite resins of choice for temperatures in the range of 350 to 450°F to bridge the temperature gap between epoxies and other polyimides. They are more expensive and tend to be more brittle than epoxies. Condensation polyimides are used for composite applications from 450 to 520°F. These materials are difficult to process and tend to be quire brittle so their application is limited to adhesive bonding with some structural composite hardware. Combination-reaction polyimides have the highest thermal and oxidative stability and are used at service temperatures to 600°F. These resins are considerably more complicated because of their reaction mechanisms and the handling difficulties of the chemicals. Their application is principally restricted to aerospace composite structures where the performance needs to justify the difficult processing and high cost.

Processing

Continuous-fiber composites are manufactured in two steps, preform fabrication and consolidation/curing. The material comes in either the dry-fiber (without resin) form or with resin included called *prepreg*. Dry fibers are used in filament winding, pultrusion, weaving, braiding. In the case of filament winding and pultrusion, resin is introduced at the same time fibers are during the time of winding and pultrusion. Resin can also be introduced to the fiber later by resin-transfer molding (RTM). Prepreg, at a higher material cost, eliminates the step of resin addition and provides the adhesion to hold the material together. Layup of prepreg, a time-consuming process, is a common method used in the aerospace industry where complex contoured surfaces are present.

Consolidation and curing consist of compacting the material to remove entrapped air, volatile, and excess resins while developing the structural properties by increasing the polymer chain length and

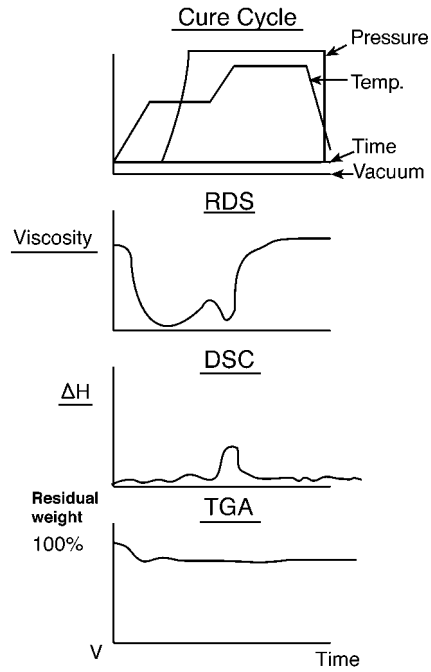


FIGURE 12.6.5 A typical curing cycle of epoxy composite with corresponding Rheometric Dynamic Spectroscopy (RDS), Differential Scanning Calorimetry (DSC), and Thermogravimetric Analysis (TGA) curves.

cross-linking. Thermosetting polymer matrices must be cured *in situ* with the fibers to form composites. The goals of a successful cure are good consolidation with low porosity and high conversion of initial monomeric constituents to polymer (degree of cure).

The cure cycle of an epoxy resin composite is shown in Figure 12.6.5. At the start of the cure cycle, the material is normally under vacuum to remove residual volatile. The temperature is then ramped to the point where the polymer is melted. The ramp rates on heating must be slow enough as not to cause unnecessary thermal gradients and to avoid dangerous exotherm (runaway reactions) but not so slow as to cause excessive process time or premature cross-linking. Normal ramp rates are generally in the 1 to 5°F/min range. An isothermal hold of about 30 to 60 min is performed at the point where the resin reaches its minimum in viscosity, during which pressure is applied and the polymer allowed to flow, thus consolidating the laminate. Following consolidation, the temperature is increased to the point where cross-linking occurs (350°F for this example). An additional hold of about 30 to 90 min is performed at this temperature to allow for the material to complete cross-linking. This event is shown as the end of the exothermic peak in the DSC curve as well as the asymptotic high-viscosity region.

Epoxies cure via addition reactions where no volatile is generated during the cross-linking process. However, many other matrix materials (e.g., phenolics and many polyimides) cure via condensation reactions.

Mechanical Properties

Mechanical properties of polymer composites depend substantially on the processing and fabrication methods used, as well as on the fiber orientation. Thus, standard materials do not exist and it is difficult to generalize regarding properties. The three most significant factors in determining properties are the type of fiber, the volume fraction fiber, and the orientation of the fiber. High-strength high-modulus graphitic carbon fibers will, of course, produce stronger, stiffer composites than those produced from fiberglass. Similarly, 60 vol % fiber composites will be stronger and stiffer than 30 vol %, and uniaxially aligned fiber composites will have maximum properties along the alignment axis, but will be highly

anisotropic and will have poor properties in off-axis orientations. Cross-ply and laminated structures are standard approaches to reducing anisotropy.

Composite stiffness, E_c, can be approximated for polymer composites by application of the rule of mixtures:

$$E_c = E_f V_f + E_m\left(1 - V_f\right)$$

where E is the modulus of elasticity and V is the volume fraction fibers. Subscripts c, f, and m refer to composite, fiber, and matrix, respectively. Similar expressions approximate other properties. Properties of commonly used commodity composite materials are given in Tables 12.6.3 to 12.6.5. Advanced engineering polymer composites have specialized properties which depend on the parameters discussed previously. Expected properties of such advanced polymer composites can only be approximated once the component and material variables are determined during the design stage. The recommended approach is to work with a composite fabricator during the design phase to establish the expected mechanical properties.

TABLE 12.6.3 Typical Properties of Glass Fiber–Reinforced Polymers

Property	Base Resin				
	Polyester	**Phenolic**	**Epoxy**	**Melamine**	**Polyurethane**
Molding quality	Excellent	Good	Excellent	Good	Good
Compression molding					
Temperature, °F	170–320	280–350	300–330	280–340	300–400
Pressure, psi	250–2000	2000–4000	300–5000	2000–8000	100–5000
Mold shrinkage, in./in.	0.0–0.002	0.0001–0.001	0.001–0.002	0.001–0.004	0.009–0.03
Specific gravity	1.35–2.3	1.75–1.95	1.8–2.0	1.8–2.0	1.11–1.25
Tensile strength, 1000 psi	25–30	5–10	14–30	5–10	4.5–8
Elongation, %	0.5–5.0	0.02	4	—	10–650
Modulus of elasticity, 10^{-5} psi	8–20	33	30.4	24	—
Compression strength, 1000 psi	15–30	17–26	30–38	20–35	20
Flexural strength, 1000 psi	10–40	10–60	20–26	15–23	7–9
Impact. Izod, ft-lb/in. or notch	2–10	10–50	8–15	4–6	No break
Hardness, Rockwell	M70–M120	M95–M100	M100–M108	—	M28–R60
Thermal expansion, per °C	$2–5 \times 10^{-5}$	1.6×10^{-5}	$1.1–3.0 \times 10^{-5}$	1.5×10^{-5}	$10–20 \times 10^{-5}$
Volume resistivity at 50% RH, 23°C, Ω-cm	1×10^{14}	7×10^{12}	3.8×10^{15}	2×10^{11}	$2 \times 10^{11}–10^{14}$
Dielectric strength $^1/_8$ in. thickness, v/mil	350–500	140–370	360	170–300	330–900
Dielectric constant					
At 60 Hz	3.8–6.0	7.1	5.5	9.7–11.1	5.4–7.6
At 1 kHz	4.0–6.0	6.9	—	—	5.6–7.6
Dissipation factor					
At 60 Hz	0.01–0.04	0.05	0.087	0.14–0.23	0.015–0.048
At 1 kHz	0.01–0.05	0.02	—	—	0.043–0.060
Water absorption, %	0.01–1.0	0.1–1.2	0.05–0.095	0.09–0.21	0.7–0.9
Sunlight (change)	Slight	Darkens	Slight	Slight	None to slight
Chemical resistance	Fair[a]	Fair[a]	Excellent	Very good[b]	Fair
Machining qualities	Good	—	Good	Good	Good

Note: Filament-wound components with high glass content, highly oriented, have higher strengths. The decreasing order of tensile strength is: roving, glass cloth, continuous mat, and chopped-strand mat.

[a] Attacked by strong acids or alkalies.

[b] Attacked by strong acids.

From Spang, C.A. and Davis, G.J., *Machine Design*, 40(29); 32, Dec. 12, 1968. With permission.

TABLE 12.6.4 Properties of Reinforced Nylon Polymer

Property	Type 6/6	Type 6	Type 6/10	Type 11	Glass-Reinforced Type 6/6, 40%	MoS_2-filled, $2^1/_2$%	Direct Polymerized, Castable
Mechanical							
Tensile strength, psi	11,800	11,800	8200	8500	30,000	10,000–14,000	11,000–14,000
Elongation, %	60	200	240	120	1.9	5–150	10–50
Tensile yield stress, psi	11,800	11,800	8500	—	30,000	—	—
Flexural modulus, psi	410,000	395,000	280,000	151,000	1,800,000	450,000	—
Tensile modulus, psi	420,000	380,000	280,000	178,000	—	450,000–600,000	350,000–450,000
Hardness, Rockwell	118R	119R	111R	55A	75E–80E	110R–125R	112R–120R
Impact strength, tensile, ft-lb/sq in.	76	—	160	—	—	50–180	80–100
Impact strength, Izod. ft-lb/in. of notch	0.9	1.0	1.2	3.3	3.7[a]	0.6	0.9
Deformation under load, 2000 psi, 122°F, %	1.4	1.8	4.2	2.02[b]	0.4[c]	0.5–2.5	0.5–1
Thermal							
Heat deflection temp, °F							
At 66 psi	360	365	300	154	509	400–490	400–425
At 264 psi	150	152	135	118	502	200–470	300–425
Coefficient of thermal expansion per °F	4.5×10^{-5}	4.6×10^{-5}	5×10^{-5}	10×10^{-5}	0.9×10^{-5}	3.5×10^{-5}	5.0×10^{-5}
Coefficient of thermal conductivity, Btu in./hr ft^3 °F	1.7	1.7	1.5	—	—	—	—
Specific heat	0.3–0.5	0.4	0.3–0.5	0.58	—	—	—
Brittleness temp, °F	–112	—	–166	—	—	—	—
Electrical							
Dielectric strength, short time, v/mil	385	420	470	425	480	300–400	500–600[d]
Dielectric constant							
At 60 Hz	4.0	3.8	3.9	—	4.45	—	3.7
At 10^3 Hz	3.9	3.7	3.6	3.3	4.40	—	3.7
At 10^6 Hz	3.6	3.4	3.5	—	4.10	—	3.7
Power factor							
At 60 Hz	0.014	0.010	0.04	0.03	0.009	—	0.02
At 10^3 Hz	0.02	0.016	0.04	0.03	0.011	—	0.02
At 10^6 Hz	0.04	0.020	0.03	0.02	0.018	—	0.02
Volume resistivity, Ω-cm	10^{14}–10^{15}	3×10^{15}	10^{14}–10^{15}	2×10^{13}	2.6×10^{15}	2.5×10^{13}	—
General							
Water absorption, 24 hr, %	1.5	1.6	0.4	0.4	0.6	0.5–1.4	0.9

TABLE 12.6.4 Properties of Reinforced Nylon Polymer (continued)

Property	Type 6/6	Type 6	Type 6/10	Type 11	Glass-Reinforced Type 6/6, 40%	MoS_2-filled, $2^1/_2$%	Direct Polymerized, Castable
Specific gravity	1.13–1.15	1.13	1.07–1.09	1.04	1.52	1.14–1.18	1.15–1.17
Melting point, °F	482–500	420–435	405–430	367	480–490	496 ± 9	430 ± 10
Flammability	Self-extinguishing	Self-extinguishing	Self-extinguishing	Self-extinguishing	Self-extinguishing	Self-extinguishing	Self-extinguishing
Chemical resistance to							
Strong acids	Poor	Poor	Poor	Poor	Poor	Poor	Poor
Strong bases	Good	Good	Good	Fair	Good	Good	Good
Hydrocarbons	Excellent	Excellent	Excellent	Good	Excellent	Excellent	Excellent
Chlorinated hydrocarbons	Good	Good	Good	Fair	Good	Good	Good
Aromatic alcohols	Good	Good	Good	Good	Good	Good	Good
Aliphatic alcohols	Good	Good	Good	Fair	Good	Good	Good

Notes: Most nylon resins listed in this table are used for injection molding, and test values are determined from standard injection-molded specimens. In these cases a single typical value is listed. Exceptions are MoS_2-filled nylon and direct-polymerized (castable) nylon, which are sold principally in semifinished stock shapes. Ranges of values listed are based on tests on various forms and sizes produced under varying processing conditions.

Because single values apply only to standard molded specimens, and properties vary in finished parts of different sizes and forms produced by various processes, these values should be used for comparison and preliminary design considerations only. For final design purposes the manufacturer should be consulted for test experience with the form being considered. Listed values should not be used for specification purposes.

[a] $^1/_2 \times {}^1/_4$-in. bar.

[b] 2000 psi, 73°F.

[c] 4000 psi, 122°F.

[d] 0.040-in. thick.

From Carswell, D.D., *Machine Design*, 40(29), 62, Dec. 12. 1968. With permission.

TABLE 12.6.5 Comparative Properties of Reinforced Plastics

Property	Polyamide		Polystyrene[a]		Polycarbonate		Styrene Acrylonitrile[b]		Polypropylene		Acetal		Linear Polyethylene	
	U	R	U	R	U	R	U	R	U	R	U	R	U	R
Tensile strength, 1000 psi	11.8	30.0	8.5	14.0	9.0	20.0	11.0	18.0	5.0	6.6	10.0	12.5	3.3	11.0
Impact strength, notched, ft-lb/in.														
At 73°F	0.9	3.8	0.3	2.5	2.0[c]	4.0[c]	0.45	3.0	1.3–2.1	2.4	60.0	3.0	—	4.5
At –40°F	0.6	4.2	0.2	3.2	1.5[c]	4.08[c]	—	4.0	—	2.2	—	3.0	—	5.0
Tensile modulus, 10^5 psi	4.0	—	4.0	12.1	3.2	17.0	5.2	15.0	2.0	4.5	4.0	8.1	1.2	9.0
Shear strength, 1000 psi	9.6	14.0	—	9.0	9.2	12.0	—	12.5	4.6	4.7	9.5	9.1	—	5.5
Flexural strength, 1000 psi	11.5	37.0	11.0	20.0	12.0	26.0	17.0	26.0	6–8	7.0	14.0	16.0	—	12.0
Compressive strength, 1000 psi	4.9[d]	24.0	14.0	17.0	11.0	19	17.0	22.0	8.5	6.0	5.2	13.0	2.7–3.6	6.0
Deformation, 4000-psi load, %	2.5	0.4	1.6	0.6	0.3	0.1	—	0.3	—	6.0	—	1.0	—	0.4[e]
Elongation, %	60.0	2.2	2.0	1.1	60–100	1.7	3.2	1.4	>200	3.6	9–15	1.5	60.0	3.5
Water absorption, in 24 hr, %	1.5	0.6	0.03	0.07	0.3	0.09	0.2	1.15	0.01	0.05	0.20	1.1	0.01	0.04
Hardness, Rockwell	M79	E75–80	M70	E53	M70	E57	M83	E65	R101	M50	M94	M90	R64	R60
Specific gravity	1.14	1.52	1.05	1.28	1.2	1.52	1.07	1.36	0.90	1.05	1.43	1.7	0.96	1.30
Heat distortion temperature, at 264 psi, °F	150	502	190	220	280	300	200	225	155	280	212	335	126	260
Coefficient of thermal expansion, per F × 10^{-5}	5.5	0.9	4.0	2.2	3.9	0.9	4.0	1.9	4.7	2.7	4.5	1.9	9.0	1.7
Dielectric strength, short time, v/mil	385	480	500	396	400	482	450	515	750	—	500	—	—	600
Volume resistivity Ω-cm × 10^{15}	450	2.6	10.0	36.0	20.0	1.4	10^{16}	43.5	17.0	15.0	0.6	38.0	10^{15}	29.0
Dielectric constant, at 60 Hz	4.1	4.5	2.6	3.1	3.1	3.8	3.0	3.6	2.3	—	—	—	2.3	2.9
Power factor, at 60 Hz	0.0140	0.009	0.0030	0.0048	0.0009	0.0030	0.0085	0.005	—	—	—	—	—	0.001
Approximate cost, ¢/in.[3]	3.0	8.0	0.5	2.5	3.6	6.5	0.9	3.5	0.6	2.1	3.3	7.8	0.7	3.1

Note: U = reinforced; R = reinforced. Multiply tabular values in psi by 6895 to obtain N/m^2.

[a] Medium-flow, general-purpose grade.

[b] Heat-resistant grade.

[c] Impact values for polycarbonates are a function of thickness.

[d] At 1% deformation.

[e] 1000-psi load.

From Lachowecki, W., *Machine Design*, 40(29), 34, Dec. 12, 1968. With permission.

Metal Matrix Composites

Introduction

The term *metal matrix composites* is usually reserved for fiber-reinforced materials, although technically it applies to particulate-reinforced systems also. The general term for particulate-, whisker-, and chopped-fiber-reinforced systems is *cermet*. A majority of commonly used systems are based upon particulate reinforcements.

Cermets and Cemented Carbides — Particulate-Reinforced MMCs

The term *cemented carbide* is applied to particulate composites based upon tungsten carbide in a cobalt matrix. This distinction is largely historical, not technical. The tungsten carbide–cobalt-based materials were first developed in Germany, while development of other cermets (including noncarbide-, oxide-, and nitride-based systems) occurred in the United States during and immediately after World War II. Thus, a distinguishing nomenclature was developed both because "carbide" was no longer completely appropriate with oxide systems available and for geopolitical reasons. This nomenclature confusion between the most common cermets, the cemented carbides, and other materials still survives. The structure consists of a continuous metal alloy functioning as cement which holds together particles of carbide, oxide, or nitride ceramic. These materials provide strength, high hardness, wear (abrasion) resistance, low sliding friction, and precise tolerance even at quite high temperatures. Cermets and cemented carbides derive their usefulness when the compressive strength, hardness, and thermal resistance of the ceramic reinforcement is coupled with the ductility, toughness, adhesion, and lubricity provided by the metal. Common applications include metal and rock cutting and grinding tools, high-temperature containers, and pouring spouts, rocket nozzles, turbine parts, flame nozzles, friction and glide parts, seals, magnetron tube cathodes, flame nozzles, and ballpoint pen tips.

A cermet is usually fabricated by mixing the ceramic and metal powders together with a binder incorporated. This often requires high-energy mixing because of density differences with tooling made of the same cermet to avoid contamination. Subsequently, the part is formed, commonly by unidirectional or isostatic pressing, to final shape (accounting for shrinkage) with a wear-resistant die and with the formed object then liquid phase sintered in a controlled-atmosphere furnace. The chemistry is often quite complex in order to control the structure and the interfacial bond between metal and ceramic. Molybdenum promotes wetting of titanium carbide by molten nickel. Tantalum and titanium carbides are usually added to tungsten carbide–cobalt cemented carbides. Some cermets are prepared by slip casting, liquid metal infiltration, hot pressing, or solid-state sintering of parts. In some cases the material is "deficient" in metal to provide porosity for the passage of ink (pen tips) or incorporation of lubricant.

As indicated in the introduction to this section, a good bond between ceramic grain and metal "cement" is required for proper load transfer and to provide a homogeneous structure for good load bearing and uniform deformation. This normally requires a good wetting between liquid metal and ceramic grains. This is affected by minor alloying additions to the metal and the firing atmosphere.

The distribution, stability, and composition of phases have profound effects on properties. The production of WC–6 $^{w}/_{o}$ Co is a good example of the need for strict compositional control. The tertiary-phase diagram shows that the composition range from 6.00 to 6.12 $^{w}/_{o}$ must be adhered to in order to prevent formation of phases other than WC and Co — a composition range of less than 0.12 $^{w}/_{o}$! Embrittlement by graphite occurs at higher carbon content and by eta phase (η–WC_{1-x}) at lower carbon levels. Precise formulation and furnace atmosphere control is required, usually assisted by using a graphite-lined furnace. About 0.1% outside the acceptable composition range there is a drop of about 25% from the optimum strength.

It is important to engineer the microstructure of cermets in order to optimize properties. A uniform dispersion of round oxide, carbide, or nitride grains in the matrix of metal binder is desired. Inhomogeneity and sharp grain corners can act as failure origins. Round grains result from powder preparation and/or partial dissolution of sharp contours in the molten metal. As the amount of binder increases (Figure 12.6.6), the indentation hardness of the composite decreases, because metal, compared with

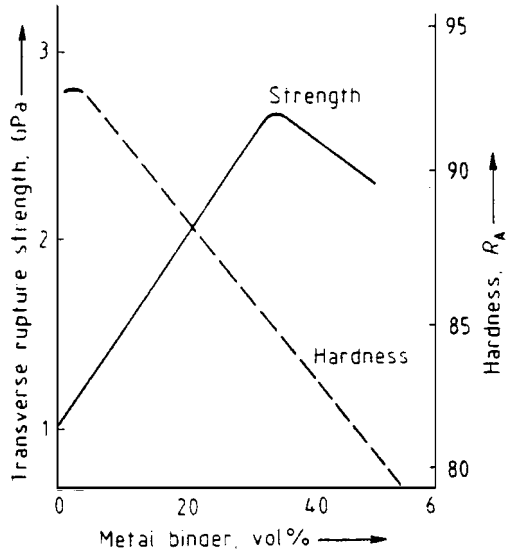


FIGURE 12.6.6 Effect of increased binder on strength and hardness of a TiC-Ni-Mo cermet with a 1 μm average carbide particle size.

ceramic, has a lower elastic modulus and will deform more easily. Fracture strength increases with binder because the metal phase confers fracture toughness. Generally, as ceramic grain size decreases, hardness and resistance to deformation increase for a fixed relative amount of ceramic. However, if the surface area increases sufficiently because of small grain size, there may not be sufficient metal to wet all the grain.

With these diverse factors a simplified way of looking at microstructural features and adjusting them for optimum performance is the mean free path, the average distance between neighboring ceramic grains (the average thickness of the binder phase). The mean free path depends on the size and shape of the grains and the amount of metal matrix available. Hardness, elevated-temperature hardness, strength, and toughness depend on the mean free path because it is the zone for absorption of crack energy and the medium for load transfer from grain-to-grain. Figure 12.6.7 shows that hardness varies with mean free path for a variety of cobalt contents and tungsten carbide grain sizes in cemented carbide. The curve does not continue below 0.1 μm because the metal binder becomes discontinuous.

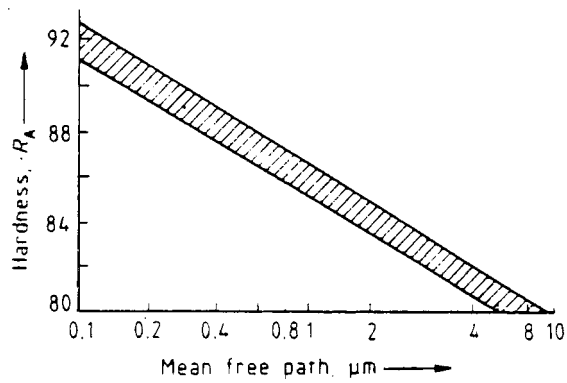


FIGURE 12.6.7 Effect of mean free path on hardness for a cemented carbide.

TABLE 12.6.6 Particle-Reinforced Metals

A. Metals Strengthened by Dispersed Powders

	Strengthener		Stress, kpsi	
Matrix Metals	**Component**	**% vol**	**Matrix Only, No Reinforcement**	**Composite Material**
Pure iron	Al_2O_3	4	2.2[a]	10[a]
Pure iron	Al_2O_3	10	2.2[a]	21[a]
Pure copper	Al_2O_3	10	2.2[b]	20[b]
Platinum	ThO_2	12.5	0.6[c]	6.1[c]
Uranium	Al_2O_3	3.5	6[d]	14[d]
Uranium	Al_2O_3	7.5	6[d]	20[d]
Copper	W (powder)	60	(E = 18)[e]	(E = 34)[e]

[a] Stress for 100-hr rupture life, 650°C.
[b] Stress for 100-hr rupture life, 450°C.
[c] Stress for 100-hr rupture life, 1100°C.
[d] Stress for 10^{-4} in./in./hr creep rate.
[e] Modulus of elasticity, E = stress/strain.

B. Strength Ratios at High Temperatures

	Dispersed Particles		Strength by Test	
Matrix Metal	**Material**	**% by Volume (Range)**	**Range of Test Temperatures, °F**	**Strength Ratio, (Referred to Matrix at Same Temperature)**
Aluminum	Al_2O_3	10–15	400–800	2–4[a]
Copper	Al_2O_3	3–10	800–1100	5–10
Iron	Al_2O_3	8–10	1100–1300	5–10
Nickel	ThO_2	3	1600–2100	Strength higher than that of many super alloys
Platinum	ThO_2	12	2000–2400	8–11
Uranium	Al_2O_3	4–7	900–1100	2–3

[a] The tensile strength of the final product is two to four times the strength of the aluminum.

Table 12.6.6 presents data for selected metal matrix systems with particulate reinforcement, including strength values at elevated temperatures. A section of Table 12.6.7 provides stiffness values for cobalt and nickel strengthened with sintered carbides.

Fiber-Reinforced MMCs

Fiber-reinforced MMCs may still be classed as an emerging materials area. A number of applications have been practiced driven by the need for a lightweight, high-stiffness material and/or for elevated-temperature resistance in structures. These applications are typically for aerospace and similar advanced applications which can use a premium material. The limited scope of current practical application and the limitations placed on technology disclosure by the Export Control Act and International Traffic-in-Arms Regulations make this section brief.

High-strength/high-elastic-modulus fibers such as boron, silicon carbide, graphite, aluminum oxide, and tungsten metal are typically incorporated in light metal matrices such as aluminum, magnesium, and titanium. The specific modulus (elastic modulus divided by density) has been shown to be two to four times that of high-strength structural metals. The specific strength may be substantially higher than the metal. These properties are maintained to many hundreds of degrees above ambient, while those of many advanced structural metals deteriorate both in immediate heating and long-term creep. Some systems have been shown to have considerable fatigue resistance, showing several times the stress for equivalent cycle lifetimes. These properties are important for structural shapes in which light weight and elevated-temperature resistance can afford a premium. Some applications have been struts and tubes

TABLE 12.6.7 Fiber- and Particle-Reinforced Metals — Test Results on Composite Metals

	Strengthener		Stress, kpsi	
Matrix Metals	Components	% Vol	Matrix Only, No Reinforcement	Composite Material
Metals Strengthened by Fibers				
Copper	W fibers	60	20	200[a]
Silver	Al_2O_3 whiskers	35	10[b]	75[b]
Aluminum	Glass fibers	50	(23%)[c]	(94%)[c]
Aluminum	Al_2O_3	35	25[d]	161[d]
Aluminum	Steel	25	25[d]	173[d]
Nickel	B	8	70[d]	384[d]
Iron	Al_2O_3	36	40[d]	237[d]
Titanium	Mo	20	80[d]	96[d]
Metals Strengthened by Sintered Carbides				
Cobalt	WC	90	(E = 30)[e]	(E = 85)[e]
Nickel	TiC	75	(E = 31)[f]	(E = 55)[f]

[a] Tensile strength with continuous fibers.
[b] Tensile strength at 350°C; modulus of elasticity: Cu = 17, composite 42 (millions of psi).
[c] Percentage of tensile strength at room temperature retained when tested at 300°C.
[d] Tensile strength, room temperature.
[e] Modulus of elasticity, E, measured in compression; hardness, 90 R-A; compressive strength, about 600,000 psi.
[f] Modulus of elasticity, E measured in compression; hardness about 85 R-A.
Compiled from various sources.

for space structures, bicycle frames, turbine parts, propellers, and engine components. The fiber may be incorporated unidirectionally, as a woven 2D laminate or a 3D structure. Fabrication methods include diffusion bonding, pultrusion (hot isostatic drawing), hot rolling, molten metal infiltration, and casting.

Table 12.6.7 gives metal matrix strengthening values for fiber reinforcements added in volume percentages from 8 to 60%.

Ceramic Matrix Composites

Richard L. Lehman and Daniel J. Strange

CMCs are ceramic matrix materials, either oxide, carbide, nitride, boride, or similar material, reinforced with particulates, whiskers, or continuous fibers. The reinforcing phase may be of any material, but most interest is directed toward ceramic reinforcement media. The following sections address fiber- and whisker-reinforced ceramic composites, the materials which offer the greatest potential for ameliorating the brittleness of ceramics and for developing exceptional mechanical and structural properties.

Ceramic Matrix Fiber Composites

Introduction. Ceramic matrix fiber composites (CMFCs) are the focus of substantial research but of limited commercial application as of this writing. The research enthusiasm is stimulated by the potential of high strength and toughness, mechanical and chemical durability, hardness, and oxidation resistance, all at elevated temperatures in the range 1000 to 1400°C. The limited commercialization, which is principally in military and aerospace applications, stems from the substantial cost of the fabricated materials, often in the range of thousands of dollars per kilogram. Fiber and matrix materials and properties used in CMFCs are presented in Table 12.6.8.

TABLE 12.6.8 Ceramic Matrix Composites: Fiber and Matrix Properties

Material	*E*	*a*	*R*	UTS	*e**	*g*(*Kc*)	*T*	*r*
Nicalon	180–200	3.1–4.0	7.5	2.5–3.0	1.4	(2.3)	1300	2.55
HM-carbon	345–414	–8–0.8	4	1.7–2.8	0.7	—	400, 3500	1.91
SCS-6 (SiC)	350–415	3.6	70	3.45	0.83	—	1300	3.3
FP-Al2O3	380	5.7	10	1.4	0.37	—	1316	3.9
Nextel 440	180	3.5	6	2.7	1.2	—	1426	3.1
Tyranno	193	3.1	5	2.76	1.5	—	1300	2.5
Ca-aluminosilicate	88–89	4.5–5.0	—	0.08–0.17	0.15	25 (2.2)	1350	2.8
Borosilicate glass	63–68	3.2–3.5	—	0.1	0.16	4.7, 40	600	2.2
Li aluminosilicate	74–88	0.9–1.5	—	0.17	0.21	10–40	1000–1200	2.1–2.4
Mg aluminosilicate	110	2.6	—	0.14	0.13	—	1200	2.6–2.8
Ba Mg aluminosilicate	106	2.7	—	—	—	40	1250	2.6
AS-1723	70–88	5.2	—	—	—	7–9, 40	700	—
7761	63	2.6	—	—	—	—	—	—
9741	50	4.9	—	—	—	—	—	—
7052	57	5.2	—	—	—	—	—	—
Reaction-bonded Si_3N_4	110	3.6	—	0.084	0.08	—	1900	3.2
SiC	340–380	4.3	—	0.35–0.92	0.1	(1.5–4.2)	1500	3.2
SiO_2	70–80	1.0	—	—	—	7–9	1150	—
ZrO_2	195	4.2	—	0.216	0.11	(2.5)	1500	6.0
Mullite	145	5.7	—	0.185	—	(2.2)	—	2.8

Note: E = Young's modulus (GPa), a = coefficient of thermal expansion ($\times 10^6$/°C), R = fiber radius (μm), UTS = ultimate tensile strength (GPa), e^* = failure strain (%), g = surface energy (J/m^2), Kc = fracture toughness (MPa m$^{1/2}$), T = maximum-use temperature (°C), r = density (g/cm^3).

Source: Karandikar, P.G. and Chou, T.-W., in *Handbook on Continuous Fiber-Reinforced Ceramic Matrix Composites*, Lehman, R.L. et al., Eds., Purdue Research Foundation, 1995, West Lafayette, IN. With permission.

Mechanical Behavior. The primary goal of reinforcing a ceramic material with ceramic fibers is to increase the fracture toughness of otherwise brittle ceramic materials. In a monolithic ceramic, a flaw will propagate through the material under tension due to stress concentration at the crack tip. CMCs are designed with $E_f/E_m > 1$ so that the reinforcing fiber absorbs stresses which tend to open and propagate cracks. Thus, the stress at the crack tip is reduced. Fiber debonding at the crack front, sliding, crack deflection, and other mechanisms also contribute to toughness. Often this is referred to as "rising R-curve" behavior, indicating that the strength of the ceramic at a crack tip increases with increasing crack length, stopping or slowing further crack growth. The most significant feature of CMFCs is the extensive fiber pullout, which occurs during failure. The fiber pullout results from low interfacial shear strengths, τ, which is designed into the composite by modifying interfacial properties and by using fibers with a *low* Weibull modulus. Loading is characterized by an initial elastic region followed by progressive failure of the matrix. The behavior is illustrated in Figure 12.6.8. During this matrix failure region, monotonically increasing loads are supported by the fibers. Ultimately, the fibers begin to fail according to bundle theory and the remaining load support is provided by the frictional resistance to fiber pullout, a highly significant effect which absorbs large amount of energy and contributes a high strain-to-failure performance previously unknown in ceramics. This "graceful" failure allows for a less catastrophic failure than normally encountered in ceramics and is of great significance in the design of such materials as turbine blades where the consequences of catastrophic failure are severe.

Continuous and Chopped Fiber Composites. The use of a continuous fiber reinforcement can have several advantages over the use of chopped fibers. There is a larger strain to pullout due to the increased fiber length, and the continuous fibers do not have stress-concentrating "flaws" as do the exposed ends of chopped fibers. Also, the prearrangement of the fibers allows very careful control of the properties in each direction, and the properties parallel to the fiber orientation will more closely mirror those of the fibers themselves than of the matrix. However, this type of composite is much more difficult to fabricate because of the problems of fiber weaving and forming, resistance to matrix infiltration, and

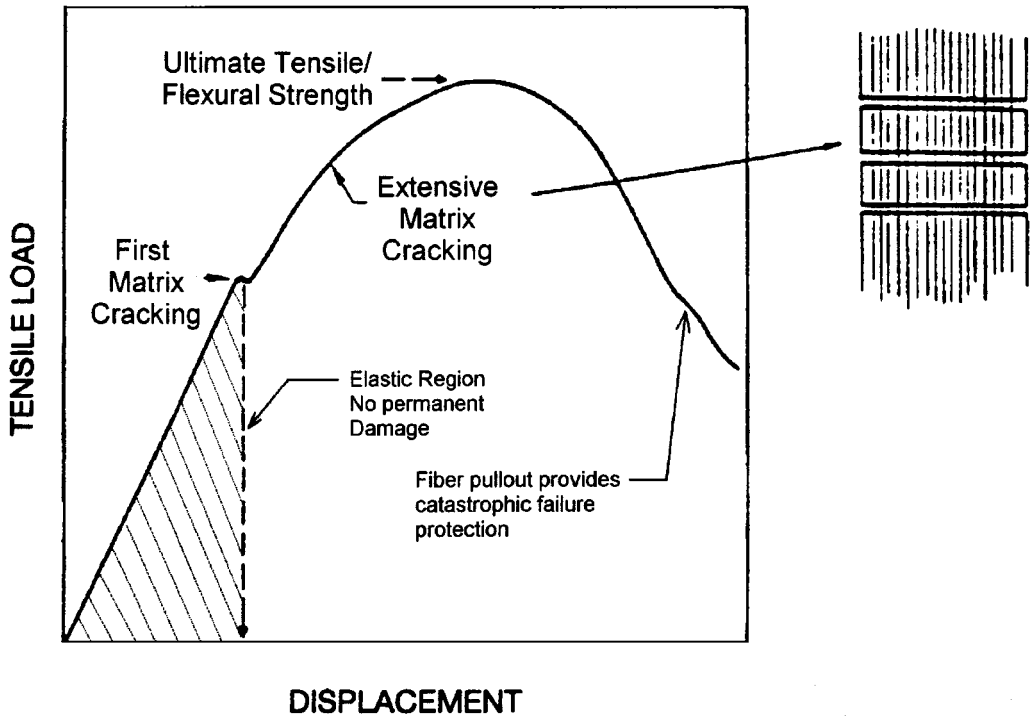


FIGURE 12.6.8 Load deflection behavior of CMFCs illustrating regions of fracture behavior.

resistance to densification, which necessitates expensive densification techniques, such as hot pressing, chemical vapor infiltration, polymer pyrolysis, directed metal oxidation, or sol–gel infiltration.

Glass Matrix CMFCs. A broad class of CMFCs has been developed and commercialized based on glassy matrix materials. The incentive for using a glassy matrix is the ease with which the matrix can be densified at low temperatures by vitreous sintering as opposed to the high-temperature solid-state sintering required for crystalline matrices. The low-temperature processing, in addition to being lower cost, preserves the high-strength properties of the fibers, which can easily be degraded at high temperatures. Unfortunately, the useful temperature of a glass-based CMC is limited to the T_g of the matrix, or slightly above, which prohibits traditional glass matrix CMCs from being used in extreme temperature applications (above approximately 1000°C). Newer glass and glass–ceramic compositions in the CaO-Al_2O_3-SiO_2 system are pushing this limit to the region of 1300°C. Two commercial glass matrix composites are Compglas™ and Blackglas™, both glass and glass/ceramic matrix materials reinforced with SiC fibers.

Polycrystalline matrix CMCs achieve higher temperature stability than the glass matrix composites, allow a wider choice of matrix materials, and generally have fewer problems with matrix reactivity. Typical crystalline matrix materials are the SiC, Si_3N_4, Al_2O_3, ZrO_2, and mixed silicates. Unfortunately, it is extremely difficult to sinter a CMC with a polycrystalline matrix to full density because of the interference of the fibers. This is particularly true with the nonoxide matrix materials. Pressureless sintering is ineffective for these materials, which necessitates the use of hot pressing or other exotic forming processes which raise material costs considerably.

Ceramic Matrix Whisker Composites (CMWCs)

Whisker reinforcement of ceramic matrices, while not as exotic, can dramatically increase toughness while preserving relatively inexpensive forming techniques. Figure 12.6.9 illustrates the "rising R-curve" behavior of alumina with the addition of SiC whiskers, while Figure 12.6.10 shows the increase in

fracture strength. The composites can generally be fully densified with hot pressing or hot isostatic pressing (HIP). Unfortunately, the toughness values achieved to date with whisker reinforcement are not as high as those for continuous-fiber reinforcement, although the lower cost of whisker composites has stimulated industrial applications. Cutting tool inserts, which must withstand high stresses at elevated temperatures (1200°C), are fabricated commercially from SiC-whisker-reinforced Al_2O_3 (SiC/Al_2O_3). These composite inserts have been a commercial product for a decade and are an excellent example of cost-effective CMCs, i.e., high-value-added material applied to a small part exposed to extreme conditions. Health concerns regarding highly durable ceramic whiskers produced from SiC or Si_3N_4 have severely limited development and commercialization of these composites in the United States and Europe.

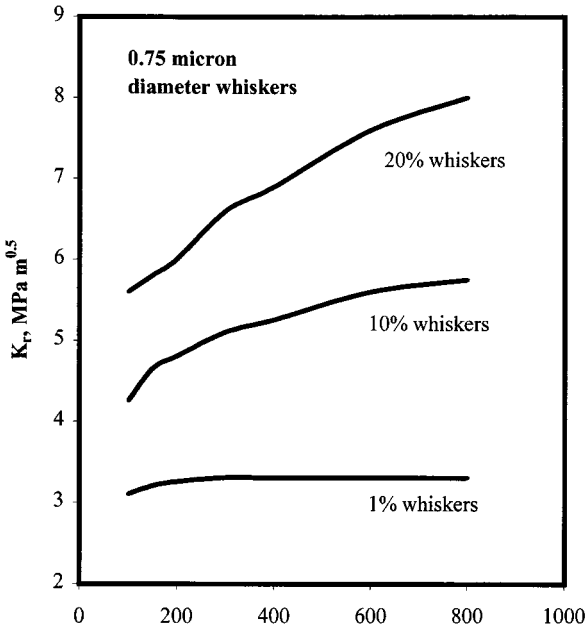


FIGURE 12.6.9 Fracture toughness and R-curve behavior of ceramic matrix whisker composites.

Carbon–Carbon Composites

Introduction

Carbon–carbon composites were originally developed for rocket nose cones in the 1960s because of their high specific strength, fracture toughness, thermal shock, and erosion resistance. Today, carbon–carbon composites are used in a broader range of applications, including brake components, fasteners, heaters, crucibles, and other assorted high-strength/high-temperature products.

High-Temperature Oxidation

Perhaps the greatest obstacle to the use of carbon–carbon composites as an engineering material, aside from high cost, is low oxidation resistance above 500°C. Although carbon–carbon maintains its mechanical properties to very high temperatures (>2000°C), it must be protected from oxidizing atmospheres at these temperatures or it will quickly degrade. Generally this is accomplished with SiC coatings with varying degrees of success. For most high-temperature oxidizing applications, carbon–carbon composite components are simply replaced when erosion reaches specified limits.

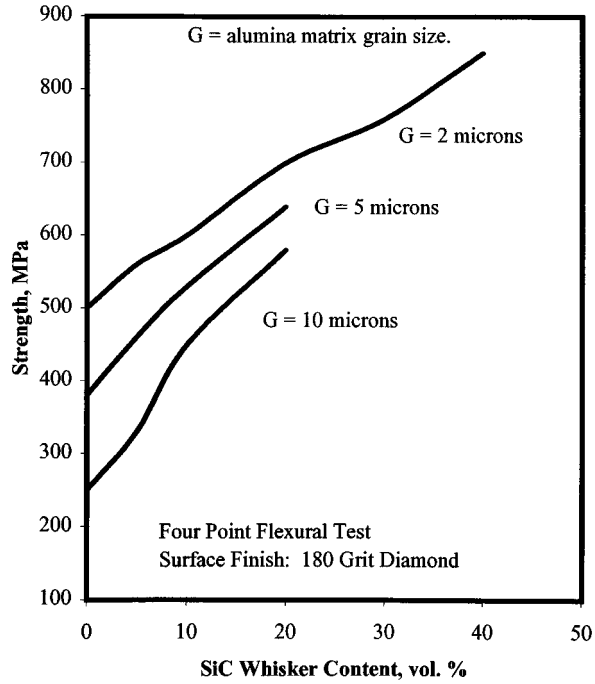


FIGURE 12.6.10 Strength behavior of SiC whisker-reinforced alumina matrix ceramic composite.

Fabrication

Most carbon–carbon composites are fabricated by polymer pyrolysis, the process of soaking carbon fiber yarns in a solution of a suitable polymer, then winding or laminating the fibers into the desired shape. Coal tar–based pitches, phenolics, and polyimides are also used as precursors.

The resulting green composite is then pressed, cured, and heated to 800 to 1200°C in an inert atmosphere, carbonizing the precursor. Some structures are heated to 2500°C to graphitize the matrix. The matrix after carbonization will contain a high porosity, typically 30%. This porosity is reduced by infiltrating the matrix with the precursor again and repeating the carbonization cycle. The densification cycle is repeated until the desired density is reached.

Carbon–carbon composites can be formed in random, unidirectional, 2D, 2.5D, and 3D weaves. Unidirectional weaves provide the highest strength along the fiber axis, but much lower strength in other directions. Unidirectional weaves are generally only used in laboratory investigations, as they have little technical value. 2D weaves (fabrics) are by far the most common due to the relative ease of manufacture and good mechanical properties in the plane of the fabric. 2.5D and 3D weaves allow the weaving of very complex structures with good properties in any desired direction, yet the weaving costs can be prohibitive and unnecessary for most industrial applications.

Mechanical Properties

Properties of some carbon–carbon composite materials are given in Table 12.6.9. The properties can vary considerably, however, depending upon factors such as the weave, tow diameter, precursor, pressing pressure, heat-treatment temperature, and number of densification cycles.

Under cyclical loading, carbon–carbon withstands 70 to 80% of its ultimate tensile strength. Carbon–carbon composites also have large strains to failure, exhibiting pseudoplastic behavior. This is unusual for a high-temperature ceramic material and is one of its prime attractions. The responsibility for this behavior lies with the low fiber–matrix interfacial bond strength and matrix microcracking. Unlike most high-temperature materials, the mechanical properties of carbon–carbon do not degrade at

TABLE 12.6.9 Properties of Selected Carbon–Carbon Composites

Property	Units	Structural Grade[a]			Fastener Grade[b]		
		Warp Direction	Fill Direction	Across Ply	Warp Direction	Fill Direction	Across Ply
Tensile strength	MPa	112	115	N/A	139	128	N/A
Tensile modulus	Gpa	89.6	91	N/A	71	717	N/A
Tensile strain to fail	%	0.126	0.126	N/A	0.271	0.242	N/A
Compressive strength	MPa	57	67	122 (min)	68	60	122 (min)
Flexural strength	MPa	86.2	N/A	N/A	106	106	N/A
Interlaminar shear	MPa	3.92	N/A	N/A	8.62	N/A	N/A
Thermal conductivity	J m^{-1} s^{-1} $°C^{-1}$	5.82	4.22	1.44	6.73	10.1	3.43
Coefficient of thermal expansion	$10^{-7}°C^{-1}$ at 371°C	–0.94	–0.94	23.4	–1.2	–1.0	20.4
Density	g/cc		1.52			1.7	
Izod impact value	J cm^{-1}		3.52			5.95	
Hardness (Shore) "D" method			69.8			79.8	
Porosity	%			<20			<8.5
Resistivity, in plane, room-temperature	Ω-cm			24×10^{-4}			21×10^{-4}
Resistivity, in plane, 1750°C	Ω-cm			24×10^{-4}			—

[a] Fiber materials C³ 16 PC — 2 × 2 twill, 1780 denier continuous PAN fibers.
[b] Fiber materials C³ 40 PS — 8 harness satin, 600 denier, staple PAN fibers.
Source: Fiber Materials Incorporated, Biddeford, Maine.

high temperatures. In fact, in inert atmospheres the strength of carbon–carbon composites increases by 40 to 50% up to a temperature of 1600°C.

Thermal conduction, electrical conduction, and thermal expansion are all much greater along the direction of the weave than perpendicular to it. Interestingly, thermal expansion is negative in 2D weaves perpendicular to the fibers. It is therefore possible to use the fiber architecture to achieve near zero thermal expansion in some directions.

Selected References and Bibliography

ASM International. 1987. *Composites, Engineered Materials Handbook,* Vol. 1, ASM, Materials Park, OH.

Brandt, A.M. and Marshall, I.H., Eds. 1986. *Brittle Matrix Composites 1,* Elsevier Applied Science, New York.

Chou, T.W. 1989. *Textile Structure Composites,* Elsevier, New York.

Lehman, R.L., El-Rahaiby, S.K., and Wachtman, J.B., Eds. 1995. *Handbook on Continuous Fiber-Reinforced Ceramic Matrix Composites,* CIAC/CINDAS, Purdue University, West Lafayette, IN.

Schwartz, M.M. 1984. *Composites Materials Handbook,* McGraw-Hill, New York.

Singh, J.P. and Bansal, N.P., Eds. 1994. *Advances in Ceramic Matrix Composites II,* American Ceramic Society, Westerville, OH.

Summerscales, J., Ed. 1987. *Non-Destructive Testing of Fiber-Reinforced Plastics Composites,* Vols. 1 and 2, Elsevier, New York.

Ziegler, G. and Huttner, W. 1991. Engineering properties of carbon–carbon and ceramic matrix composites, in *ASM Engineered Materials Handbook,* Vol. 4, ASM International, Materials Park, OH.

12.7 Ceramics and Glass

Richard L. Lehman, Daniel J. Strange, and William F. Fischer, III

Traditional Ceramics

Traditional ceramics encompass many materials, not simply pottery, dinnerware, tile, and sanitaryware, but also technical whitewares, structural clay products, refractories, glazes, and enamels. The product and processing technology of these materials has advanced substantially in recent years which has improved products, reduced costs, and expanded application. Refer to Table 12.7.1 for properties of selected ceramics in this group.

TABLE 12.7.1 (Part I) Physical Properties of Selected Ceramics

Material	Porcelain	Cordierite Refractory	Alumina, Alumina Silicate Refractories	Magnesium Silicate
Specific gravity	2.2–2.4	1.6–2.1	2.2–2.4	2.3–2.8
Coefficient of linear thermal expansion, ppm/°C, 20–700°	5.0–6.5 × 10^6	2.5–3.0 × 10^6	5.0–7.0 × 10^6	11.5 × 10^6
Safe operating temperature, °C	~400	1,250	1,300–1,700	1,200
Thermal conductivity (cal/cm^2/cm/sec/°C)	0.004–0.005	0.003–0.004	0.004–0.005	0.003–0.005
Tensile strength (psi)	1,500–2,500	1,000–3,500	700–3,000	2,500
Compressive strength (psi)	25,000–50,000	20,000–45,000	13,000–60,000	20,000–30,000
Flexural strength (psi)	3,500–6,000	1,500–7,000	1,500–6,000	7,000–9,000
Impact strength (ft-lb; $^1/_2$″ rod)	0.2–0.3	0.2–0.25	0.17–0.25	0.2–0.3
Modulus of elasticity (psi)	7–10 × 10^6	2–5 × 10^6	2–5 × 10^6	4–5 × 10^6
Thermal shock resistance	Moderate	Excellent	Excellent	Good
Dielectric strength, (V/mil; 0.25″ specimen)	40–100	40–100	40–100	80–100
Resistivity (Ω/cm^2, 22°C)	10^2–10^4	10^2–10^4	10^2–10^4	10^2–10^5
Power factor at 10^6 Hz	0.010–0.020	0.004–0.010	0.002–0.010	0.008–0.010
Dielectric constant	6.0–7.0	4.5–5.5	4.5–6.5	5.0–6.0

Whitewares

Whitewares are principally comprised of traditional ceramic bodies that are white, cream, ivory, or light gray in appearance. Most whiteware materials are formulated from clay, flint, and feldspar (triaxial compositions) although other additives may be incorporated. The engineering properties of whitewares are strongly affected by porosity, a characteristic that reduces frost resistance and cleanability but is essential for certain aesthetic effects.

Vitreous. Vitreous whiteware bodies are translucent and have no open porosity, while earthenware and wall tile, for example, have substantial porosity, and semivitreous bodies bridge the gap in both porosity and translucency. The firing temperature is largely responsible for the differences in properties of the bodies by affecting the degree of vitrification. The general categories of products within the whiteware group are given in Table 12.7.2.

Earthenware. Earthenware materials are defined as a nonvitreous clay-based ceramic ware of medium porosity (4 to 20%). They can be glazed or unglazed in their finished form. There are four primary subclasses of earthenware, natural clay, fine, semivitreous, and talc earthenware. Natural clay earthenware is derived from a single, unbeneficiated clay, whereas fine earthenware possesses beneficiated clays, as well as nonplastic materials, to comprise a triaxial body. Semivitreous earthenware is also a triaxial body, but it is fired to a higher temperature to form a more glassy phase, thereby creating a body with

TABLE 12.7.1 (Part II) Physical Properties of Selected Ceramics

Material	High-Voltage Porcelain	Alumina Porcelain	Steatite	Forsterite	Zirconia Porcelain	Lithia Porcelain	Titania/Titanate Ceramics
Specific gravity	2.3–2.5	3.1–3.9	2.5–2.7	2.7–2.9	3.5–3.8	2.34	3.5–5.5
Coefficient of linear thermal expansion, ppm/°C, 20–700°	$5.0–6.8 \times 10^6$	$5.5–8.1 \times 10^6$	$8.6–10.5 \times 10^6$	11×10^6	$3.5–5.5 \times 10^6$	1×10^6	$7.0–10.0 \times 10^6$
Safe operating temperature, °C	1,000	1,350–1,500	1,000–1,100	1,000–1,100	1,000–1,200	1,000	—
Thermal conductivity (cal/cm^2/cm/sec/°C)	0.002–0.005	0.007–0.05	0.005–0.006	0.005–0.010	0.010–0.015	—	0.008–0.01
Tensile strength (psi)	3,000–8,000	8,000–30,000	8,000–10,000	8,000–10,000	10,000–15,000	—	4,000–10,000
Compressive strength (psi)	25,000–50,000	80,000–25,000	65,000–130,000	60,000–100,000	80,000–150,000	60,000	40,000–120,000
Flexural strength (psi)	9,000–15,000	20,000–45,000	16,000–24,000	18,000–20,000	20,000–35,000	8,000	10,000–22,000
Impact strength (ft-lb; $^1/_2$″ rod)	0.2–0.3	0.5–0.7	0.3–0.4	0.3–0.4	0.4–0.5	0.3	0.3–0.5
Modulus of elasticity (psi)	$7–14 \times 10^6$	$15–52 \times 10^6$	$13–15 \times 10^6$	$13–15 \times 10^6$	$20–30 \times 10^6$	—	$0.3–0.5 \times 10^6$
Thermal shock resistance	Moderate–good	Good	Moderate	Poor	Good	Excellent	Poor
Dielectric strength, (V/mil; 0.25″ specimen)	250–400	250–400	200–350	200–300	250–350	200–300	50–300
Resistivity (Ω/cm^2, 22°C)	$10^{12}–10^{14}$	$10^{14}–10^{15}$	$10^{13}–10^{15}$	$10^{13}–10^{15}$	$10^{13}–10^{15}$	—	$10^8–10^{15}$
Power factor at 10^6 Hz	0.006–0.010	0.001–0.002	0.008–0.035	0.0003	0.0006–0.0020	0.05	0.0002–0.050
Dielectric constant	6.0–7.0	8–9	5.5–7.5	6.2	8.0–9.0	5.6	15–10,000

TABLE 12.7.2 Whiteware Materials

Class and Subclass	Percent Water Absorption	Example Product Type
Earthenware		
Natural clay	>15	Artware and tableware
Fine earthenware	10–15	Tableware, kitchenware, and artware
Semivitreous Earthenware	4–9	Tableware and artware
Talc earthenware	10–20	Ovenware and artware
Stoneware		
Natural stoneware	<5	Kitchware, artware, and drainage pipes
Fine stoneware	<5	Cookware, tableware and artware
Jasper stoneware	<1	Artware
Basalt stoneware	<1	Artware
Technical vitreous stoneware	<0.2	Chemicalware
China		
Vitreous china	0.1–0.3	Sanitaryware
Hotel china	0.1–0.3	Tableware
Cookware	1–5	Ovenware and stoveware
Technical china	<0.5	Chemicalware and ball mill jars and media
Fine china	<0.5	Tableware and artware
Porcelains		
Technical porcelains	<0.2	Chemicalware
Triaxial electrical porcelains	<0.2	Low-frequency insulators
High-strength electrical porcelains	<0.2	Low-frequency insulators
Dental porcelains	<0.1	Dental fixtures

the lowest porosity of the earthenware group, usually between 4 and 9%. The final earthenware body is talc earthenware, produced principally from raw talc, with porosity ranging up to 20%. Earthenware bodies range in color from white for the talc and triaxial bodies, to tan and brown for many artware bodies, to a rusty red for terra-cotta.

Stoneware. Stoneware bodies can be either vitreous or semivitreous. They are primarily composed of nonrefractory fireclays or a combination of triaxial materials that matches the forming, firing, and finished properties of a natural stoneware body, bodies made from a single, naturally occurring, largely unbeneficiated clay-bearing material. Fine stonewares incorporate beneficiated clays, as well as nonplastics. Jasper stonewares are composed primarily from barium-containing compounds, while basalt stonewares contain large amounts of iron oxide.

Vitreous stoneware bodies are made from blends of a variety of beneficiated materials that are fired to higher temperatures to achieve low porosity levels (0 to 5%) necessary for many applications. Stoneware bodies are usually quite durable and resistant to chipping. However, translucency is less than that of china and the colors are not as white because of the presence of iron and other impurities.

China and Porcelain. China and porcelain are nearly synonymous terms which refer to fully vitreous (no porosity) clay, flint, feldspar compositions which are typically glazed, fired to high temperatures, and exhibit strength, hardness, and chemical durability. The term *china* is used to describe exceptionally fine materials prepared from low-impurity raw materials and used in artware and dinnerware. In modern times there has been a trend toward highly vitreous and highly translucent china compositions. Porcelain is used to describe mostly technical ceramics of the triaxial composition which are used as electrical insulators, sanitaryware, and chemical ware.

Subclassifications of china exist, such as vitreous china, hotel china, cookware, technical ceramics, fine chinas, and porcelains. Body formulations are usually based on the triaxial body, clays, flint/silica and fluxing agents, most usually feldspathic materials. However, there are a large number of bodies that are composed of a large fraction of other materials. Inclusive of these are alumina, bone ash, cordierite,

other fluxes, and/or lithium compounds. Vitreous china is a category of traditional ceramics referring to the various sanitaryware plumbing fixtures and accompaniments. Hotel chinas, as the name implies, are generally used in commercial food establishments. Both bodies are glazed in a single firing operation in which both the body and glaze mature at the same time.

Technical whitewares account for a wide variety of vitreous ceramics used in the chemical, dental, refractory, mechanical, electrical, and structural areas. The compositions of most of these materials are similar to that used in the hotel chinas, with the possible substitution of alumina and zircon for some or all of the silica. These materials can be either glazed or unglazed with water absorption less than 0.5%.

Fine china bodies, including bone china, are highly vitrified and translucent materials that are usually fired in two or more separate operations. The first, higher-temperature firing matures the body and a second, lower-temperature firing matures the newly applied glaze. The separate firing conditions allow for the use of high-gloss glazes. Subsequent firings are used to apply decals and metallic decorations.

Porcelain ceramics are mostly used in technical applications. The typical body is triaxial (Table 12.7.2), although some or all of the silica can be replaced with alumina to increase the mechanical properties. Aside from triaxial porcelains, compositions in the $MgO \cdot Al_2O_3 \cdot SiO_2$ composition range are popular for electronic applications due to the absence of mobile alkali ions.

Refractories

Introduction. Refractory ceramic materials are by nature inert, high-melting-point compounds that are resistant to corrosion throughout the temperature range of use. Refractories must also withstand thermal cycling, thermal shock, mechanical fatigue, and a range of chemical attack from the elevated-temperature environments typical of most applications. Refractory materials are used in the processing of metals (75% of all refractories), glass, cement, and in the processing of nearly all ceramics.

Temperature Tolerance. A quality refractory must be stable at the intended use temperature. Refractoriness is a measure of the highest use temperature the material can withstand and is limited by the softening or melting point of the constituent oxides. Most refractories are a mixture of phases, and, as such, do not display a distinct melting point, but have a range of temperatures where the material starts to soften or melt. Most frequently, a refractory is categorized by an upper use temperature, but sometimes the refractoriness is quantified by the PCE or pyrometric cone equivalent. The PCE is a measure of the heat content that the refractory can withstand before beginning to soften, which is determined by the slumping of pyrometric cones during thermal cycle testing. This value may correspond to different temperatures under different environments or atmospheres and is thus a good indicator of maximum-use conditions. Refractory suppliers can provide PCEs for specific products. Another measure of a refractory quality is the failure under load temperature. The temperature where a refractory sags or deforms is part of all refractory specifications and is related to the amount and composition of the glassy phase within the material.

Dimensional Stability and Spalling. Dimensional stability and resistance to spalling are important performance criteria for most refractories. Spalling is the cracking or flaking of the refractory which usually results from thermal cycling, thermal gradients within the refractory, or compression effects due to differing thermal expansion of the different system materials. Spalling reduces the effectiveness and lifetime of the refractory. The dimensional stability of the refractory is also important. Since the refractory is subjected to both heating and cooling cycles, as well as thermal gradients in use, the expansion of the material is very important when choosing a refractory. Large changes in the size of a refractory set up stresses that can reduce the effectiveness of the refractory and may result in its failure.

Porosity. Refractory porosity is closely controlled in manufacturing since it leads to a reduction in the mechanical strength of the material and allows for the penetration and chemical attack of liquids or gases to the internal surface of the refractory. However, on the positive side, the presence of internal pores reduces the thermal conductivity of the material and increases fracture toughness, K_{IC}.

Fireclay Refractories. Fireclay refractories are composed of hydrated aluminosilicates with silica content of up to 75% or more with alumina and other minor contents of less than 40%, although alumina-fortified fireclay refractories are made with considerably higher alumina contents. Properties vary greatly over this wide range of compositions — generally, the higher the alumina content, the higher the performance. Fireclay refractories based on kaolin have a high refractoriness and high load resistance. Resistance to chemical attack and thermal conductivity decrease with increasing porosity, whereas spalling resistance increases. Increased alumina content raises the resistance of the material to attack in molten environments. Fireclay refractories are the most widely used refractory and find application in many industries.

Alumina Refractories. High-alumina refractories contain 80 to 99% or more aluminum oxide. As with fireclay materials, the higher the alumina content, the higher the refractoriness and the higher the load-bearing capacity. The chemical resistance of alumina refractories is greater than that of the fireclay refractories. Alumina brick is used to replace fireclay brick in more severe applications in the steel industry. Alumina bricks with phosphate bonding are used in the production of aluminum because of their refractoriness and the resistance of the phosphate bonding to chemical attack by the molten aluminum. Mullite ($3Al_2O_3 \cdot 2SiO_2$) refractories are similar to alumuna refractories in performance, but are cheaper and more resistant to thermal shock at the expense of some temperature capability.

Silica. Silica brick refractories, used in the glassmaking industry, range in composition from almost pure silica to mixtures containing lime, iron oxide, and alumina, depending on the degree of beneficiation of the raw materials. These refractories show a high degree of volume stability below 650°C (1200°F) and a lower degree of spalling than the fireclay bricks, largely because of the low thermal expansion of the silica.

Basic Refractories. Basic refractories are constituted from magnesite (MgO), calcia (CaO), and chrome (Cr_2O_3) and are used widely in metallurgical industries where basic slags predominate. They are not suitable in acid, or high-silica, environments. Refractories in this group, such as pure magnesite, dolomite (CaO-MgO), and chrome magnesite materials, have high refractoriness, volume stability, and are very resistant to chemical attack. Magnesite refractories have between 80 and 95% MgO content, a level at which the refractoriness and resistance to chemical attack is extremely high. Chromite refractories also have excellent chemical resistance in basic environments, with moderate resistance in acidic environments. The chrome content in these refractories ranges between 30 and 45%. Chrome–magnesite refractories are composed of over 60% MgO. Compositions and properties of selected basic and high-duty refractories are given in Table 12.7.3.

Glazes and Enamels

A glaze is a continuous glassy layer that is bonded to the surface of a ceramic. The glaze is typically hard, impervious to moisture, and easily cleaned. An enamel is similar in properties to a glaze, but the substrate is metallic. The surface finish of either can be altered from glossy to matte by varying the composition or the firing conditions. A glaze is usually composed of an aqueous suspension of ceramic particles that is applied to the surface of a material, dried, and fired. During firing, the materials within the glaze react and melt, forming a thin glassy layer on the surface of the ceramic material. Some materials are prefired and then the glaze is applied, or, as is becoming more common, the glaze can be fired along with the body. The maturing temperature of most glazes is on the order of 500 to 1500°C, or 930 to 2730°F. Glazes can be either clear, transparent, or opaque. Some glazes are formulated to form crystals within the glaze for a variety of optical effects, such as opalescence.

There are three main types of glazes used. The first of these are the raw glazes. Raw glazes can be further broken down into leaded, leadless, zinc-containing, slip, and porcelain glazes. Lead promotes the processing of the glaze via low viscosity and surface tension and imparts a high refractive index to the finished glaze. However, due to the health hazards associated with free lead, there has been a movement toward lead-free glazes in certain applications. Lead-free glazes require an increase in the

TABLE 12.7.3 Compositions and Properties of Selected Basic and High-Duty Refractories

Refractory	Composition	Maximum Use Temperature in O_2 (°C)	Maximum Use Temperature in O_2 (°F)	Thermal Conductivity 100°C/212°F	Thermal Conductivity 500°C/930°F	Thermal Conductivity 100°C/1830°F
Basic Refractories						
Silica	93–96% SiO_2	1700	3090	0.8–1.0	1.2–1.4	1.5–1.7
Fireclay	55–80% SiO_2 and 15–45% Al_2O_3	1300–1450	1370–2640	0.8–0.9	0.9–1.1	2.4–2.6
Magnesite	80–95% MgO Fe_2O_3 Al_2O_3	1800	3270	3.8–9.7	2.7–4.7	2.2–2.6
Chromite	30–45% Cr_2O_3, 14–19% MgO, 10–17% Fe_2O_3, 15–33% Al_2O_3	1700	3090	1.3	1.5	1.7
Chromite–magnesite	60+% Fe_2O_3 Al_2O_3	1800	3270	1.9–3.5	1.2–2.3	1.6
High-Duty Refractory						
Alumina	100% Al_2O_3	1950	3540	26	9.4	5.3
Magnesia	100% MgO	2400	4350	31	12	6
Silica	100% SiO_2	1200	2190	0.8	1.4	1.8
Mullite	72% Al_2O_3, 28% SiO_2	1850	3360	5.3	3.8	3.4

firing temperature of approximately 150°C, from 1030 to around 1190°C. Porcelain glazes mature at temperatures in the same regime as the underlying body from which they get their name. Zinc-containing glazes are similar to porcelain glazes except that they mature at lower temperatures. Slip glazes are used in artware glazing and high-tension electrical insulators. Fritted glazes are in the form of prereacted glass which has been ground to form a powder. Special glazes offer special optical properties in the finished surface. Salt glazes are formed by injecting salts into the firing kiln, with the resulting glaze having a complex pattern of crystalline and glassy phases. Crystalline glazes are often zinc based and produce crystals within the glaze, again for artistic value. Luster glazes form a metallic coating on the glaze.

Structural Clay Products

Ceramic materials are used in a wide variety of applications in the construction industry, ranging from concrete and cement for buildings and highways, to structural clay materials for use in piping and roofing. For a discussion of concrete materials refer to Section 12.5.

Structural clay products have been used for millennia. Original uses included tile and clay brick for building construction, as well as pipes in water supply and sewer applications. These materials are still used today because of their high compressive strength and imperviousness to water. Structural ceramics are coarse-grained materials with the one exception of ceramic tile. Typical raw material compositions for these bodies are 35 to 55% clay, 25 to 45% filler, usually silica, and 25 to 55% fluxing material. Colors range from white for the kaolinitic clay products, to a buff for the fireclay materials, to red for the illitic materials. Properties vary widely between the different materials. Most concern is placed on the water absorption, which relates to freeze/thaw durability and compressive strength.

Advanced Ceramics

Classes of Advanced Ceramics

Advanced ceramic materials are materials which have been engineered to possess exceptional levels of mechanical, optical, thermal, or other property. Most often the materials possess high strength, high stiffness, or are chemically inert. Typical materials contain oxides, nitrides, or carbides which may be monolithic structures or reinforced with various particulate and/or fibrous materials. Refer to Section

12.6 for a detailed discussion of composites. The reinforcement phase usually, but not necessarily, differs from the matrix material. Phases used in designing engineered ceramics are listed in Table 12.7.4 with specific property characteristics as indicated. Advanced ceramic materials are currently much more varied and less standardized than metals, and materials from different manufacturers, or even from different production lots, will have varied properties.

TABLE 12.7.4 Selected Properties of Crystalline Phases Used in Engineered Ceramics

Crystalline Phases	Formula	Melting Temperature (C)	Thermal Expansion α (ppm/°C)	Dielectric Constant κ
Magnesia	MgO	2852	14	5.5
Magnesia spinel	$MgO \cdot Al_2O_3$	2135	8	8.0
Alumina	Al_2O_3	2072	9	10.0
Mullite	$3Al_2O_3 \cdot 2SiO_2$	1920	5.5	4.5
Silica	SiO_2	1723	—	3.8
Protoenstatite	$MgO \cdot SiO_2$	1557	8.0	6.0
Forsterite	$2MgO \cdot SiO_2$	1910	12.0	6.0
Cordierite	$2MgO \cdot 2Al_2O_3 \cdot 5SiO_2$	1450	2.0	5.0
Carbon	C	3652	4.4	—
Silicon carbide	SiC	2700	4.4	—
Silicon nitride	Si_3N_4	1900	—	—
Zirconium oxide	ZrO_2	5000	—	—
Zircon	$ZrO_2 \cdot SiO_2$	2550	4.5	6.5
Wollastonite	$CaO \cdot SiO_2$	1540	5.5	6.0
Titania	TiO_2	1830	—	90
Calcium titanate	$CaO \cdot TiO_2$	1975	—	180
Strontium titanate	$SrO \cdot TiO_2$	—	—	360
Magnesium titanate	$MgO \cdot TiO_2$	—	—	14
Barium titanate	$BaO \cdot TiO_2$	—	—	2000
Magnesium ferrite	$MgO \cdot Fe_2O_3$	—	—	—
Zinc ferrite	$ZnO \cdot Fe_2O_3$	—	—	—

Structural Ceramics

Required Properties. Structural applications involve the use of ceramic materials in load-bearing situations. Material properties required for these conditions include strength over a wide temperature range, often as high as 1400°C, stiffness, creep resistance, resistance to corrosion and oxidation, and, ideally, damage tolerance or toughness. The major difficulty in the use of ceramics for structural applications is their low fracture toughness compared with metals. Conversely, ceramics excel at high-temperature behavior, have low density/weight ratios, high stiffness, and chemical inertness. At the present level of technology, there is a trade-off between high strength and high toughness. To get a high-strength material usually requires a fine-grain-sized ceramic, while a tough ceramic material often has elongated grains or reinforcement phases that are usually quite large.

Applications. Existing and potential applications include automotive, biomedical, power generation, heat exchangers, wear materials, aerospace and military applications, cutting tools, and various other technologies. Table 12.7.5 provides the mechanical properties of selected advanced ceramics. The property values listed are guidelines; exact properties are difficult to specify since the exact properties depend largely on processing. Material selection is governed by the environmental conditions of each application.

Electronic and Magnetic Ceramics

Electronic and magnetic ceramic materials have a variety of useful functions. Alumina, alumina titanate, and aluminum nitride are used as substrate materials, zirconia is used in oxygen sensors, lead zirconate

TABLE 12.7.5 Mechanical Properties of Selected Advanced Ceramic Materials

Material	Composition	Density (g/cc)	Elastic Modulus (GPa)	Fracture Strength (MPa)	Fracture Toughness K_{IC} (MPa · $m^{0.5}$)	Hardness (Vickers)
Alumina	Al_2O_3	3.9	380	Up to 400	4–9	2000
Beryllia	BeO	2.8–2.9	340	125	5	1100–1400
Chromic oxide	Cr_2O_3	4.2–4.4	—	—	4–9	—
Magnesia	MgO	3.5	300	—	3–5	500–600
Spinel	$MgAl_2O_4$	3.2	260	50–100	2–5	1200–1500
Zircon	ZrO_2 SiO_2	4.25	160	50–100	2–4	—
Zirconia	ZrO_2 stabilized with CaO	5.5	200	500	Up to 13	1200–1500
Zirconia	ZrO_2 stabilized with MgO	5.5	200	500	Up to 13	1200–1500
Zirconia	ZrO_2 stabilized with Y_2O_3	5.6	200	500–600	Up to 13	1200–1500
Zirconia (tetragonal)	ZrO_2	6.0	200	750	Up to 13	—
Zirconia (monoclinic)	ZrO_2	5.5	200	450	Up to 10	—
Reaction-bonded silicon nitride (RBSN)	Si_3N_4	1.9–2.8	150–250	300–400	Up to 12	750
Silicon nitride (hot pressed)	Si_3N_4	3.1–3.2	310	400–700	5–9	1600–2800
Silicon carbide (sintered)	SiC	3.0–3.2	400	400–500	6–9	2400–2800
Silicon carbide (hot pressed)	SiC	3.0–3.2	440	550–650	7–9	2500
Silicon carbide (RBSC)	SiC	3.0–3.15	350–400	300–400	4–8	2000
Boron carbide	B_4C	2.3–2.5	450	400–600	—	2800–3200
Boron nitride	BN	2.0–2.1	20–100[a]	—	—	Soft anisotropic
Graphite	C	1.9	3–15[a]	<50	—	Soft
Tungsten carbide	WC	15	600	450–750	Up to 20[b]	1300–1600
Titanium nitride	TiN	4.9	—	—	—	—
Titanium carbide	TiC	4.9	—	—	—	2800–3700

[a] Anisotropic.
[b] With Co additions.

titanate (PZT) and lead magnesium niobate are common actuator and transducer materials, and barium titanate and related materials are used in capacitors. Ceramic materials are used not only in military and aerospace applications, but also in consumer electronics, computers, automotive and transportation systems, and power generation systems. Electronic/magnetic ceramics are useful as a result of a variety of properties. Refer to Table 12.7.1 for dielectric property data of selected compositions.

Optical Ceramics

Zirconia and transparent alumina, or sapphire, are often used as high-temperature windows, due to their high melting points, chemical inertness, and high transparency. They are also used as watch crystals, due to their high scratch resistance. Mirrors made from silicon carbide are being evaluated because of the higher strength-to-weight ratio compared with glass materials and the relatively low coefficient of thermal expansion of SiC. Laser crystals are a group of materials which are often based on doped crystals — yttrium aluminum garnet (YAG) is representative, Infrared (IR) transparent windows and shields are used principally in military and aerospace applications to protect IR sensors from damage. Missile nose cones, or randomes, must possess high mechanical strength and be resistant to mechanical erosion and thermal shock. Typical materials include aluminum oxynitride, spinel, zinc sulfide, calcium fluoride, yttria, and sapphire. Refer to Table 12.7.6 for IR absorption ranges of important ceramic materials.

Effect of Finishing and Machining on Properties

Final finishing of many advanced ceramics is required to obtain the optimum mechanical properties or to meet design tolerances. Since sharp corners act as stress concentrators and are not associated with good design, they are usually machined to a several-millimeter radius if they occur in the as-manufactured product. Other operations include surface grinding, polishing, or lapping operations. Ceramics are very hard materials, which makes final matching machining extremely time-consuming and expensive. Every effort should be made to fabricate components to close to net shape tolerances, or to machine the piece as much as possible prior to final firing.

Traditional Glasses

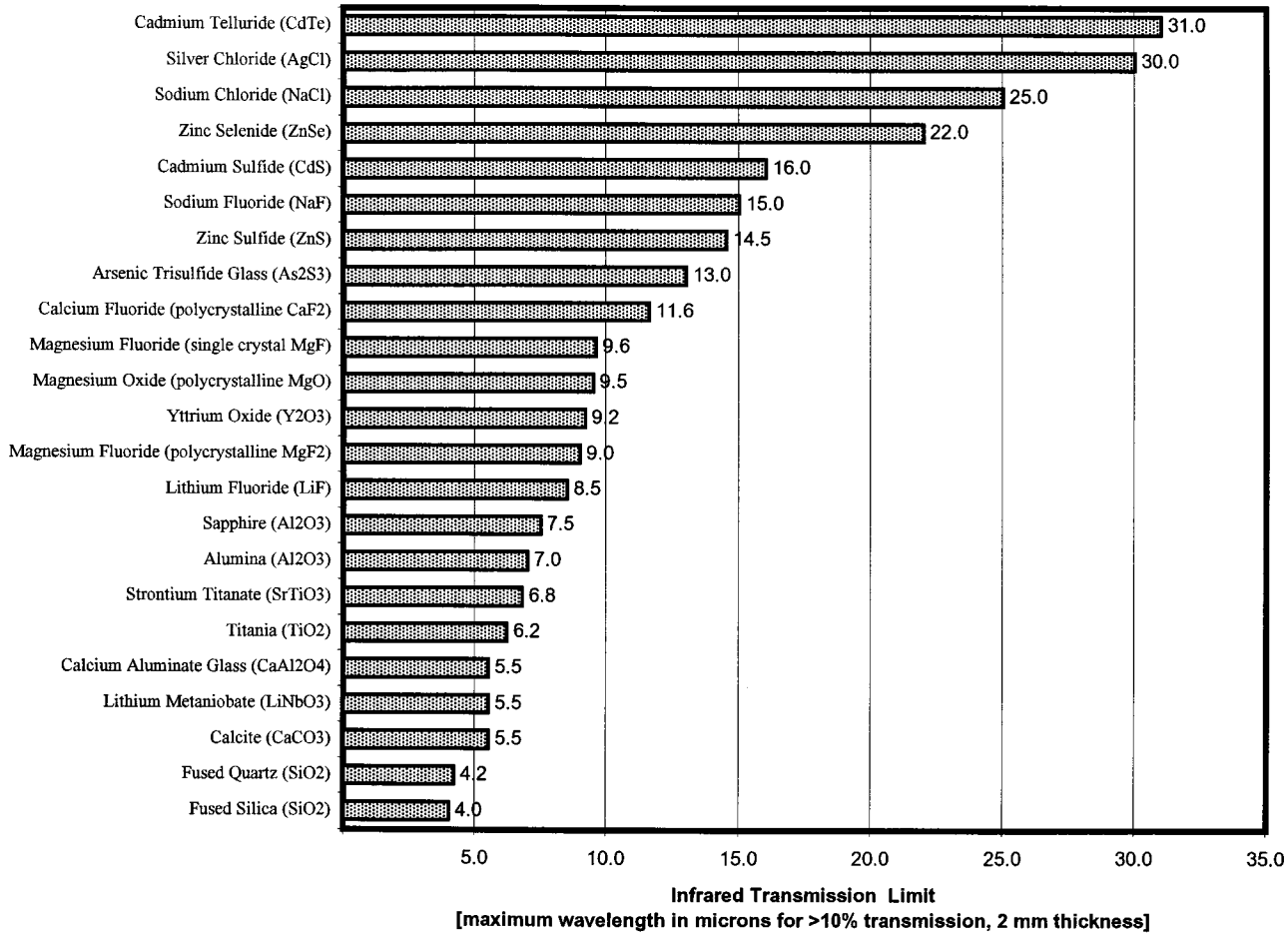
Definition and Introduction

Traditionally, glass is defined as an inorganic product of fusion which has cooled to a rigid state without crystallizing. This broad definition says nothing about the chemistry of the glass but rather refers to the traditional concept of cooling liquid melts below the melting point to form substances which are rigid elastic mechanical solids but which have not crystallized. The historical model for this process is the melting of soda–lime silicate glass by fusing the raw materials at high temperature and cooling quickly to form the well-known window and container glass still in wide use today. Although this definition is still generally valid, processes exist today in which glass is made without fusion. Chemical and physical vapor deposition, and sol–gel processes are examples. Many organic and metallic materials form glasses and are in commercial use today, thus rendering the “inorganic” part of the historical definition inappropriate.

A more applicable definition of glass is an X-ray diffraction amorphous material which exhibits a glass transition. A glass transition is the temperature point at which the amorphous solid exhibits a continuous change in thermodynamic quantities such as heat capacity and expansion coefficient. Because the structure of the glass is frozen before it reaches equilibrium, a glass is metastable with respect to its corresponding crystalline phase. Unlike the crystalline phase, the glass can have a range of physical properties depending upon the rate of cooling (thermal history) and these properties are universally isotropic.

Although glass can be made from an enormous range of polymers, oxides, metals, and inorganic and organic salts, the principle glasses of interest to a mechanical engineer will be glasses with oxide network formers such as SiO_2, B_2O_3, P_2O_5. These oxides, among others, are known as glass formers. They are capable of forming the three-dimensional network essential to the existence of the glassy state. Other

TABLE 12.7.6 IR Transmission Limits for Selected Materials



oxides, such as alkalis, alkaline earths, and various transition elements simply serve to modify the basic network to produce varying properties.

The most common glass in use today is the very old soda–lime–silicate glass composition based on the weight composition: 74% SiO_2, 16% Na_2O, and 10% CaO. In this composition the silica is the network former, the sodium breaks up the silicate network to permit melting and fabrication at reasonable temperatures, and the calcia provides ionic bonding within the network to provide chemical durability. This composition is used today, with minor variations, for early all container, window, and tableware glass articles.

Composition/Properties

Traditional glasses are silicate glasses, i.e., glass which contain SiO_2 as the principal network-forming oxide. Table 12.7.7 give compositions for a representative range of commercial glasses, and Table 12.7.8 lists properties for flat glass. All glasses in Table 12.7.7 contain at least 50% SiO_2 and many are quite close to the traditional 74/16/10 composition mentioned above. The types and amount of modifier oxides are determined by the properties required in use. A more specific list of technical glass compositions from Corning, Inc., and identified by Corning glass codes, is given in Table 12.7.9. Principal glass types are discussed in the following sections.

Pure Silica Glass. Pure silica glass (fused silica) has excellent thermal, mechanical, chemical, and optical properties. Thermal expansion is $5.5 \times 10^{-7}°C^{-1}$, use temperature is above 1000°C, and it has excellent visible and ultraviolet (UV) transmission. However, the melting point is greater than 2000°C, which makes manufacture extremely difficult and expensive. Therefore, fused silica is used primarily for such demanding applications as optical fibers, semiconductor process equipment, UV lamps, optics, and space shuttle windows. If 7 wt% titanium dioxide (TiO_2) is added, the thermal expansion coefficient is reduced to zero in normal temperature ranges. This glass (ULE, or ultra low expansion) is used in applications which require exceptional dimensional stability. Telescope mirrors are an example.

High-Silica Glass. The high melting temperatures needed to produce pure SiO_2 glass can be avoided by first melting a special sodium borosilicate glass which can be formed into the desired shape. This glass (Vycor®) phase separates and after acid leaching results in a high (>96%) silica glass with nearly the same properties as pure vitreous silica.

Soda Silica. In order to lower the melting point, a flux is added. The most common flux is sodium oxide (Na_2O), which makes a low-melting glass that is soluble in water and is called water glass. Dissolved glass is used to seal or bond materials such as low-speed grinding wheels.

Soda–Lime–Silica. This is the most common glass and accounts for more than 90% of all the glasses made today. In this glass, lime (CaO) has been added to improve the durability of simple soda–silica glass. Small amounts of other oxides such as alumina (Al_2O_3), magnesium oxide (MgO), and potassium oxide (K_2O) are also added to further enhance the desired properties. The lime is usually obtained from limestone, and the sodium monoxide is obtained from any number of sodium compounds, notably sodium bicarbonate or common baking soda.

Aluminosilicate Glass. Aluminosilicate glass is produced by adding up to 20% Al_2O_3 to soda–lime glass. Aluminosilicate glasses are resistant to thermal shock and high temperatures, and they are not as difficult to produce as silica glass. They find wide use in electronics and in high-temperature laboratory equipment.

Borosilicate Glass. In borosilicate glass much of the soda and lime of ordinary glass is replaced by boric oxide. The result is a glass that has a low thermal expansion and is thus resistant to heat and to sudden changes in temperature. The first borosilicate glass was developed by the Corning Glass Works under the trade name Pyrex. Borosilicate glasses are used in laboratory glassware, as well as in home ovenware. Because of their low thermal expansion they are suitable for applications requiring dimensional stability.

TABLE 12.7.7 Commercial Glass Compositions by Application (Wt% oxide)

Oxide	(Optical) Vitreous silica	(High silica) Vycor	Plate	Window	Con-tainer	Light bulb	Tubing	Lime table-ware	Low-expansion boro-silicate	Thermo-meter	Boro-silicate Crown	Lead table-ware	Halogen lamp	(Textile fiber) E glass	S glass	Optical flint
SiO_2	100.0	94.0	72.7	72.0	74.0	73.6	72.1	74.0	81.0	72.9	69.6	67.0	60.0	52.9	65.0	49.8
Al_2O_3			0.5	0.6	1.0	1.0	1.6	0.5	2.0	6.2		0.4	14.3	14.5	25.0	0.1
B_2O_3		5.0							12.0	10.4	9.9			9.2		
SO_3			0.5	0.7	tr	5.2							0.3			
CaO			13.0	10.0	5.4	3.6	5.6	7.5		0.4			6.5	17.4		
MgO				2.5	3.7		3.4			0.2				4.4	10.0	
BaO					tr						2.5		18.3			13.4
PbO												17.0				18.7
Na_2O		1.0	13.2	14.2	15.3	16.0	16.3	18.0	4.5	9.8	8.4	6.0	0.01			1.2
K_2O					0.6	0.6	1.0			0.1	8.4	9.6	tr	1.0		8.2
ZnO																8.0
As_2O_3			tr	tr	tr	tr		tr		tr	0.3	tr				0.4

Source: Varshneya, A.K., *Fundamentals of Inorganic Glasses,* Academic Press, New York, 1994. With permission.

TABLE 12.7.8 Properties of Clear and Tinted Flat Glass (Applicable Federal Specification Standard DD-G-451c)

Property	Value
Specific gravity	2.5
Specific heat	0.21
Hardness (Moh's)	5–6
Softening point, °C	729
Refractive index, sodium D-line	1.52
Modulus of elasticity, GPa	70
Tensile strength, MPa	45
Poisson's ratio	0.23
Coefficient of linear expansion, ppm/°C	8.8
Dielectric constant, 1 Mhz	7.1

Lead Silicates. Lead silicate glass has a higher index of refraction and also a higher dispersion than soda–lime–silicate glass and thus finds use in optical applications. The high gloss resulting from high Fresnel reflection makes lead glasses of 24 to 35% PbO popular for consumer products in the form of artware and lead crystal glass. Lead glasses of different compositions are used widely in electronic applications since low-melting sealing glasses can be formed with little or no alkali, a constituent which promotes high electrical loss.

Strength of Glass

Theoretical Strength. Glass below the glass transition temperature (T_g) is a brittle solid, with failure originating at flaws (scratches, defects, minute compositional differences) which act as stress concentrators. Without flaws of any kind the strength of glass approaches theoretical levels of about 17 GPa (2.5×10^6 psi). Unfortunately, unless glass is processed under the utmost pristine conditions and then immediately coated to prevent surface abrasion by dust or other environmental agents, the glass will contain flaws which decrease the strength by several orders of magnitude from theoretical levels. Synthetic silica glass optical fibers, prepared under meticulous conditions and tested at 77 K, are among the few glass materials which exhibit nearly theoretical strength.

Nominal and Design Strength. The exact failure stress for a specific piece of glass will depend upon the configuration and size of the defect at the crack origin. Since these defects vary in size over a wide range, the standard deviation in the strength of glass will also be large. The glass design engineer must allow large safety factors, often 20 to 50%, to account for the statistical variation. For most types of glass a nominal strength of 70 MPa and a design stress of 7 MPa are typical. Table 12.7.10 summarizes strength and variability in the strength of glasses.

Strengthening and Tempering. Glass can be substantially strengthened, or tempered, either by rapid cooling or by ion-exchange of the surface to develop compressive stresses. Fast, uniform cooling of glass plates heated to the softening point will introduce surface compressive stresses on the glass (with corresponding tensile stress in the center) which become permanently frozen into place upon cooling to room temperature. These compressive stresses, typically 70 to 200 MPa, will resist externally applied tensile stresses and help to prevent crack propagation. Tempering will increase the strength of glass by as much as three to six times. Tempered glass will shatter violently when it fails as a result of the sudden release of stored elastic energy, although the broken pieces will not be sharp. Generally, only simple shapes can safely be thermally tempered.

Ion-exchange tempering also improves the strength of the glass by introducing surface compression. In the ion-exchange process large ions are "stuffed" into the interstices in the glass structure previously occupied by smaller ions. This is done by immersing the glass in a molten bath of the alkali salt (typically KNO_3 for a sodium-containing glass) at an elevated temperature. The larger ions will introduce a compressive strain as they force their way into the glass network. The strains obtainable are much higher

TABLE 12.7.9 Compositions of Silicate Glasses — Corning Glass Types by Number (approximate wt%)

Glass No.[a]	SiO_2, Silica	Na_2O, Soda	K_2O, Potash	PbO, Lead	CaO, Lime	B_2O_3, Boric Oxide	Al_2O_3, Aluminum Oxide	Other
0010	63	7	7	22	—	—	1	—
0080	73	17	—	—	5	—	1	4% MgO
0120	56	4	9	29	—	—	2	—
1720	62	1	—	—	8	5	17	7% MgO
1723	57	—	—	—	10	5	15	6% BaO, 7 MgO
1990	41	5	12	40	—	—	—	2% Li_2O
2405	70	5	—	—	—	12	1	11% ZnO + CdS, Se
2475	67	10	7	—	—	—	—	12% ZnO, 2% CdO + F^-
3320	76	4	2	—	—	14	3	1% U_3O_8
6720	60	9	2	—	5	1	10	9% ZnO + 4% F^-
6750	61	15	—	—	—	1	11	9% BaO + 3% F^-
6810	56	7	1	3	4	1	10	12% ZnO + 6% F^-
7040	67	4	3	—	—	23	3	—
7050	67	7	—	—	—	24	2	—
7052	65	2	3	—	—	18	7	3% BaO + F^-, 1% Li_2O
7056	70	1	8	—	—	17	3	1% Li_2O
7070	71	0.5	1	—	—	26	1	0.5% Li_2O
7250	78	5	—	—	—	15	2	—
7570	3	—	—	75	—	11	11	—
7720	73	4	—	6	—	15	2	—
7740	81	4	—	—	—	13	2	—
7760	79	2	2	—	—	15	2	—
7900	96	—	—	—	—	3	0.3	—
7913	96.5	—	—	—	—	3	0.5	—
7940	99.9	—	—	—	—	—	—	0.1% H_2O
8160	56	3	10	23	1	—	2	5% BaO + F^-
8161	40	—	5	51	—	—	—	2% BaO + 2% Rb_2O
8363	5	—	—	82	—	10	3	—
8871	42	2	6	49	—	—	—	1% Li_2O
9010	67	7	7	2	—	—	4	12% BaO + Co_3O_4 + NiO + F^-, 1% Li_2O
9606	56	—	—	—	—	—	20	9% TiO_2, 15% MgO
9700	80	5	—	—	—	13	2	—
9741	66	2	—	—	—	24	6	1% F^-, 1% Li_2O

[a] See Table C.12 in Appendix.

Source: Hutchins, J.R., III and Harrington, R.V., *Kirk-Othmer Encyclopedia of Chemical Technology*, Vol. 10, p. 542. Copyright © 1966 by John Wiley & Sons, Inc. Reprinted by permission.

than those from thermal tempering, generating stresses as high as 700 MPa, although the surface compressive layer is quite shallow and subject to penetration. Ion-exchanged glasses can be more than 10 to 20 times as strong as normal glass and are used, for example, in aircraft windshields to resist bird impacts.

Behavior at Elevated Temperatures

Glasses do not have a clearly defined melting temperature. Instead, there is a temperature range where the viscosity of the glass changes smoothly from solid (greater than 10^{14} Pa · sec) to liquid (less than 10 Pa · sec). It is useful to define some points on the viscosity–temperature curve to create a clearer

TABLE 12.7.10 Ideal and Practical Strengths of Glass, Glass Fibers, and Glass–Ceramics

Type of Glass	Tensile Strength, psi	Strength/Weight Ratio, psi/lb per cu. in.
Untreated Glass		
Theoretical strength	1,000,000–4,000,000	45,000,000
Fibers, protected in vacuum	Up to 2,000,000	—
Fibers, in air, commercially available	250,000 average	6,000,000 average
Fibers, effective strength in plastic	150,000 average	1,800,000–4,500,000
Bulk glass, protected in vacuum	Up to 500,000	—
Blown ware, unabraded	Up to 100,000	90,000
Pressed ware, unabraded	8,000 average	55,000
Bulk glass, abraded	4,000–8,000 average	30,000
Bulk glass, abraded, 1000-hr stress	2,000 minimum	—
Bulk glass, abraded, design strength	500–1,500	—
Tempered Glass		
Bulk glass, abraded	15,000–35,000	—
Normal design strength	1,500–6,000	—
Chemically Strengthened Glass		
Bulk glass, abraded	100,000 and more	—
Glass–Ceramics		
Bulk material, unabraded	20,000–35,000	27,000
Bulk material, abraded	10,000–24,000	—
Design strength	3,000–6,000	—
Chemically Strengthened Glass–Ceramics		
Bulk material, abraded	200,000 and more	—

Compiled from several sources.

picture of the viscosity behavior of a glass. The strain point is defined as the temperature at which a glass will release 95% of its stresses within a period of 6 hr. This occurs at a viscosity of approximately $10^{13.5}$ Pa · sec. The annealing point occurs at 10^{12} Pa · sec a viscosity at which 95% of stresses will be released in 15 min. At $10^{6.6}$ Pa · sec, the softening point, the glass will deform under its own weight. The working point is defined as 10^3 Pa · sec, and is a typical minimum viscosity for machine working during forming. Refer to Table C.12 in the Appendix, Properties of Silicate Glasses. Many properties of glasses are presented in this comprehensive table, including maximum application temperatures (upper working temperature) and viscosity data for most common glasses. Generally, the extreme upper working temperature corresponds to the strain point of the glass. Normal service conditions are typically 50 to 60% of the strain point on the Kelvin temperature scale. Under normal service conditions glass is a brittle, mechanical solid, it possesses no mechanical characteristics of a liquid, and it will not flow. Specific heat and thermal conductivity data are presented in Table 12.7.11. Below the glass transition, glass is not a supercooled liquid, it is a glass. Old windows are not thicker at the bottom because of viscous flow.

Chemical Durability

Although glass is often considered an extraordinarily inert material, even the most durable glasses undergo some environmental degradation and many glasses are rapidly attacked by strong acid or basic solutions. Corrosive aqueous environments can cause the ions in the glass to be extracted by a leaching process. Chemical durability behavior is difficult to generalize and trial-and-error testing is usually required for each application. The most severe attack is experienced at extreme high and low pH values; neutral pH solutions rarely attack glass at significant rates. Glasses high in SiO_2, Al_2O_3, and CaO are generally most durable. Glasses high in alkali, such as Na_2O or K_2O, are less durable. B_2O_3 and PbO are intermediate.

A notable form of chemical degradation which occurs under simple, apparently benign, conditions is weathering. Standard soda–lime–silicate window glass can be permanently damaged in a short time (~30 days) if held under high humidity with a means to retain adsorbed moisture on the surface. Such a means

TABLE 12.7.11 Thermal Properties of Several Glass Types

Material	Specific Heat			Thermal Conductivity, cal/cm sec°C × 10^4[a]			
	25°C	500°C	1000°C	–100°C	0°C	100°C	400°C
Fused silica	0.173	0.268	0.292	25.0	31.5	35.4	—
7900	0.18	0.24	0.29	24	30	34	—
7740	0.17	0.28	—	21	26	30	—
1723	0.18	0.26	—	—	29	33	—
0311 (chemically strengthened)	0.21	0.28	—	—	27	29	35
Soda–lime window glass	0.190	0.300	0.333	19	24	27	—
Heavy flint, 80% PbO, 20% SiO_2	—	—	—	10	12	14	—
Foamglass insulation	0.20	—	—	(0.97)	1.3	1.73	(2.81)
Fibrous glass	—	—	—	—	(0.8)	—	—
9606 glass–ceramic	0.185	0.267	0.311	—	90	86	75
9608 low-expansion glass–ceramic	0.195	0.286	—	—	48	51	55

Notes: Parentheses indicate extrapolated values. Specific heat increases with temperature and approaches zero at 0° K. There are no critical temperatures or phase changes. Thermal conductivity increases with temperature and is very high for glass ceramics.

[a] To convert to SI units, 1 cal/cm sec °C = 418.6 w/mk.

Source: Hutchins, J.R. III and Harrington, R.V., *Kirk-Othmer Encyclopedia of Chemical Technology*, Vol. 10, p. 598. Copyright © 1966 by John Wiley & Sons, Inc. Reprinted by permission.

can be a dirt film or perhaps a sheet of newspaper. The moisture ion-exchanges with sodium ions in the glass, which form a high-pH surface film which accelerates the attack on the glass. Pitting and adherent surface deposits can quickly form, making the glass unusable.

For container applications requiring extremely high purity, it is common to use a durable glass such as a borosilicate or to increase the chemical resistance of the glass using surface treatments of SO_2, SnO_2, or TiO_2.

Optical Properties

Traditional glasses are principally used because they are transparent in the visible region of the spectrum; windows, containers, and precision optics are examples. Although pure silica is highly transmissive overt the entire visible spectrum, impurities impart coloration which detract from the performance of some glasses but which can also be used to produce beautiful and useful colored glasses. Table 12.7.12 lists impurity ions and their resulting colors. Iron is the most common impurity, imparting a blue-green-yellow tint depending on the oxidation state. It is possible to neutralize but not bleach impurity colors by "decolorization," the addition of complementary coloring oxides to produce an overall neutral gray absorption. Traditional glasses do not transmit well in the UV or IR range, with the exception of certain specialty glasses. Pure silica is the best example, enabling high transmission levels (>10%, 2 mm thickness) from 160 nm to 4 μm.

The refractive index of common silicate glasses are in the range of 1.5 to 1.7 and specific values are given in Table C.12 in the Appendix, *Properties of Silicate Glasses*.

Specialty Glasses

Non-Silica-Oxide Glasses

Glasses made from B_2O_3 and P_2O_3 glass formers rather than SiO_2 possess some special thermal, optical, and chemical properties which make them of interest in certain narrow engineering fields. Borate glasses, often in combination with PbO, are useful solder and sealing glasses for electronic applications; phosphate glasses have special optical properties and are also used as water-soluble chemicals in industry. Tellurite glasses have high refractive indexes and, hence, are used in some demanding refractive optic applications. Glasses based on Bi_2O_3, Sb_2O_3, TeO_2, or V_2O_5 have very low melting points, suggesting their use as low-temperature electrical seals. Boro-aluminates have very high electrical resistivities. Alkaline earth aluminates have excellent IR transmitting properties and make excellent high-temperature

TABLE 12.7.12 Coloring Additives to Glass

Color	Additive
Red	Colloidal Au or Cu, Cd-Se (S, Te)
Pink	$MnO_2 \cdot CeO_2$, Se^{2-}
Orange	CdS (Se)
Amber	FeS_x, $Fe_2O_3 \cdot TiO_2$
Yellow	UO_2, $CeO_2 \cdot TiO_2$, CdS
Green	Cr_2O_3, Fe_2O_3, CuO, U_2O_3
Blue	CoO, FeO, CuO
Violet	NiO, Mn_2O_3
Gray	$Co_3O_4 \cdot NiO$
Black	$Mn_2O_3 \cdot Cr_2O_3$, PbS, FeS, $CoSe_x$
UV absorption	CeO_2, TiO_2, Fe_2O_3, V_2O_5, CrO_3
IR absorption	FeO, CuO
Decolorization, i.e., mask Fe_2O_3 color	MnO, Se^{2-}, NiO, Co_3O_4
Opacification, i.e., white opals	CaF_2, NaF, ZnS, $Ca_2(PO_4)_3$
Solarization	Cerium, arsenic

lamp seals. Despite their unique properties, non-silica-oxide glasses are costly and make up only a very small percentage of the glass produced annually.

Chalcogenide Glasses

A class of excellent IR-transmitting glasses, the chalcogenides, are obtained by combining group VI elements with group V and IV elements. Glasses in this group also exhibit photoconductivity and semiconductivity. Applications for these glasses consist of IR-transmitting optical waveguides (to 20 μm), high-performance IR optical applications, and specialty applications which utilize their photoconductivity properties. Most notable of these applications is the photosensitive coating applied to photocopy drums. Purity issues have limited applications in optical fibers as of this writing.

Heavy Metal Fluoride Glasses

Heavy metal fluoride glasses (HMFG) are an important new (1975) composition group because of their extremely low theoretical optical attenuation (10^{-3} dB/km at 3.5 μm), which makes them candidates for repeaterless transoceanic communication links. Unfortunately, this magnitude of transmission has not been obtained in practice because of problems with high oxygen impurities and crystallization. Furthermore, HMFGs are readily attacked by water, and current development efforts are aimed at improving chemical durability. At present, HMFGs are limited to short transmission distance IR optical applications.

Amorphous Metals

Certain metal compositions can be fabricated as glasses by subjecting streams of the molten metal to extremely rapid quenching rates (10^5 to 10^8°C/sec). The resulting glasses possess intriguing properties. Strengths approach theoretical limits, and electrical resistivities are greater than their crystalline counterparts yet decrease with temperature. Most importantly, they have extremely low B–H hysteresis curves. For this reason they are used commercially as power transformer core laminations.

Amorphous Semiconductors

Many elements and compounds which exhibit semiconducting properties in the crystalline state are also semiconductors in the amorphous state. Si, Ge, P, As, $CdGe_xAs_2$ (x = 0 to 1.2), $Si_{1-x}H_x$ (x = 0.1 to 0.2) are important examples. These materials are use in fabrication of inexpensive vapor deposition fabrication of photovoltaic cells.

Glass Ceramics

A useful group of materials is made by batching, melting, and forming a product as a glass followed by heat treatment to nucleate and grow crystalline phases from the glass to produce a ceramic with up to 99% crystalline phase content. The microstructure contains crystals of about 1 μm size, a glassy matrix, and no porosity. The processing route is a principal advantage since high-speed glass-forming methods can be used, no porosity exists, the formed shape can be inspected as a transparent glass, and rejects at the forming stage can be recycled.

Glass ceramics are typically stronger than most ceramics as a result of zero porosity, and they are tougher than glass because of the deflection of crack fronts around the crystals. High-temperature properties are generally not good due to the glassy phase and the nature of the process, thus limiting most glass ceramics to low- and intermediate-temperature applications (<1000°C). A wide range of products has been made from glass ceramics, from home cookware to industrial bearings and aerospace radomes.

Selected References and Bibliography

1. Doremus, R.H. 1994. *Glass Science*. 2nd ed. John Wiley & Sons, New York.
2. Fanderlik, I. 1983. *Optical Properties of Glass*. Elsevier, New York.
3. Henkes, V.E., Onoda, G.Y., and Carty, W.M. 1996. *Science of Whitewares*. American Ceramic Society, Westerville, OH.
4. Jones, J.T. and Berard, M.F. 1972. *Ceramics — Industrial Processing and Testing*. Iowa State University Press, Ames.
5. Kingery, W.D., Bowen, H.K., and Uhlmann, D.R. 1976. *Introduction to Ceramics*. 2nd ed. John Wiley & Sons, New York.
6. Reed, J. 1988. *Introduction to the Principles of Ceramic Processing*. John Wiley & Sons, New York.
7. Richerson, D.W. 1982. *Modern Ceramic Engineering.* Marcel Dekker, New York.
8. Schneider, S.J., Ed. 1991. *Engineered Materials Handbook, Vol. 4: Ceramics and Glasses*. ASM International, Materials Park, OH.
9. Shand, E.B. 1982. *Glass Engineering Handbook*, 2nd ed. McGraw-Hill, New York.
10. Tooley, F.V. 1974. The *Handbook of Glass Manufacture*, Vol. I and II. Ashlee Publishing Company, New York.
11. Tooley, F.V. 1988. *Handbook on Glass Manufacturing*. Ashlee Publishing Company, New York.
12. Varshneya, A.K. 1994. *Fundamentals of Inorganic Glasses*. Academic Press, New York.

Lee, J.; et. al. “Modern Manufacturing”
Mechanical Engineering Handbook
Ed. Frank Kreith
Boca Raton: CRC Press LLC, 1999

Modern Manufacturing

Jay Lee
National Science Foundation

Robert E. Schafrik
National Research Council

Steven Y. Liang
Georgia Institute of Technology

Trevor D. Howes
University of Connecticut

John Webster
University of Connecticut

Ioan Marinescu
Kansas State University

K. P. Rajurkar
University of Nebraska-Lincoln

W. M. Wang
University of Nebraska-Lincoln

Talyan Altan
Ohio State University

Weiping Wang
General Electric R & D Center

Alan Ridilla
General Electric R & D Center

Matthew Buczek
General Electric R&D Center

S. H. Cho
Institute for Science and Technology, Republic of Korea

Ira Pence
Georgia Institute of Technology

Toskiaki Yamaguchi
NSK Ltd.

Yashitsuga Taketomi
NSK Ltd.

(continued on next page)

Carl J. Kempf
NSK Ltd.

John Fildes
Northwestern University

Yoram Koren
University of Michigan

M. Tomizuka
University of California-Berkeley

Kam Lau
Automated Precision, Inc.

Tai-Ran Hsu
San Jose State University

David C. Anderson
Purdue University

Tien-Chien Chang
Purdue University

Hank Grant
University of Oklahoma

Tien-I. Liu
California State University at Sacramento

J. M. A. Tanchoco
Purdue University

Andrew C. Lee
Purdue University

Su-Hsia Yang
Purdue University

Takeo Nakagawa
University of Tokyo

H. E. Cook
University of Illinois at Urbana-Champaign

James J. Solberg
Purdue University

Chris Wang
IBM

13.1 Introduction

Jay Lee and Robert Shafrik

Manufacturing is the means by which the technical and industrial capability of a nation is harnessed to transform innovative designs into well-made products that meet customer needs. This activity occurs through the action of an integrated network that links many different participants with the goals of developing, making, and selling useful things.

Manufacturing is the conversion of raw materials into desired end products. The word derives from two Latin roots meaning *hand* and *make*. Manufacturing, in the broad sense, begins during the design phase when judgments are made concerning part geometry, tolerances, material choices, and so on. Manufacturing operations start with manufacturing planning activities and with the acquisition of required resources, such as process equipment and raw materials. The manufacturing function extends throughout a number of activities of design and production to the distribution of the end product and, as necessary, life cycle support. Modern manufacturing operations can be viewed as having six principal components: materials being processed, process equipment (machines), manufacturing methods, equipment calibration and maintenance, skilled workers and technicians, and enabling resources.

There are three distinct categories of manufacturing:

- *Discrete item manufacturing*, which encompasses the many different processes that bestow physical shape and structure to materials as they are fashioned into products. These processes can be grouped into families, known as unit manufacturing processes, which are used throughout manufacturing.
- *Continuous materials processing*, which is characterized by a continuous production of materials for use in other manufacturing processes or products. Typical processes include base metals production, chemical processing, and web handing. Continuous materials processing will not be further discussed in this chapter.
- *Micro- and nano-fabrication*, which refers to the creation of small physical structures with a characteristic scale size of microns (millionths of a meter) or less. This category of manufacturing is essential to the semiconductor and mechatronics industry. It is emerging as very important for the next-generation manufacturing processes.

Manufacturing is a significant component of the U.S. economy. In 1995, 19% of the U.S. gross domestic product resulted from production of durable and nondurable goods; approximately 65% of total U.S. exports were manufactured goods; the manufacturing sector accounted for 95% of industrial research and development spending; and manufacturing industries employed a work force of over 19 million people in 360,000 companies. In the modern economy, success as a global manufacturer requires the development and application of manufacturing processes capable of economically producing high-quality products in an environmentally acceptable manner.

Modern Manufacturing

Manufacturing technologies address the capabilities to design and to create products, and to manage that overall process. Product quality and reliability, responsiveness to customer demands, increased labor productivity, and efficient use of capital were the primary areas that leading manufacturing companies throughout the world emphasized during the past decade to respond to the challenge of global competitiveness. As a consequence of these trends, leading manufacturing organizations are flexible in management and labor practices, develop and produce virtually defect-free products quickly (supported with global customer service) in response to opportunities, and employ a smaller work force possessing multidisciplinary skills. These companies have an optimal balance of automated and manual operations.

To meet these challenges, the manufacturing practices must be continually evaluated and strategically employed. In addition, manufacturing firms must cope with design processes (e.g., using customers' requirements and expectations to develop engineering specifications, and then designing components),

production processes (e.g., moving materials, converting materials properties or shapes, assembling products or components, verifying processes results), and business practices (e.g., turning a customer order into a list of required parts, cost accounting, and documentation of procedures). Information technology will play an indispensable role in supporting and enabling the complex practices of manufacturing by providing the mechanisms to facilitate and manage the complexity of manufacturing processes and achieving the integration of manufacturing activities within and among manufacturing enterprises. A skilled, educated work force is also a critical component of a state-of-the-art manufacturing capability. Training and education are essential, not just for new graduates, but for the existing work force.

Manufacturing is evolving from an art or a trade into a science. The authors believe that we must understand manufacturing as a technical discipline. Such knowledge is needed to most effectively apply capabilities, quickly incorporate new developments, and identify the best available solutions to solve problems. The structure of the science of manufacturing is very similar across product lines since the same fundamental functions are performed and the same basic managerial controls are exercised.

13.2 Unit Manufacturing and Assembly Processes

Robert E. Schafrik

There are a bewildering number of manufacturing processes able to impart physical shape and structure to a workpiece. However, if these processes are broken down into their basic elements and then examined for commonality, only a few fundamental processes remain. These are the building blocks, or unit processes, from which even the most complicated manufacturing system is constructed. This section describes these unit processes in sufficient detail that a technically trained person, such as a design engineer serving as a member of an integrated product and process design team comprised of members from other specialties, could become generally knowledgeable regarding the essential aspects of manufacturing processes. Also, the information presented in this section will aid such an individual in pursuing further information from more specialized manufacturing handbooks, publications, and equipment/tool catalogs.

Considering the effect that a manufacturing process has on workpiece configuration and structure, the following five general types of unit manufacturing process can be identified (Altan et al., 1983; NRC, 1995):

Material removal processes — Geometry is generated by changing the mass of the incoming material in a controlled and well-defined manner, e.g., milling, turning, electrodischarge machining, and polishing.

Deformation processes — The shape of a solid workpiece is altered by plastic deformation without changing its mass or composition, e.g., rolling, forging, and stamping.

Primary shaping processes — A well-defined geometry is established by bulk forming material that initially had no shape, e.g., casting, injection molding, die casting, and consolidation of powders.

Structure-change processes — The microstructure, properties, or appearance of the workpiece are altered without changing the original shape of the workpiece, e.g., heat treatment and surface hardening.

Joining and assembly processes — Smaller objects are put together to achieve a desired geometry, structure, and/or property. There are two general types: (1) consolidation processes which use mechanical, chemical, or thermal energy to bond the objects (e.g., welding and diffusion bonding) and (2) strictly mechanical joining (e.g., riveting, shrink fitting, and conventional assembly).

Unit Process Selection

Each component being manufactured has a well-defined geometry and a set of requirements that it must meet. These typically include

- Shape and size
- Bill-of-material
- Accuracy and tolerances
- Appearance and surface finish
- Physical (including mechanical) properties
- Production quantity
- Cost of manufacture

In order to satisfy these criteria, more than one solution is usually possible and trade-off analyses should be conducted to compare the different approaches that could be used to produce a particular part.

Control and Automation of Unit Processes

Every unit process must be controlled or directed in some way. The need for improved accuracy, speed, and manufacturing productivity has spurred the incorporation of automation into unit processes regarding both the translation of part design details into machine instructions, and the operation of the unit process

itself and as a subsystem of the overall production environment. The section of this chapter on computer-aided design/computer-aided manufacturing (CAD/CAM) discusses the technology involved in creating and storing CAD files and their use in CAM. The expectations of precision are continuing to change, as indicated in Figure 13.2.1. This drive for ever-tighter tolerances is helping spur interest in continual improvements in design and manufacturing processses.

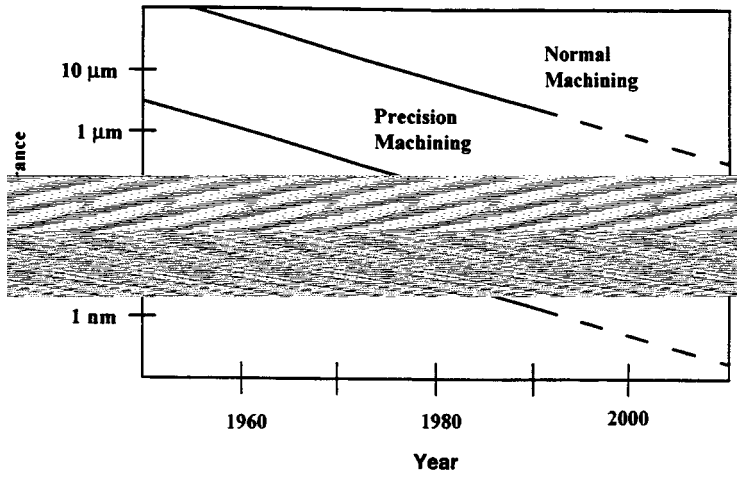


FIGURE 13.2.1 Precision machining domains. (From NRC, *Unit Manufacturing Processes*, National Academy Press, Washington, D.C., 1995, 169. With permission.)

Modern machine tool controls are emphasizing two areas: adaptive control and communication. For *adaptive control* the controller must adapt its control gains so that the overall system remains at or near the optimal condition in spite of varying process dynamics. Expanded *communication* links the data collected by a unit process controller to other segments of the manufacturing operation. Data regarding production time and quantity of parts produced can be stored in an accessible database for use by inventory control and quality monitoring. This same database can then be used by production schedulers to avoid problems and costs associated with redundant databases.

At the factory level, machining operations employing two or more numerically controlled (NC) machine tools may use a separate mainframe computer that controls several machine tools or an entire shop. The system is often referred to as *distributed numerical control* (DNC).

Today many factories are implementing *flexible manufacturing systems* (FMS), an evolution of DNC. An FMS consists of several NC unit processes (not necessarily only machine tools) which are interconnected by an automated materials handling system and which employ industrial robots for a variety of tasks requiring flexibility, such as loading/unloading the unit process queues. A single computer serves as master controller for the system, and each process may utilize a computer to direct the lower-order tasks. Advantages of FMS include:

- A wide range of parts can be produced with a high degree of automation
- Overall production lead times are shortened and inventory levels reduced
- Productivity of production employees is increased
- Production cost is reduced
- The system can easily adapt to changes in products and production levels

Unit Processes

In the following discussion, a number of unit processes are discussed, organized by the effect that they have on workpiece configuration and structure. Many of the examples deal with processing of metals

since that is the most likely material which users of this handbook will encounter. However, other materials are readily processed with the unit processes described in this chapter, albeit with suitable modifications or variations.

Mechanical assembly and material handling are also discussed in this section. On average, mechanical assembly accounts for half of the manufacturing time, and processes have been developed to improve the automation and flexibility of this very difficult task. Material handling provides the integrating link between the different processes — material-handling systems ensure that the required material arrives at the proper place at the right time for the various unit processes and assembly operations.

The section ends with a case study that demonstrates how understanding of the different unit processes can be used to make engineering decisions.

- Material removal (machining) processes
  - Traditional machining
    Drill and reaming
    Turning and boring
    Planing and shaping
    Milling
    Broaching
    Grinding
    Mortality
  - Nontraditional machining
    Electrical discharge machining
    Electrical chemical machining
    Laser beam machining
    Jet machining (water and abrasive)
    Ultrasonic machining
- Phase-change processes
  - Green sand casting
  - Investment casting
- Structure-change processes
  - Normalizing steel
  - Laser surface hardening
- Deformation processes
  - Die forging
  - Press-brake forming
- Consolidation processes
  - Polymer composite consolidation
  - Shielded metal-arc welding
- Mechanical assembly
- Material handling
- Case study: Manufacturing and inspection of precision recirculating ballscrews

References

Altan, T., Oh, S.I., and Gegel, H. 1983. *Metal Forming — Fundamentals and Applications*, ASM International, Metals Park, OH.

ASM Handbook Series, 10th ed., 1996. ASM International, Metals Park, OH.

Bakerjian, R., Ed. 1992. *Design for Manufacturability,* Vol. VI, *Tool and Manufacturing Engineers Handbook,* 4th ed., Society of Manufacturing Engineers, Dearborn, MI.
DeVries, W.R. 1991. *Analysis of Material Removal Processes,* Springer-Verlag, New York.
Kalpakjian, S. 1992. *Manufacturing Engineering and Technology,* Addison-Wesley, Reading, MA.
National Research Council (NRC), 1995. *Unit Manufacturing Processes — Issues and Opportunites in Research,* National Academy Press, Washington, D.C.

Material Removal Processes

These processes, also known as machining, remove material by mechanical, electrical, laser, or chemical means to generate the desired shape and/or surface characteristic. Workpiece materials span the spectrum of metals, ceramics, polymers, and composites, but metals, and particularly iron and steel alloys, are by far the most common. Machining can also improve the tolerances and finish of workpieces previously shaped by other processes, such as forging. Machining is an essential element of many manufacturing systems (ASM, 1989b; Bakerjian, 1992).

Machining is important in manufacturing because

- It is precise. Machining is capable of creating geometric configurations, tolerances, and surface finishes that are often unobtainable by other methods. For example, generally achievable surface roughness for sand casting is 400 to 800 μin. (10 to 20 μm), for forging 200 to 400 μin. (5 to 10 μm), and for die casting 80 to 200 μin. (2 to 5 μm). Ultraprecision machining (i.e., super-finishing, lapping, diamond turning) can produce a surface finish of 0.4 μin (0.01 μm) or better. The achievable dimensional accuracy in casting is 1 to 3% (ratio of tolerance to dimension) depending on the thermal expansion coefficient and in metal forming it is 0.05 to 0.30% depending on the elastic stiffness, but in machining the achievable tolerance can be 0.001%.
- It is flexible. The shape of the final machined product is programmed and therefore many different parts can be made on the same machine tool and just about any arbitrary shape can be machined. In machining, the product contour is created by the path, rather than the shape, of the cutter. By contrast, casting, molding, and forming processes require dedicated tools for each product geometry, thus restricting their flexibility.
- It can be economical. Small lots and large quantities of parts can be relatively inexpensively produced if matched to the proper machining process.

The dominating physical mechanism at the tool/workpiece interface in conventional machining is either plastic deformation or controlled fracture of the workpiece. Mechanical forces are imposed on the workpiece by the application of a tool with sharp edges and higher hardness than the workpiece. However, many new materials are either harder than conventional cutting tools or cannot withstand the high cutting forces involved in traditional machining. Nontraditional manufacturing (NTM) processes can produce precision components of these hard and high-strength materials. NTM processes remove material through thermal, chemical, electrochemical, and mechanical (with high impact velocity) interactions.

Machinability is defined in terms of total tool life, power requirements, and resultant workpiece surface finish. To date, no fundamental relationship incorporates these three factors and thus machinability must be empirically determined by testing.

Process Selection

Machine tools can be grouped into two broad categories:

- Those that generate surfaces of rotation
- Those that generate flat or contoured surfaces by linear motion

Selection of equipment and machining procedures depends largely on these considerations:

- Size of workpiece
- Configuration of workpiece
- Equipment capacity (speed, feed, horsepower range)
- Dimensional accuracy
- Number of operations
- Required surface condition and product quality

For example, Figure 13.2.2 graphically indicates the various tolerance levels that can be typically achieved for common machining unit processes as a function of the size of the workpiece. Such data can help in identifying candidate unit processes that are capable of meeting product requirements.

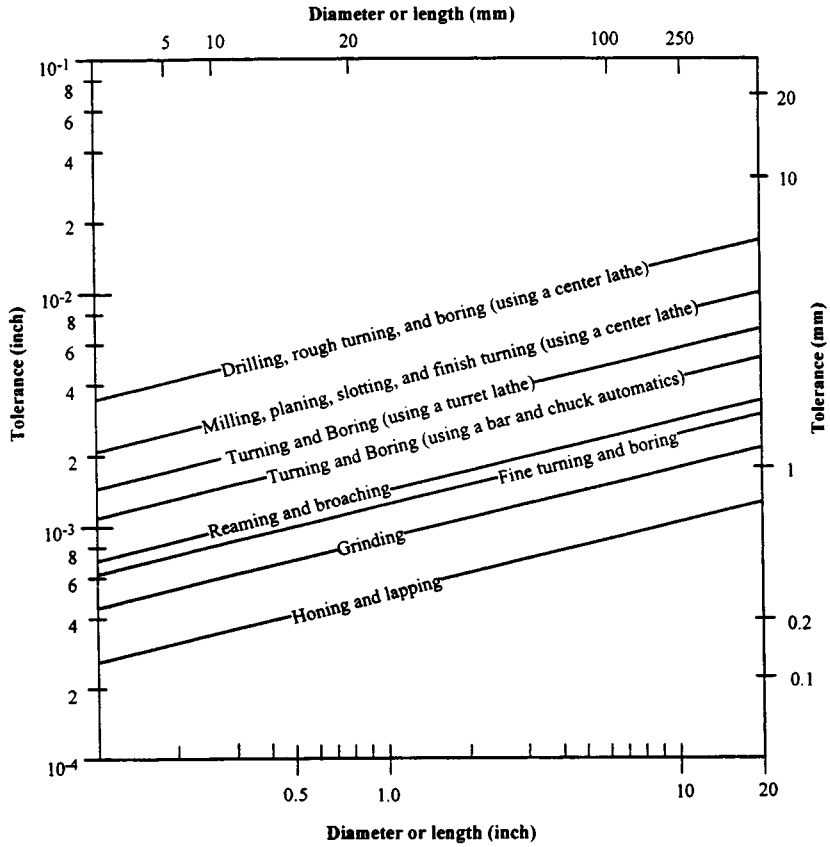


FIGURE 13.2.2 Tolerance vs. dimensional data for machining processes. (From NRC, *Unit Manufacturing Processes*, National Academy Press, Washington, D.C., 1995, 168. With permission.)

Traditional Machining

Steven Y. Liang

Traditional machining processes remove material from a workpiece through plastic deformation. The process requires direct mechanical contact between the tool and workpiece and uses relative motion between the tool and the workpiece to develop the shear forces necessary to form machining chips. The tool must be harder than the workpiece to avoid excessive tool wear. The unit processes described here are a representative sample of the types most likely to be encountered. The reference list at the end of

the section should be consulted for more detailed information on the unit processes discussed below, plus those that are not included here.

Process Kinematics in Traditional Machining. In all traditional machining processes, the surface is created by providing suitable relative motion between the cutting tool and the workpiece. There are two basic components of relative motion: primary motion and feed motion. Primary motion is the main motion provided by a machine tool to cause relative motion between the tool and workpiece. The feed motion, or the secondary motion, is a motion that, when added to the primary motion, leads to a repeated or continuous chip removal. It usually absorbs a small proportion of the total power required to perform a machining operation. The two motion components often take place simultaneously in orthogonal directions.

The functional definitions of turning, milling, drilling, and grinding are not distinctively different, but machining process specialists have developed terminology peculiar to a given combination of functions or machine configurations. Commonly used metal-cutting machine tools, however, can be divided into three groups depending upon the basic type of cutter used: single-point tools, multipoint tools, or abrasive grits.

Dynamic Stability and Chatter. One of the important considerations in selecting a machine tool is its vibrational stability. In metal cutting, there is a possibility for the cutter to move in and out of the workpiece at frequency and amplitude that cause excessive variations of the cutting force, resulting in poor surface quality and reduced life of the cutting tool (Welboum, 1970).

Forced vibrations during cutting are associated with periodic forces resulting from the unbalance of rotating parts, from errors of accuracy in some driving components, or simply from the intermittent engagement of workpiece with multipoint cutters. *Self-excited vibrations* occur under conditions generally associated with an increase in machining rate. These vibrations are often referred to as chatter. All types of chatter are caused by a *feedback loop* within the machine tool structure between the cutting process and the machine frame and drive system. The transfer function of the machine tool, in terms of the stiffness and damping characteristics, plays a critical role in the stability of the overall feedback system. The static stiffness of most machine tools, as measured between the cutting tool and the workpiece, tends to be around 10^5 lb-ft/in. A stiffness of 10^6 lb-ft/in. is exceptionally good, while stiffness of 10^4 lb-ft/in. is poor but perhaps acceptable for low-cost production utilizing small machine tools.

Basic Machine Tool Components. Advances in machine-tool design and fabrication philosophy are quickly eliminating the differences between machine types. Fifty years ago, most machine tools performed a single function such as drilling or turning, and operated strictly stand-alone. The addition of automatic turrets, tool-changers, and computerized numerical control (CNC) systems allowed lathes to become *turning centers* and milling machines to become *machining centers*. These multiprocess centers can perform a range of standard machining functions: turning, milling, boring, drilling, and grinding (Green, 1992).

The machine tool *frame* supports all the active and passive components of the tool — spindles, table, and controls. Factors governing the choice of frame materials are: resistance to deformation (hardness), resistance to impact and fracture (toughness), limited expansion under heat (coefficient of thermal expansion), high absorption of vibrations (damping), resistance to shop-floor environment (corrosion resistance), and low cost.

Guide ways carry the workpiece table or spindles. Each type of way consists of a *slide* moving along a track in the frame. The slide carries the workpiece table or a spindle. The oldest and simplest way is the *box way.* As a result of its large contact area, it has high stiffness, good damping characteristics, and high resistance to cutting forces and shock loads. Box slides can experience stick-slip motion as a result of the difference between dynamic and static friction coefficients in the ways. This condition introduces positioning and feed motion errors. A *linear way* also consists of a rail and a slide, but it uses a rolling-element bearing, eliminating stick-slip. Linear ways are lighter in weight and operate with less friction,

so they can be positioned faster with less energy. However, they are less robust because of the limited surface contact area.

Slides are moved by hydraulics, rack-and-pinion systems, or screws. *Hydraulic pistons* are the least costly, most powerful, most difficult to maintain, and the least accurate option. Heat buildup often significantly reduces accuracy in these systems. Motor-driven *rack-and-pinion* actuators are easy to maintain and are used for large motion ranges, but they are not very accurate and require a lot of power to operate. Motor-driven screws are the most common actuation method. The screws can either be lead screws or ballscrews, with the former being less expensive and the latter more accurate. The *recirculating ballscrew* has very tight backlash; thus, it is ideal for CNC machine tools since their tool trajectories are essentially continuous. A disadvantage of the ballscrew systems is the effective stiffness due to limited contact area between the balls and the thread. (*Note:* a case study at the end of this section discusses the manufacture of precision ballscrews.)

Electric motors are the prime movers for most machine tool functions. They are made in a variety of types to serve three general machine tool needs: spindle power, slide drives, and auxiliary power. Most of them use three-phase AC power supplied at 220 or 440 V. The design challenge with machine tools and motors has been achieving high torque throughout a range of speed settings. In recent years, the operational speed of the spindle has risen significantly. For example, conventional speeds 5 years ago were approximately 1600 rpm. Today, electric motors can turn at 12,000 rpm and higher. Higher speeds cause vibration, which makes use of a mechanical transmission difficult. By virtue of improvement in motor design and control technology, it is now possible to quickly adjust motor speed and torque. Mechanical systems involving more than a three-speed transmission are becoming unnecessary for most high-speed and low-torque machines. Spindle motors are rated by horsepower, which generally ranges from 5 to 150 hp (3.7 to 112 kW) with the average approximately 50 hp (37 kW). Positioning motors are usually designated by torque, which generally ranges from 0.5 to 85 lb-ft (0.2 to 115 Nm).

The *spindle* delivers torque to the cutting tool, so its precision is essential to machine tool operation. The key factors influencing precision are bearing type and placement, lubrication, and cooling.

Cutting Tool Materials. The selection of cutting tool materials is one of the key factors in determining the effectiveness of the machining process (ASM, 1989b). During cutting, the tool usually experiences high temperatures, high stresses, rubbing friction, sudden impact, and vibrations. Therefore, the two important issues in the selection of cutting tool materials are hardness and toughness. *Hardness* is defined as the endurance to plastic deformation and wear; hardness at elevated temperatures is especially important. *Toughness* is a measure of resistance to impact and vibrations, which occur frequently in interrupted cutting operations such as milling and boring. Hardness and toughness do not generally increase together, and thus the selection of cutting tool often involves a trade-off between these two characteristics.

Cutting tool materials are continuously being improved. Carbon steels of 0.9 to 1.3% carbon and tool steels with alloying elements such as molybdenum and chromium lose hardness at temperatures above 400°F (200°C) and have largely been replaced by *high-speed steels* (HSS). HSS typically contains 18% tungsten or 8% molybdenum and smaller amounts of cobalt and chromium. HSSs retain hardness up to 1100°F (600°C) and can operate at approximately double the cutting speed with equal life. Both tool steels and HSS are tough and resistive to fracture; therefore, they are ideal for processes involving interrupted engagements and machine tools with low stiffness that are subject to vibration and chatter.

Powder metallurgy (P/M) high-speed tool steels are a recent improvement over the conventionally cast HSSs. Powder metallurgy processing produces a very fine microstructure that has a uniform distribution of hard particles. These steels are tougher and have better cutting performance than HSS. Milling cutters are becoming a significant application for these cutting tool materials.

Cast cobalt alloys, popularly known as Stellite tools, were introduced in 1915. These alloys have 38 to 53% cobalt, 30 to 33% chromium, and 10 to 20% tungsten. Though comparable in room temperature hardness to HSS tools, cast cobalt alloy tools retain their hardness to a much higher temperature, and they can be used at 25% higher cutting speeds than HSS tools.

Cemented carbides offered a four- or fivefold increase in cutting speeds over conventional HSS. They are much harder, but more brittle and less tough. The first widely used cemented carbide was tungsten carbide (WC) cemented in a ductile cobalt binder. Most carbide tools in use now are a variation of the basic WC-Co material. For instance, WC may be present as single crystals or a solid solution mixture of WC-TiC or WC-TiC-TaC. These solid solution mixtures have a greater chemical stability in the cutting of steel. In general, cemented carbides are good for continuous roughing on rigid machines, but should avoid shallow cuts, interrupted cuts, and less rigid machines because of likely chipping.

A thin layer of TiC, TiN, or Al_2O_3 can be applied to HSS or carbide substrate to improve resistance to abrasion, temperature, friction, and chemical attacks. The *coated tools* were introduced in the early 1970s and have gained wide acceptance since. Coated tools have two or three times the wear resistance of the best uncoated tools and offer a 50 to 100% increase in speed for equivalent tool life.

Ceramic tools used for machining are based on alumina (Al_2O_3) or silicon nitride (Si_3N_4). They can be used for high-speed finishing operations and for machining of difficult-to-machine advanced materials, such as superalloys (Komanduri and Samanta, 1989). The alumina-based materials contain particles of titanium carbide, zirconia, or silicon carbide whiskers to improve hardness and/or toughness. These materials are a major improvement over the older ceramic tools. Silicon nitride–based materials have excellent high-temperature mechanical properties and resistance to oxidation. These materials also have high thermal shock resistance, and thus can be used with cutting fluids to produce better surface finishes than the alumina tools.

These tools can be operated at two to three times the cutting speeds of tungsten carbide, usually require no coolant, and have about the same tool life at higher speeds as tungsten carbide does at lower speeds. However, ceramics lack toughness; therefore, interrupted cuts and intermittent application of coolants can lead to premature tool failure due to poor mechanical and thermal shock resistance.

Cermets are titanium carbide (TiC) or titanium carbonitride particles embedded in a nickel or nickel/molybdenum binder. These materials, produced by the powder metallurgy process, can be considered as a type of cemented carbide. They are somewhat more wear resistant, and thus can be used for higher cutting speeds. They also can be used for machining of ferrous materials without requiring a protective coating.

Cubic boron nitride (CBN) is the hardest material at present available except for diamond. Its cost is somewhat higher than either carbide or ceramic tools but it can cut about five times as fast as carbide and can hold hardness up to 200°C. It is chemically very stable and can be used to machine ferrous materials.

Industrial *diamonds* are now available in the form of polycrystalline compacts for the machining of metals and plastics with greatly reduced cutting force, high hardness, good thermal conductivity, small cutting-edge radius, and low friction. Recently, diamond-coated tools are becoming available that promise longer-life cutting edges. Shortcomings with diamond tools are brittleness, cost, and the tendency to interact chemically with workpiece materials that form carbides, such as carbon steel, titanium, and nickel.

Wear of Cutting Tool Materials. Cutting tools are subjected to large forces under conditions of high temperature and stress. There are many mechanisms that cause wear.

- Adhesion. The tool and chip can weld together; wear occurs as the welded joint fractures and removes part of the tool material, such as along a tool cutting edge.
- Abrasion. Small particles on the wear surface can be deformed and broken away by mechanical action due to the high localized contact stresses; these particles then abrade the cutting tool. Typically, this is the most common wear mode.
- Brittle fracture. Catastrophic failure of the tool can occur if the tool is overloaded by an excessive depth of cut and/or feed rate.
- Diffusion. Solid-state diffusion can occur between the tool and the workpiece at high temperatures and contact pressures, typically at an area on the tool tip that corresponds to the location of

maximum temperature, e.g., cemented carbide tools used to machine steel. High-speed machining results in higher chip temperatures, making this an increasingly important wear mode.

- Edge chipping.
- Electrochemical. In the presence of a cutting fluid, an electrochemical reaction can occur between the tool and the workpiece, resulting in the loss of a small amount of tool material in every chip.
- Fatigue.
- Plastic deformation.

Single-Point Cutting Tool Geometry. Figure 13.2.3 depicts the location of various angles of interest on a single-point cutting tool. The most significant angle is the *cutting-edge* angle, which directly affects the shear angle in the chip formation process, and therefore greatly influences tool force, power requirements, and temperature of the tool/workpiece interface (ASM, 1989a). The larger the positive value of the cutting-edge angle, the lower the force, but the greater the load on the cutting tool. For machining higher-strength materials, negative rake angles are used. *Back rake* usually controls the direction of chip flow and is of less importance than the side rake. Zero back rake makes the tool spiral more tightly, whereas a positive back rake stretches the spiral into a longer helix. *Side rake* angle controls the thickness of the tool behind the cutting edge. A thick tool associated with a small rake angle provides maximum strength, but the small angle produces higher cutting forces than a larger angle; the large angle requires less motor horsepower.

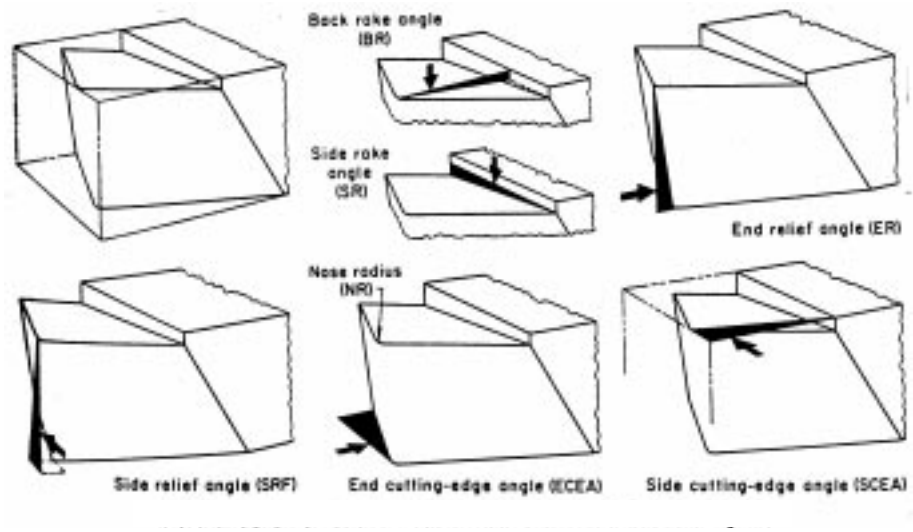


FIGURE 13.2.3 Standard nomenclature for single-point cutting tool angles. (From *ASM Handbook, Machining*, Vol. 16, 9th ed., ASM International, Metals Park, OH, 1989, 141. With permission.)

The *end relief angle* provides clearance between the tool and the finished surface of the work. Wear reduces the angle. If the angle is too small, the tool rubs on the surface of the workpiece and mars the finish. If the angle is too large, the tool may dig into the workpiece and chatter, or show weakness and fail through chipping. The *side relief angle* provides clearance between the cut surface of the work and the flank of the tool. Tool wear reduces the effective portion of the angle closest to the workpiece. If this angle is too small, the cutter rubs and heats. If the angle is too large, the cutting edge is weak and the tool may dig into the workpiece. The *end cutting-edge angle* provides clearance between the cutter and the finished surface of the work. An angle too close to zero may cause chatter with heavy feeds, but for a smooth finish the angle on light finishing cuts should be small.

Machinability. Optimum speed and feed for machining depend on workpiece material, tool material, characteristics of the cut, cutting tool configuration, rigidity of setup, tolerance, and cutting fluid. Consequently, it is not possible to recommend universally applicable speeds and feeds.

Drilling and Reaming

Description and Applications. *Drilling* is the most widely used process for making circular holes of moderate accuracy. It is often a preliminary step to other processes such as tapping, boring, or reaming. *Reaming* is used to improve the accuracy of a hole while increasing its diameter. Holes to be reamed are drilled undersize.

Key System Components. Drills are classified by the material from which they are made, method of manufacture, length, shape, number and type of helix or flute, shank, point characteristics, and size series (Table 13.2.1). Selection of drill depends on several factors (ASM, 1989b):

TABLE 13.2.1 Common Drill Types

Drill Type	Description	Application
Core	Has large clearance for chips	Roughing cuts; enlarging holes
General-purpose (jobber)	Conventional two-flute design; right-hand (standard) or left-hand helix. Available with flute modification to break up long chips	General-purpose use, wide range of sizes
Gun	Drill body has a tube for cutting fluid; drill has two cutting edges on one side and counter-balancing wear pads on the other side	Drill high-production quantities of holes without a subsequent finishing operation
High helix	Wide flutes and narrow lands to provide a large bearing surface	Soft materials, deep holes, high feed rates
Low helix	Deep flutes facilitate chip removal	Soft materials, shallow holes
Oil hole	Has holes through the drill body for pressurized fluid	Hard materials, deep holes, high feed rate
Screw-machine	Short length, short flutes, extremely rigid	Hard materials; nonflat surfaces
Step	Two or more drill diameters along the drill axis	Produce multiple-diameter holes, such as for drilling/countersinking
Straight flute	Flutes parallel to the drill axis minimize torquing of the workpiece	Soft materials; thin sheets

- Hardness and composition of the workpiece, with hardness being more important
- Rigidity of the tooling
- Hole dimensions
- Type of drilling machine
- Drill application — originating or enlarging holes
- Tolerances
- Cost

The most widely used drill is the general-purpose twist drill, which has many variations. The flutes on a twist drill are helical and are not designed for cutting but for removing chips from the hole. Typical twist drills are shown in Figure 13.2.4.

Machining forces during *reaming* operations are less than those of drilling, and hence reamers require less toughness than drills and often are more fragile. The reaming operation requires maximum rigidity in the machine, reamer, and workpiece.

Most *reamers* have two or more flutes, either parallel to the tool axis or in a helix, which provide teeth for cutting and grooves for chip removal. The selection of the number of flutes is critical: a reamer with too many flutes may become clogged with chips, while a reamer with too few flutes is likely to chatter (Table 13.2.2).

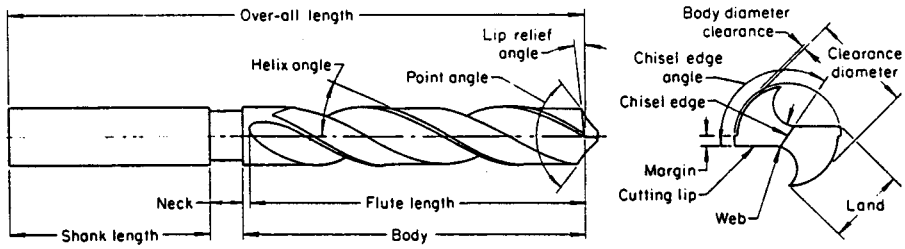


FIGURE 13.2.4 Design features of a typical straight-shank twist drill. (From *ASM Handbook, Machining*, Vol. 16, 9th ed., ASM International, Metals Park, OH, 1989, 218. With permission.)

TABLE 13.2.2 Common Reamer Types

Reamer Type	Description	Application
Adjustable	Tool holder allows adjustment of the reamer diameter to compensate for tool wear, etc.	High-rate production
End-cutting	Cutting edges are at right angles to the tool axis	Finish blind holes, correct deviations in through-holes
Floating blade	Replaceable and adjustable cutting edges to maintain tight tolerances	High-speed production (workpiece rotated, tool stationary)
Gun	Hollow shank with a cutting edge (e.g., carbide) fastened to the end and cutting fluid fed through the stem	High-speed production (workpiece rotated, tool stationary)
Shell	Two-piece assemblies, mounted on arbors, can be adjusted to compensate for wear	Used for finishing operations (workpiece rotated, tool stationary)
Spiral flute	Flutes in a helix pattern, otherwise same as straight-flute reamer	Difficult to ream materials, and holes with irregularities
Straight flute	Flutes parallel to the tool axis, typically pointed with a 45° chamfer	General-purpose, solid reamer

Machining Parameters. The optimal speed and feed for drilling depend on workpiece material, tool material, depth of hole, design of drill, rigidity of setup, tolerance, and cutting fluid. For *reaming* operations, hardness of the workpiece has the greatest effect on machinability. Other significant factors include hole diameter, hole configuration (e.g., hole having keyways or other irregularities), hole length, amount of stack removed, type of fixturing, accuracy and finish requirements, size of production run, and cost. Most reamers are more easily damaged than drills; therefore, the practice is to ream a hole at no more than two thirds of the speed at which it was drilled.

Capabilities and Process Limitations. Most drilled holes are $^1/_8$ to 1 in. (3.2 to 40 mm) in diameter. However, drills are available for holes as small as 0.001 in. (0.03 mm) (microdrilling), and special drills are available as large as 6 in. (150 mm) in diameter. The range of length-to-diameter (L/D) of holes that can be successfully drilled depends on the method of driving the drill and the straightness requirements. In the simplest form of drilling in which a rotating twist drill is fed into a fixed workpiece, best results are obtained when L/D is <3. But by using special tools, equipment, and techniques, straight holes can be drilled with L/D = 8 or somewhat greater. Nonconventional machining processes can also generate high-aspect-ratio holes in a wide variety of materials.

Reaming and boring are related operations. Hole diameter and length, amount of material to be removed, and required tolerance all influence which process would be most efficient for a given application (ASM, 1989b). Most holes reamed are within the size range of $^1/_8$ to 1 inch (3.2 to 40 mm), although larger and smaller holes have been successfully reamed. For most applications with standard reamers, the length of a hole that can be reamed to required accuracy ranges from slightly longer to much shorter than the cutting edges of the reamer, but there are many exceptions to this general rule-of-thumb. Tolerances of 0.001 to 0.003 in. (0.03 to 0.08 mm) with respect to the diameter are readily

achievable in production reaming operations. Surface finish for annealed steels can be held within the range of 100 to 125 µin. (2.50 to 3.20 µm), but a surface as smooth as 40 µin. (1 µm) can be obtained under appropriate processing conditions (ASM, 1989b).

Turning and Boring

Description and Applications. *Turning* produces external cylindrical surfaces by removing material from a rotating workpiece, usually with a single-point cutting tool in a lathe. *Boring* is this same process applied for enlarging or finishing internal surfaces of revolution.

Key System Components. The basic equipment for turning is an *engine lathe* that consists of a bed, a headstock, a carriage slide, a cross slide, a tool holder mounted on the cross slide, and a source of power for rotating the workpiece (Table 13.2.3). Engine lathes are often modified to perform additional types of machining operations through the use of attachments. Most turning machines can also perform boring operations, but boring machines may not be able to perform turning operations. Sizes of lathes range from fractional horsepower to greater than 200 hp.

TABLE 13.2.3 Typical Lathes Used for Turning

Lathe	Description	Applications
Bench lathe	An engine lathe that can be placed on a workbench	Small workpieces and prototype parts
Engine lathe	Has a leadscrew that moves the slide uniformly along the bed; available with chucking or centering headstock	Chucking type allows centering and clamping for rotation, e.g., holding castings or forgings Centering type secures workpiece between pointed centers, e.g., for turning long workpieces, such as shafts
Gap-frame lathe	Modified engine lathe for turning larger diameter parts	Workpieces requiring off-center mounting or irregular protuberances
Numerically controlled lathe	Uses a computer program to control the lathe to generate the desired shape	Produces consistent parts in a CAD/CAM environment
Tracer-controlled lathe	A duplicating lathe that uses a stylus moving over a template to control the cutting tool	Manufacture of prototype parts and low-rate production

TABLE 13.2.4 Typical Machines Used for Boring

Boring Machine	Description	Applications
Bar (screw) machine	Modified turret lathe to handle bars and tubes	Parts made from bars or tubes
Engine lathe	Versatile machine; essentially same machine as used for turning	Bores one hole at a time in a single part; limitations regarding workpiece size and configuration
Horizontal boring mill	Workpiece remains stationary and tool rotates	Wide variety of parts; cost-effective for a relatively high production volume
Precision boring machine	Vertical and horizontal models	Parts requiring extreme tolerances
Special-purpose machines	Boring machine modified for specialized application	Single-purpose applications with high production rates
Turret lathe	Has rotating turret on a lathe, tooled for multiple machining operations	More versatile than engine lathe; supports high production rates
Vertical boring mill	Same basic components as a lathe	Very large, heavy, or eccentric workpieces
Vertical turret lathe	Same features as vertical boring mill; may also have a second vertical head	Flexible machine, useful in CAD/CAM environment; simultaneous multiple machining operations possible

Machines used for boring are noted for their rigidity, adaptability, and ability to maintain a high degree of accuracy (Table 13.2.4). For extremely large workpieces, weighing thousands of pounds, the boring cutting tool is rotated and the workpiece is fixed.

Machine Tool and Machining Parameters. In turning and boring operations, a single-point tool is traversed longitudinally along the axisymmetric workpiece axis parallel to the spindle. A tangential force is generated when the cutting tool engages the rotating work. This force is generally independent of the cutting speed and directly proportional to the depth of cut for a particular material, tool shape (particularly side rake angle), and feed rate. That force, when multiplied by the surface speed of the workpiece, estimates the net horsepower required to remove material. The extent to which workpiece material affects required machining power is illustrated in Figure 13.2.5. Moving the tool longitudinally requires much less power (ASM, 1989b).

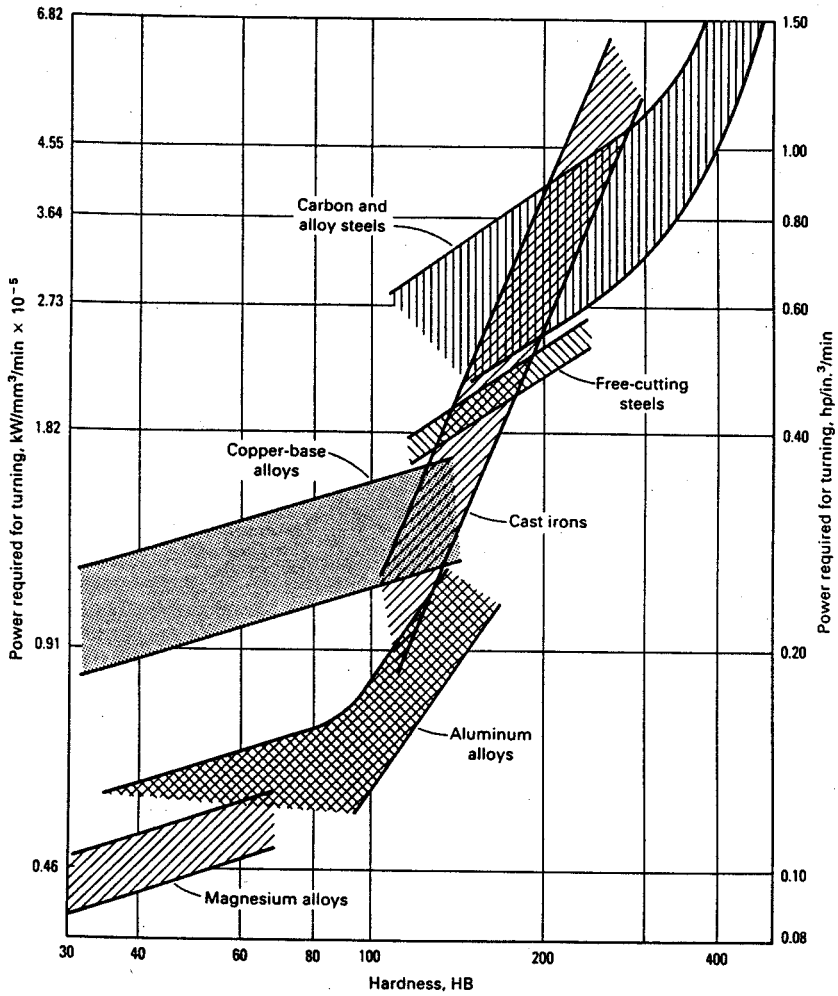


FIGURE 13.2.5 Effect of workpiece composition and hardness on power required for turning. (From *ASM Handbook, Machining*, Vol. 16, 9th ed., ASM International, Metals Park, OH, 1989, 136. With permission.)

To minimize the number of cuts required, the depth of cut should be as great as possible, which is limited by the strength of the part and the fixturing, and the power output of the machine tool. The feed rate is a function of the finish desired and the strength and rigidity of the part and machine tool.

Capabilities and Limitations. Components that range in size from those used in watches up to large steel propeller shafts more than 80 ft (24 m) long are regularly turned. Aluminum parts over 10 ft (3 m) in diameter have been turned. In practice, the weight of the workpiece per unit of volume determines the size of the workpiece that is practical to turn. Large, heavy parts can be turned in a vertical boring machine. Irregular-shaped parts, such as crankshafts, may require the use of counter-weighting to achieve dynamic balance for vibration-free turning.

For both turning and boring, the rotation speed, feed, and depth of cut determine the rate of material removal and resulting surface quality. Feed rate for most applications falls between 0.005 and 0.020 in./rev (0.13 and 0.51 mm/rev). Finishing cuts have a significantly lower feed rate (e.g., 0.001 in./rev, 0.03 mm/rev), and roughing cuts are made at a significantly higher feed rate (e.g., 0.25 in./rev, 6.35 mm/rev). Boring is not limited by the L/D ratio of a hole — this ratio can be as great as 50 if the tool bar and workpiece are adequately supported.

There are many potential sources of tolerance error in turning and boring operations. The more common errors are summarized in Tables 13.2.5 and 13.2.6.

TABLE 13.2.5 Factors Affecting Dimensional Accuracy in Turning

Quantity	Basic Cause of Inaccuracy	Corrective Actions
Diametrical roundness	Lathe spindle runout	Preload angular bearings to eliminate side and end movement
Diameter variation (taper)	Parallelism of spindle to longitudinal travel	Establish true parallelism between the axis of rotation of the workpiece and the longitudinal travel of the cutting tool
Face flatness	Lack of true normality of cross slide to axis of rotation	Precision align cross slide to axis of rotation
Length dimensions parallel to the axis of rotation	Improper positioning of longitudinal slide	Adjust positioning of longitudinal slide
Diameter accuracy	Inadequate gauging	Employ proper measurement technique

TABLE 13.2.6 Factors Affecting Dimensional Accuracy in Boring

Quantity	Basic Cause of Inaccuracy	Corrective Actions
Diameter accuracy	Inadequate gauging, or heat generated by the machining action	Employ proper measurement techniques; use cutting fluid to control temperature
Taper of the cylindrical bore	Deflection of the boring bar	Reduce the boring bar unsupported length; use a higher-stiffness material for the boring bar
Roundness, as determined by the variation in radius about a fixed axis	Finish cut is not concentric with the previous cut; or out-of-balance workpiece or holder	Begin with a semifinish cut, followed by a finish cut; carefully balance the initial setup
Concentricity of one surface with another	Too great a clamping force on the workpiece	Redesign clamping fixture, or use a precision boring machine
Squareness and parallelism of holes in relation to other features of the work	Dimensional changes in machine components	Maintain constant ambient temperature; maintain cutting fluid at constant temperature; or stabilize oil temperature

Planing and Shaping

Description and Applications. *Planing* is a widely used process for producing flat, straight surfaces on large workpieces. A variety of contour operations and slots can be generated by use of special attachments. It is often possible to machine a few parts quicker by planing than by any other method. *Shaping* is a process for machining flat and contour surfaces, including grooves and slots.

Principle of Operation. *Planers* develop cutting action from straight-line reciprocating motion between one or more single-point tools and the workpiece; the work is reciprocated longitudinally while the tools

are fed sideways into the work. Planer tables are reciprocated by either mechanical or hydraulic drives, with mechanical drives predominating.

Shapers use a single-point tool that is supported by a ram which reciprocates the tool in a linear motion against the workpiece. The workpiece rests on a flat bed and the cutting tool is driven toward it in small increments by ram strokes. Shapers are available with mechanical and hydraulic drives, with mechanical drives predominating.

Machine Tool and Machining Parameters. Planing and shaping are rugged machining operations during which the workpiece is subjected to significant cutting forces. These operations require high clamping forces to secure the workpiece to the machine bed.

In general, it is advisable to plane steel with as heavy a feed and as high a speed as possible to promote good chip-formation conditions so that chip breakers are not needed. Carbide cutters allow cutting speed to be increased from 225 to 300 surface feet per minute (sfm) (70 to 90 m/min). For best results, uniform cutting speed and feed are maintained throughout the entire stoke (ASM, 1989b).

In general, speeds are related to workpiece material characteristics and associated machinability. Feeds are influenced by the workpiece machinability, but also by ram speed, depth of cut, and required dimensional accuracy and surface finish. Common practice in shaping is to make roughing cuts at as high a feed and slow a speed as practical, and make finishing cuts at a low feed rate and high speed (ASM, 1989b). For low carbon steel, a typical speed for a roughing cut is 50 sfm (15 m/min), while for a finishing cut it is 80 sfm (25 m/min) using a conventional cutting tool. Similarly for aluminum, a roughing cut of 150 sfm (45 m/min) is typically followed by a finishing cut of 200 sfm (60 m/min).

There is a practical lower bound on minimum feed rate. Feed rates that are too low will cause the tool to chatter; feed rates less than 0.005 in. (0.125 mm) are seldom used in shaping. Similarly, shallow cuts (less than 0.015 in., 0.38 mm) will cause chatter during shaping.

Key System Components. Planers are available in a wide range of sizes (Table 13.2.7). Tools are available in a variety of configurations for undercutting, slotting, and straight planing of either horizontal or vertical surfaces (ASM, 1989b).

TABLE 13.2.7 Types of Planers

Planer Type	Description	Application
Double housing	Two vertical uprights support the crossrail which in turn supports the tools	Rigid machine; restricts the width of workpiece
Open side	A single upright column supports a cantilevered crossrail; less rigid than the double-housing type	Accommodates wide workpieces which can overhang one side of the table without interfering with the planer operation

Shapers are available in a large variety of sizes (Table 13.2.8), ranging from small models with a maximum stroke length of less than 6 in. (150 mm) to large machines with a maximum stroke of 36 in. (914 mm). On each machine, the length of stroke can be varied from its maximum to slightly less than 1 in. (25 mm) for the largest machine, and to $^1/_8$ in. (3.2 mm) for the smallest machine.

TABLE 13.2.8 Types of Shapers

Shaper Type	Description	Application
Horizontal	The ram drives the tool in the horizontal direction; uses plane or universal table (rotates on three axes)	Gears, splined shafts, racks, and so on; not used for rate production
Vertical	The ram operates vertically, cutting on the downstroke	Slots, grooves, keyways; matching die sets, molds, fixtures; not used for rate production

Capabilities and Process Limitations. Planing is a precision process in which flatness can be held within 0.0005 in. (0.013 mm) total indicated runout (TIR) on workpieces up to 4 ft^2 (0.4 m^2). Although planing is most widely used for machining large areas, it is also used for machining smaller parts, although 12 in. is about the minimum distance for a planing stoke. Size of the workpiece that can be planed is limited by the capacity of the planing equipment.

Shaping is a versatile process in which setup time is short and relatively inexpensive tools can be used. Under good conditions, a shaper can machine a square surface of 18 in. on a side (0.2 m^2) to a flatness within 0.001 in. (0.025 mm); under optimum conditions this can be improved to 0.0005 in. (0.013 mm). The size of the workpiece that can be shaped is limited by the length of the stoke, which is usually about 36 in. (914 mm). Shaping should be considered for machining flat surfaces in these instances:

- Required flatness cannot be achieved by another method
- Production quantity is insufficient to justify the tooling costs of milling or broaching

Planing and shaping are interrupted cutting processes, and are comparatively inefficient means of metal removal; for example, shaping costs five times that of milling, exclusive of the tooling and setup costs.

Milling

Description and Applications. Milling is a versatile, efficient process for metal removal. It is used to generate planar and contour surfaces through the action of rotating multiple-tooth cutters. Surfaces having almost any orientation can be machined because both the workpiece and cutter can move in more than one direction at the same time.

Principle of Operation. Cutters with multiple cutting edges rotate in a spindle. The machining process is interrupted as the teeth of the milling cutter alternately engage and disengage from the workpiece.

Key System Components. Most milling is done in machines designed for milling (Table 13.2.9). Milling can also be done by any machine tool that can rigidly hold and rotate a cutter while feeding a workpiece into the cutter. Milling machines are usually classified in terms of their appearance: knee-and-column, bed-type, planar-type, and special purpose. The knee-and-column configuration is the simplest milling machine design. The workpiece is fixed to a bed on the knee and the tool spindle is mounted on a column, as depicted in Figure 13.2.6. For very large workpieces, gantry or bridge-type milling machines are used. Machines having two columns can provide greater stability to the cutting spindle(s). Special-purpose machines are modifications of the three basic models.

The usual power range for knee-and-column machines is 1 to 50 hp (0.75 to 37 kW). Bed-type machines are available in a wide range of sizes, up to 300 hp (225 kW). Planar-type machines are available from 30 to 100 hp (22 to 75 kW).

There is a wide variety of milling cutters, using the full range of cutting tool materials; there are three basic constructions (ASM, 1989b):

- Solid — Made from a single piece of HSS or carbide; cutters can be tipped with a harder material; teeth can designed for specific cutting conditions; low initial cost.
- Inserted blade — Usually made from HSS, carbide, or cast alloy; individual blades can be replaced as they wear out, saving replacement cost; ideal for close-tolerance finishing.
- Indexable insert — Cutter inserts are made from carbide, coated carbide, ceramic, or ultrahard material such as diamond; each insert has one or more cutting edges; as inserts wear, they are repositioned to expose new cutting surface or indexed to bring another cutting insert on line. These inserts, widely used in computer-controlled machines due to their performance and flexibility, can produce a rougher surface than the other tool constructions and require somewhat higher cutting forces to remove metal.

TABLE 13.2.9 Types of Milling Machines

Milling Machine	Description	Application
Knee-and-column	Six basic components: • Base — the primary support • Column — houses spindle and drive • Overarm — provides support for the arbor-mounted cutting tools • Knee — supports the table, saddle, and workpiece; provides vertical movement • Saddle — provides 1° of horizontal motion • Table — directly supports the workpiece and provides a second degree of horizontal motion	Widely used for low production milling; provides three-axis movement; primary drawback is lack of rigidity due to the number of joints
Bed-type	Table and saddle mounted on a bed in fixed vertical position; vertical motion obtained by movement of the spindle carrier; available with horizontal or vertical spindle	Very rigid machine; permits deep machining cuts and close dimensional control
Planar-type (adjustable rail)	Can accommodate almost any type of spindle for driving cutters and boring bars; utilizes several milling heads	Use for mass-production milling; can perform simultaneous milling and boring operations
Special purpose	Many possible configurations involving major modifications or combinations of the basic types of milling machines; adapted for automated control	Optimized for high-volume production; includes profilers and machining centers; these machines are capable of performing multiple simultaneous machining operations

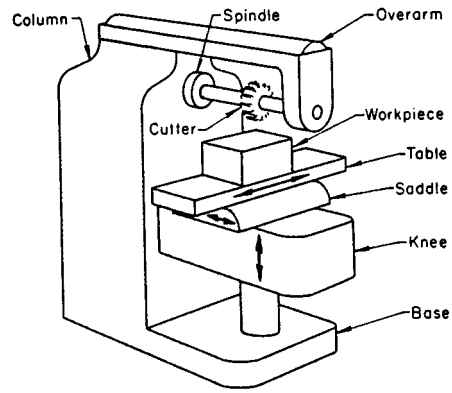


FIGURE 13.2.6 Principal components of a plain knee-and-column milling machine with a horizontal spindle. (From *ASM Handbook, Machining*, Vol. 16, 9th ed., ASM International, Metals Park, OH, 1989, 304. With permission.)

Milling cutters are also described by the location of the cutting edges, as described in Table 13.2.10. Several cutters are depicted in Figure 13.2.7.

Machine Tool and Machining Parameters. The angular relationships of the cutting edge greatly influence cutting efficiency, analogous to single-point cutting tools. A milling cutter should have enough teeth to ensure uninterrupted contact with the workpiece, yet not so many so as to provide too little space between the teeth to make chip removal difficult.

Milling speed varies greatly depending on workpiece material composition, speed, feed, tool material, tool design, and cutting fluid. Speeds as low as 20 sfm (6.1 m/min) are employed for milling low machinability alloys, while speeds as high as 20,000 sfm (6100 m/min) have been reported for milling aluminum (ASM, 1989b). If the setup is sufficiently rigid, carbide or carbide-tipped cutters can be operated three to ten times faster than HSS cutters; top speed is usually constrained by onset of tool chatter.

TABLE 13.2.10 Types of Milling Cutters

Milling Cutters	Description	Application
Peripheral mills	Cutting is primarily done by teeth on the periphery of the cutting tool; mounted on an arbor having its axis parallel to the machined surface	Removing metal from simple flat surfaces; milling contoured surfaces and surfaces having two or more angles or complex forms
Face mills	Machining action is accomplished by the bevel cutting edge located along the circumference of the mill; driven by a spindle on an axis perpendicular to the surfaced being milled	Can be more efficient at removing material than peripheral milling; very rigid tool setup possible; can achieve tight tolerances
End mills	Incorporate cutting edges on both the face and the periphery; can be used for face cuts and periphery cuts	Allow multiple operations without changing cutters; cutters can have difficulty in maintaining dimensional accuracy due to long unsupported length
Special mills	Can be almost any design	Optimized for a particular task

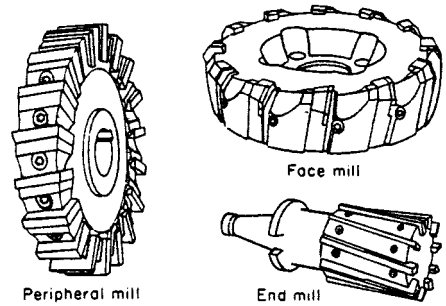


FIGURE 13.2.7 Three typical milling cutters. (From *ASM Handbook, Machining*, Vol. 16, 9th ed., ASM International, Metals Park, OH, 1989, 311. With permission.)

For highest efficiency in removing metal while minimizing chatter conditions, the feed per tooth should be as high as possible. The optimum feed rate is influenced by a number of factors (ASM, 1989b): type of cutter, number of teeth on the cutter, cutter material, workpiece machinability, depth of cut, width of cut, speed, rigidity of the setup, and machine power. The surface finish obtainable by milling can be quite good. A finish of 125 μin. (3.2 μm) can be readily achieved under normal circumstances with HSS mills, and finishes of 63 μin. (1.6 μm) are common if carbide tools are used. With careful selection of cutters and stringent control of process conditions, a finish of 10 min (0.25 mm) can be produced.

Process Limitations. The initial cost of a milling machine is considerably greater than that of a planar or a shaper that can machine workpieces of similar size to similar finishes. Milling tools usually cost up to 50 times as much as tools for planers and shapers, and the setup time is usually longer. However, milling is far more efficient in removing material, and milling machines are commonly highly automated. Therefore, milling is preferred for production operations.

Grinding is often preferred to milling when the amount of metal to be removed is small and the dimensional accuracy and surface finish are critical. Milling and grinding are frequently used in combination.

Broaching

Description and Applications. Broaching is a precision machining process. It is very efficient since both roughing cuts and finishing cuts are made during a single pass of the broach tool to produce a smooth surface, and further finishing is usually not necessary. Consequently, close tolerances can be readily achieved at a reasonable cost for high rates of production.

Broaches are expensive multitoothed cutting tools. Thus, the process is usually employed for low or high production when broaching is the only practical method to produce the required dimensional tolerance and surface quality. An example of the latter case is the dovetail slots in jet engine turbine disks.

Principle of Operation. Broaching is a machining process similar to planing. A broach is essentially a tapered bar into which teeth are cut, with the finishing teeth engaging last on the end with the larger diameter. A single broach has teeth for rough cutting, semifinishing, and finishing. Broaching involves pushing or pulling a broach in a single pass through a hole or across a surface. As the broach moves along the workpiece, cutting is gradual as each successive tooth engages the workpiece, removing a small amount of material. Overall machining forces are much greater than that of other machining methods, and consequently broaching is considered to be the most severe of all machining operations.

Key System Components. Broaching machines are categorized as horizontal or vertical, depending on the direction of broach travel. Industry usage is almost evenly divided between these two categories. The selection of machine type depends heavily on the configuration of the workpiece and available space in the factory, considering both floor space and vertical clearance requirements.

Broaches can be categorized by the method through which they are actuated (push or pull), by type of cut (internal or external), and by the construction of the broach body. Figure 13.2.8 depicts typical internal and external broaching operations. Table 13.2.11 describes broaches according to their construction.

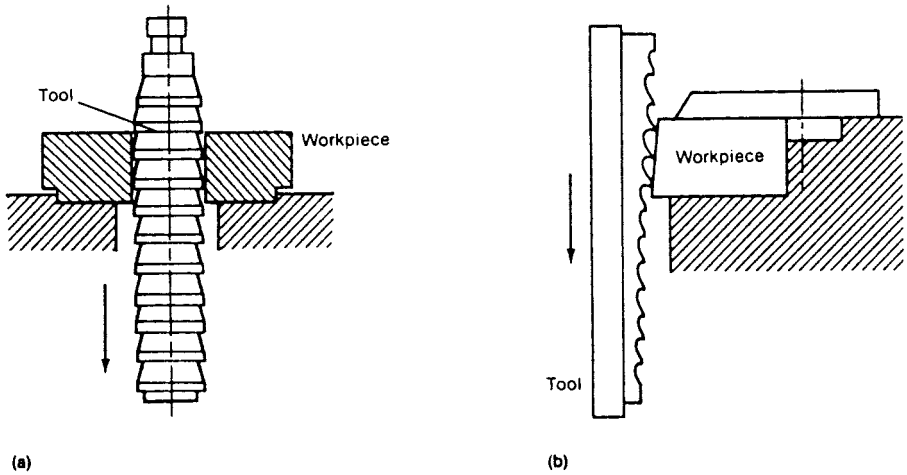


FIGURE 13.2.8 Internal (a) and external (b) broaching. (From *ASM Handbook, Machining*, vol. 16, 9th ed., ASM International, Metals Park, OH, 1989, 195. With permission.)

TABLE 13.2.11 Types of Broaches

Broaches	Description	Application
Solid	One-piece broach produced from tapered bar stock; repair of broken teeth is difficult	Parts which require high dimensional accuracy and concentricity
Shell	Multipiece broach consisting of a main body, an arbor section over which a removable shell fits, and a removable shell containing the cutting edges; worn or damaged sections can be replaced	Internal and selected external broaching; sacrifices some accuracy and concentricity as the tool is not as stiff as a solid broach
Insert-type	Effectively a tool holder with inserts to perform the actual cutting; inserts are typically made from HSS or carbides; worn or damaged inserts can be readily replaced	Broaching large, flat surfaces

Machine Tool and Machining Parameters. Length and depth of cut have the most influence on determining the required broaching tool length. For internal cutting operations, as the cut length increases, more chip storage capacity must be provided between the cutting edges for the same amount of tooth advance. Cutting fluids are useful in preventing the work metal from adhering to the broach, and thus result in higher-quality surface finishes and increased broach life.

The primary consideration in the selection of optimum broaching speed is the trade-off between speed and wear rate. In general, steels are broached at 10 to 30 sfm (3 to 9 m/min); the harder the steel, the slower the broach speed (ASM, 1989b).

Capabilities and Process Limitations. Broaching can maintain tight tolerances during long production runs since metal-cutting operations are distributed among the different roughing and finishing teeth. Also, broach teeth can be repeatedly sharpened, allowing cutting efficiency and accuracy to be maintained.

Broaching is an extremely fast, precise machining operation. It is applicable to many workpiece materials over a wide range of machinability, can be accomplished in seconds, is readily automated, and can easily be done manually. For example, for low-carbon steels, tolerances of 0.002 in. (0.05 mm) can be readily attained with a surface finish of 60 μin. (1.55 μm); if desired, tighter tolerances and surface finishes of 30 μin. (0.8 μm) are possible without much additional effort. For difficult-to-machine superalloys, tolerances of 0.001 in. (0.025 mm) and surface finishes of 30 μin. (0.8 μm) are commonly achieved in production (ASM, 1989b).

Broaching is rarely used for removing large amounts of material since the power required would be excessive. It is almost always more effective to use another machining method to remove the bulk of material and use broaching for finishing.

Since a broach moves forward in a straight line, all surface elements along the broach line must be parallel to the direction of travel. Consequently, the entire surface of a tapered hole cannot be broached. Also, cutting is done sequentially with the finishing teeth engaging last. Therefore, a blind hole can be broached only if a sufficiently long recess is provided to permit full travel of the broach.

The direction of travel of the broach cannot realistically be changed during a broaching stroke, except for rotating the tool. Thus, surfaces having compound curves cannot be broached in a single operation. On external surfaces, it is impossible to broach to a shoulder that is perpendicular to the direction of broach movement.

Grinding

Trevor D. Howes, John Webster, and Ioan Marinescu

Description and Applications. Grinding, or abrasive machining, refers to processes for removing material in the form of small chips by the mechanical action of irregularly shaped abrasive grains that are held in place by a bonding material on a moving wheel or abrasive belt (Green, 1992). In surface-finishing operations (e.g., lapping and honing) these grains are suspended in a slurry and then are embedded in a roll-on or reference surface to form the cutting tool. Although the methods of abrasion may vary, grinding and surface-finishing processes are used in manufacturing when the accuracy of workpiece dimensions and surface requirements are stringent and the material is too hard for conventional machining.

Grinding is also used in cutoff work and cleaning of rough surfaces, and some methods offer high material-removal rates suitable for shaping, an area in which milling traditionally has been used.

Grinding is applied mainly in metalworking because abrasive grains are harder than any metal and can shape the toughest of alloys. In addition, grinding wheels are available for machining plastics, glass, ceramics, and stone. Conventional precision metal and ceramic components and ultraprecision electronic and optical components are produced using grinding.

Mechanics of Grinding. Three types of energy are involved in grinding (Andrew et al., 1985). *Rubbing energy* is expended when the grains (cutting edges) of the grinding wheel wear down. As they wear,

they cut less and produce increasing friction, which consumes power but removes less material. *Plowing energy* is used when the abrasive does not remove all of the material but rather plows some of it aside plastically, leaving a groove behind. *Chip-formation energy* is consumed in removing material from the workpiece as the sharp abrasive grain cuts away the material (or chip) and pushes it ahead until the chip leaves the wheel.

The grinding wheel experiences *attritious wear* as the abrasive grains develop wear flats from rubbing on the workpiece, or when grains break free from the bond material. Attritious wear gives rise to rubbing energy resulting from friction, and thus can lead to thermal damage as power consumption increases without an increase in material removal rate. The wheel can wear through *fracture*, predominating at relatively high in-feed rates. In this case, pieces of the abrasive grain break free and expose a new, sharp surface.

Materials can be classified as either easy to grind or difficult to grind. For *easy-to-grind materials*, most of the power consumption becomes invested in chip formation; thus, rubbing and plowing energy are minimal. *Difficult-to-grind materials* involve considerable rubbing and plowing energy since the force required to remove chips is comparatively high.

Types of Grinding. In *surface grinding*, the grinding wheel traverses back and forth across the workpiece. Grinding can take place by using either the periphery or side face of the wheel. The table holding the part may also reciprocate. Surface grinding is done most commonly on flat surfaces and surfaces with shapes formed of parallel lines, e.g., slots.

Creep-feed grinding is a form of surface grinding in which the wheel feeds into the workpiece at a low rate (0.4 to 40 in./min, 10 to 1000 mm/min) while grinding at a large depth of cut (0.04 to 0.4 in., 1 to 10 mm, or deeper). A large amount of material can be removed with one pass of the wheel, compared with conventional surface grinding in which the wheel makes many quick passes over the workpiece at slight depths of cut. This process is limited by the large amount of heat generated at the grinding arc which can result in thermal damage (grinding "burn"). Application of coolant is critical in creep-feed operations. CBN wheels, with their good heat transfer property, can also reduce the severity of burn (King and Hahn, 1986).

Cylindrical grinding produces round workpieces, such as bearing rings, although some machines can also grind tapered parts. The workpiece is mounted to a spindle and rotates as the wheel grinds it. The workpiece spindle has it own drive motor so that the speed of rotation can be selected. Both inner surfaces (internal cylindrical grinding) and outer (external cylindrical grinding) can be worked, although usually the same machine cannot do both.

There are three variants for *external grinding*:

- *Plain grinding* — the wheel carriage is brought to the workpiece on its spindle and in-feeds until the desired dimensions are reached
- *Traverse grinding* — the rotating workpiece is mounted on a table that reciprocates under the wheel; the grinding wheel is stationary except for its downward feed into the workpiece
- *Plunge grinding* — the table with the rotating workpiece is locked while the wheel moves into the workpiece until the desired dimensions are attained

Centerless grinding is a form of cylindrical grinding. In this method workpieces are not held in a centering chuck but instead rotate freely between a support, regulating wheel, and the grinding wheel. The force of the rotating grinding wheel holds the workpiece against the support. The supports are usually stationary and so a flow of lubricant is required to reduce friction between workpiece and support. An example of centerless grinding is shown in Figure 13.2.9.

Abrasive belt machines use a flexible fabric coated with an abrasive stretched between two rollers, one of which is driven by a motor. Usually, the abrasive coating is aluminum oxide for steels and bronzes and silicon carbide for hard or brittle materials. In the metal industries, common use of such machines is for dry grinding of metal burrs and flash and polishing of surfaces. However, some fabrics permit use

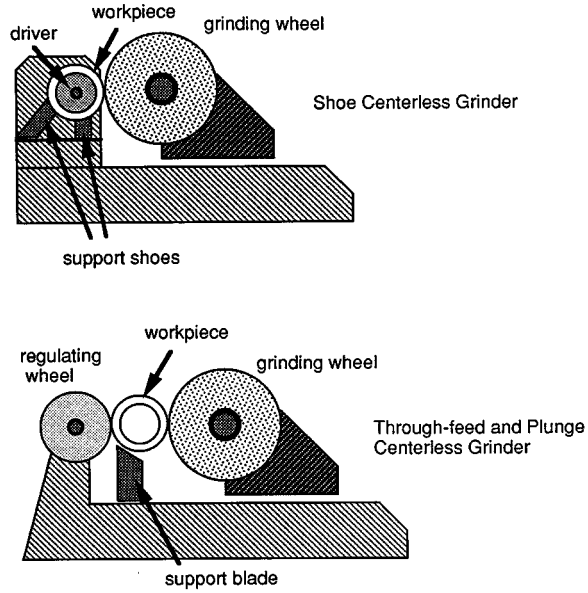


FIGURE 13.2.9 Through-feed and plunge centerless grinding. (Courtesy of T. D. Howes, John Webster, and I. Marinescu.)

of grinding fluids to enhance chip removal and provide cooling and lubrication, which results in better cutting action and longer belt life.

Honing, lapping, and polishing use abrasives to improve the accuracy of the form of a workpiece or the surface finish beyond the capabilities of grinding.

- *Honing* is a low-surface-speed operation, usually performed on an internal, cylindrical surface but possible also on external ones. Stock is removed by the shearing action of abrasive grains: a simultaneous rotary and reciprocating motion of fixed abrasive in the form of a stone or stick. Finishes range from under 1 to 50 μin. (0.025 to 1.3 μm). The development of CBN has revolutionized the honing process because this material easily outperforms conventional abrasives such as aluminum oxide, lasting up to 100 times longer.
- *Superfinishing*, like honing, uses fixed abrasives in the form of a stone. Unlike honing, which has a helical motion inside a bore, superfinishing uses high-speed, axial reciprocation combined with slow rotation of the outside diameter of the cylindrical component being processed. The geometry produced by a previous operation generally is not improved.
- *Lapping* is a fine-finishing, abrasive machining process used to obtain superior finish and dimensional accuracy. Lapping is unlike other finishing processes because some of the abrasive is loose rather than bonded. In general, lapping occurs when abrasive grains in a liquid vehicle (called a slurry) are guided across a workpiece by means of a rotating plate.
- *Polishing* uses free abrasive, as in lapping, but requires a soft support unlike the relatively hard support used in lapping. The total depth of cut during polishing can be as little as nanometers where chemical interactions will play a stronger role than mechanical or physical interactions. When the depth of cut is greater than 1 μin. (0.025 μm), the interactions are usually of a mechanical nature. Many industrial components, especially electronic and optical, required highly polished surfaces.

Key System Components. A *grinding wheel* consists of thousands of small, hard grains of abrasive material held on the surface in a matrix of bond material (Table 13.2.12). The bond material is matched to the characteristics of the grain to retain the grain sufficiently to maximize its use before shedding it.

TABLE 13.2.12 Common Grinding Wheel Abrasives

Abrasive	Characteristic	Grinding Application
Aluminum oxide	Friable	Steel: soft or hardened, plain or alloyed
Seeded gel aluminum oxide	More friable and expensive than aluminum oxide	Steel: soft or hardened, plain or alloyed; use at higher stock removal rates
CBN	Tough; increased life at higher speeds; high accuracy and finish	Hardened steels; tough superalloys
Synthetic diamond	Hardest of all abrasives; can be friable; seldom need dressing/truing	Grinding hardened tool steels, cemented carbides, ceramics, and glass; cutting and slicing of silicon and germanium wafers
Silicon carbide	Friable	Cast iron; nonferrous metals; nonmetallics

The structure of the wheel formed by specific types of grains and bonds determines its characteristics. The grains are spaced apart depending on the cutting required. Widely spaced grains (open structure) cut aggressively, which is useful for hard materials or high rates of material removal, but which tends to produce coarse finishes. Closely packed grains (dense structure) make fine and precise cuts for finish grinding.

Grain spacing is also important for temporary storage of chips of material removed from the workpiece. An open structure is best for storing chips between the grains, which are then released after wheel rotation moves the grains away from the workpiece. An open structure also permits more coolant to enter the spaces to dissipate heat.

The bonding material is important to grinding performance. This material is weaker than the cutting grains so that ideally, during grinding, the grain is shed from the wheel surface when it becomes dull, exposing new sharp grains. For instance, wheels with friable abrasives that fracture to expose new, sharp grains must retain the grains longer to maximize the use of the abrasive; these wheels use stronger bonding materials. The four types of bond material are vitrified, resinoid, rubber, and metal. Table 13.2.13 shows their properties and uses.

TABLE 13.2.13 Characteristics of Grinding Wheel Bonds

Bond Type	Characteristic	Application
Rubber	Relatively flexible	Wet cut-off wheels; high-finish work; regulating wheels for centerless grinding
Resinoid (thermoplastic)	Relatively flexible	Rough grinding; portable grinders
Vitrified (glasslike)	Endure high temperatures; resist chemical effects of coolants; sensitive to impacts	Most widely used of all bonding materials
Metal	Electrodeposited nickel or sintered metal powder often used to bond CBN and diamond abrasives; has long wheel life; electrically conductive	Aggressive cutting operations such as creep-feed and deep grinding; electrically conductive grinding (e.g., electrochemical methods)

Grinding wheels must be resharpened on occasion. *Dressing*, not always required, sharpens the grains before grinding. *Truing* operations ensure the wheel conforms to the required cutting shape and will rotate concentrically to its spindle.

Because grinding wheels have relatively high mass and high operating speed, they must be *precisely balanced*. Imbalance causes vibrations that reduce the quality of the workpiece, hasten the wear of the spindle and bearings of the machine, affect other devices mounted on the grinder, and possibly transmit vibration from the grinder through the shop floor to other machines. Mounting of the wheel on the spindle and subsequent wheel wear can degrade the balance of the rotating system. Wheels are balanced by moving counterweights on balancing flanges. Some machines have an automatic balancer that shifts internal counterbalance masses.

Coolants are usually sprayed on the grinding zone to cool the wheel and workpiece, lubricate the surface to reduce grinding power, and flush away the chips. Excessive heat can damage both the wheel and workpiece by inducing undesirable physical changes in materials, such as metallurgical phase changes or residual stresses, or softening of the bond material in the grinding wheel. Coolant application is especially important in creep-feed grinding where the wheel-to-workpiece contact arc is long, heat generation is high, and the chips produced and abrasive lost from the wheel must be flushed away.

Selecting the type of *coolant system* depends on many factors, including the grinding wheel speed, material removal rate, depth of cut, and wheel/workpiece materials. The type of fluids used in this system requires consideration of both physical and environmental issues. Use of oil fluids can favor the formation of preferred residual stress patterns and better surface finish, and these oils can be recycled for long periods. However, oils present health risks, potential for groundwater contamination, and fire risks (especially with high-sparking superalloys). Water-based fluids offer far fewer environmental problems. Disadvantages of water-based fluids lie in their limited life expectancy of 3 to 12 months. Also, the relatively low viscosity of a water-based fluid at high velocity promotes a dispersed jet which reduces cooling capacity.

Capabilities and Process Limitations. Surface grinding can be a cheaper, faster, and more precise method than milling and planing operations. For profiled shapes, the grinding wheel can be dressed with less cost and inconvenience than changing milling setups for different parts. Grinding can be used as a high-stock-removal process; for example, creep-feed grinding has a depth of cut more typical of milling operations (0.1 in., 2.54 mm, and deeper). Creep-feed grinding is used for machining materials that are too difficult to work by other machining methods.

High-speed grinding can be extremely efficient. CBN abrasive allows high rates of material removal because CBN transfers heat away from the grinding zone due to its relatively high thermal conductivity, and CBN does not react with steel.

Considerable effort has been expended on modeling and testing the thermal limitations of grinding (Malkin, 1989). Nearly all models depend on a fundamental model which depends on sliding contact theory. All models confirm the following guidelines for grinding with conventional abrasives when burn is a limitation: decrease wheel speed, increase workpiece speed, use softer-grade wheels.

Grinding operations can be limited by two types of vibration:

- *Forced vibration.* Typical causes are out-of-balance wheels, nonuniform wheels, couplings, belts, noise from hydraulic systems, bearing noise, and forces transmitted from the floor to the grinding machine.
- *Self-excited vibration.* Typical causes are wheel wear, workpiece surface error regeneration, wheel loading, and wear flats. Typical solutions include softer grinding wheels, flexible (for dampening) wheel structures, and stiffer machine structures.

References

Altan, T., Oh, S.I., and Gegel, H. 1983. *Metal Forming — Fundamentals and Applications*, ASM International, Metals Park, OH.

Andrew, C., Howes, T.D., and Pearce, T.R.A. 1985. *Creep Feed Grinding*, Holt, Rinehart, and Winston, London.

ASM. 1989a. Turning, in *Machining, ASM Handbook*, Vol. 16, 9th ed., ASM International, Metals Park, OH, 142–149.

ASM. 1989b. *Machining, ASM Handbook*, Vol. 16, 9th ed., ASM International, Metals Park, OH.

Bakerjian, R., Ed. 1992. *Design for Manufacturability*, Vol. VI, *Tool and Manufacturing Engineers Handbook*, 4th ed., Society of Manufacturing Engineers, Dearborn, MI.

DeVries, W.R. 1991. *Analysis of Material Removal Processes*, Springer-Verlag, New York.

Green, R.E., Ed. 1992. *Machinery's Handbook*, 24th ed., Industrial Press, New York.

Kalpakjian, S. 1991. *Manufacturing Processes for Engineering Materials*, Addison-Wesley, Reading, MA.
Kalpakjian, S. 1992. *Manufacturing Engineering and Technology*, Addison-Wesley, Reading, MA.
King, R.I. and Hahn, R.S., Eds. 1986. *Handbook of Modern Grinding Technology*, Chapman and Hall, New York.
Komanduri, R. and Samanta, S.K. 1989. Ceramics, in *Machining, ASM Handbook*, Vol. 16, 9th ed., ASM International, Metals Park, OH, 98–104.
Malkin, S. 1989. *Grinding Technology — Theory and Applications of Machining with Abrasives*, John Wiley, New York.
Shaw, M.C. 1984. *Metal Cutting Principles*, Oxford Science Publications. New York.
Welboum, D. 1970. *Machine-Tool Dynamics*, Cambridge University Press, New York.

Nontraditional Machining

K. P. Rajurkar and W. M. Wang

The processes described below are representative of the types most likely to be encountered. The references listed at the end of the section contain detailed information on these processes, plus those that are not described here.

Electrical Discharge Machining (EDM)

Description and Applications. The EDM process machines hard materials into complicated shapes with accurate dimensions. EDM requires an electrically conductive workpiece. Process performance is unaffected by the hardness, toughness, and strength of the material. However, process performance is a function of the melting temperature and thermal conductivity. EDM is currently widely used in aerospace, machinery, and die and mold industries.

There are two types of EDM processes:

- Die-sinking EDM uses a preshaped tool electrode to generate an inverted image of the tool on the workpiece; commonly used to generate complex-shaped cavities and to drill holes in different geometric shapes and sizes on hard and high-strength materials.
- Wire EDM (WEDM) uses a metal wire as the tool electrode; it can generate two- or three-dimensional shapes on the workpiece for making punch dies and other mechanical parts.

Principle of Operation. EDM removes workpiece materials by harnessing thermal energy produced by pulsed spark discharges across a gap between tool and workpiece. A spark discharge generates a very small plasma channel having a high energy density and a very high temperature (up to 10,000°C) that melts and evaporates a small amount of workpiece material. The spark discharges always occur at the highest electrical potential point that moves randomly over the machining gap during machining. With continuous discrete spark discharges, the workpiece material is uniformly removed around the tool electrode. The gap size in EDM is in the range of 400 μin. to 0.02 in. (0.01 to 0.5 mm), and is determined by the pulse peak voltage, the peak discharge current, and the type of dielectric fluid.

EDM Power System. The discharge energy during EDM is provided by a direct current pulse power generator. The EDM power system can be classified into RC, LC, RLC, and transistorized types. The transistorized EDM power systems provide square waveform pulses with the pulse on-time usually ranging from 1 to 2000 msec, peak voltage ranging from 40 to 400V, and peak discharge current ranging from 0.5 to 500 A. With the RC, LC, or RLC type power system, the discharge energy comes from a capacitor that is connected in parallel with the machining gap. As a result of the low impedance of plasma channel, the discharge duration is very short (less than 5 msec), and the discharge current is very high, up to 1000 A. The peak voltage is in the same range of transistorized power systems.

The transistorized power systems are usually used in die-sinking EDM operations because of their lower tool wear. Capacitive power systems are used for small hole drilling, machining of advanced materials, and micro-EDM because of higher material removal rate and better process stability. WEDM

power generator usually is a transistor-controlled capacitive power system that reduces the wire rupture risk. In this power system, the discharge frequency can be controlled by adjusting the on-time and off-time of the transistors that control the charging pulse for the capacitor connected in parallel with the machining gap.

Key System Components. The machining gap between tool and workpiece during EDM must be submerged in an electrically nonconductive *dielectric fluid*. In die-sinking EDM, kerosene is often used as a dielectric fluid because it provides lower tool wear, higher accuracy, and better surface quality. Deionized water is always used as a dielectric fluid in WEDM to provide a larger gap size and lower wire temperature in order to reduce the wire rupture risk. This fluid also serves to flush debris from the gap and thus helps maintain surface quality.

Copper and graphite are commonly used as die-sinking *EDM tool materials* because of the high electrical conductivity and high melting temperature and the ease of being fabricated into complicated shapes. The wire electrode for WEDM is usually made of copper, brass, or molybdenum in a diameter ranging from 0.01 to 0.5 mm. Stratified copper wire coated with zinc brass with diameter of 0.25 mm is often used.

In the traditional die-*sinking EDM process*, the tool is fabricated into a required shape and mounted on a ram that moves vertically. The spark discharges can only occur under a particular gap size that determines the strength of electric field to break down the dielectric. A servo control mechanism is equipped to monitor the gap voltage and to drive the machine ram moving up or down to obtain a dischargeable gap size and maintain continuous sparking. Because the average gap voltage is approximately proportional to the gap size, the servo system controls the ram position to keep the average gap voltage as close as possible to a preset voltage, known as the servo reference voltage.

In a WED machine, the wire electrode is held vertically by two wire guides located separately above and beneath the workpiece, with the wire traveling longitudinally during machining. The workpiece is usually mounted on an *x-y* table. The trajectory of the relative movement between wire and workpiece in the *x-y* coordinate space is controlled by a CNC servo system according to a preprogrammed cutting passage. The CNC servo system also adjusts the machining gap size in real time, similar to the die-sinking EDM operation. The dielectric fluid is sprayed from above and beneath the workpiece into the machining gap with two nozzles.

The power generators in WED machines usually are transistor-controlled RC or RLC systems that provide higher machining rate and larger gap size to reduce wire rupture risks. In some WED machines, the machining gap is submerged into the dielectric fluid to avoid wire vibration to obtain a better accuracy. The upper wire guide is also controlled by the CNC system in many WED machines. During machining, the upper wire guide and the *x-y* table simultaneously move along their own preprogrammed trajectories to produce a taper and/or twist surface on the workpiece.

Machining Parameters. The polarity of tool and workpiece in EDM is determined in accordance with the machining parameters. When the discharge duration is less than 20 μsec, more material is removed on the anode than that on the cathode. However, if the discharge duration is longer than 30 μsec, the material-removal rate on the cathode is higher than that on the anode. Therefore, with a transistorized power system, if the pulse on-time is longer than 30 μsec, the tool is connected as anode and the workpiece is connected as cathode. When the on-time is less than about 20 μsec, the polarity must be reversed. With an RC, LC, or RLC power system, since the discharge duration is always shorter than 20 μsec, the reversed polarity is used.

With transistorized EDM power systems, the machining rate and surface finish are primarily influenced by the peak current. The machining rate increases with the peak current. The relationship between the machining rate and pulse on-time is nonlinear, and an optimal pulse on-time exists. Reducing peak current improves the surface finish, but decreases the machining rate.

Capabilities and Process Limitations. Die-sinking ED machines with transistorized power systems under good gap-flushing conditions can attain a material-removal rate as high as 12 mm^3/min/amp (for

a steel workpiece). The wire cut EDM process can cut ferrous materials at a rate over 100 mm^2/min. A surface roughness value of 0.01 in. (0.2 mm) can be obtained with a very low discharge current. The tool wear ratio can be controlled within 1% during rough machining and semifinishing with the transistorized power generator. Dimensional tolerance of ±118 μin. (3 μm) and taper accuracy of 20 to 40 μin. (0.5 to 1 μm/mm) with both die-sinking and WEDM can be obtained.

EDM can machine materials having electrical conductivity of 10^{-2} W^{-1} cm^{-1} or higher. An average current density more than 4 A/cm^2 tends to cause substantial tool wear and unstable machining, and may lead to dielectric fire. This factor largely limits the productivity of EDM. During machining a deep cavity using die-sinking EDM under difficult flush condition, arc discharges occur, and the resultant thermal damage on workpiece substantially limits the productivity and the machined surface quality. Dielectric properties also impose additional constraints.

Electrical Chemical Machining (ECM)

Description and Applications. The ECM process uses the electrochemical anodic dissolution effect to remove workpiece material. Like die-sinking EDM, the tool electrode of ECM is preshaped according to the requirements of the workpiece. During machining, the inverted shape of the tool is gradually generated on the workpiece. ECM machines complex contours, irregular shapes, slots, and small, deep, and/or noncircular holes. Typical applications of ECM include machining nickel-based superalloy turbine blade dovetails, slots in superalloy turbine disks, engine castings, gun barrel rifles, and forging dies. ECM is also used for deburring, surface etching, and marking.

Principle of Operation. Electrolyte fluid is forced through the gap between tool and workpiece during ECM. A low-voltage and high-current DC power system supplies the machining energy to the gap. The tool electrode of ECM must be connected as cathode and the workpiece must be connected as anode. The electrochemical anodic dissolution phenomenon dissolves workpiece surface material into metal ions. The electrolyte fluid flushes the metal ions and removes heat energy generated by the deplating actions. The gap size in ECM is in the range of 0.004 to 0.04 in. (0.1 to 1 mm). ECM process performance is independent of the strength, hardness, and thermal behavior of workpiece and tool materials.

Key System Components. Copper, brass, stainless steel, and titanium are commonly used as *ECM tool electrode materials* due to their good electrical conductivity, resistance to chemical erosion, and ease of being machined into desired shapes. The geometric dimensions of the machined surface generated by ECM depend on the shape of tool and the gap size distribution.

The structure of an EC machine varies with the specific applications. An ECM must have a tool feed system to maintain the machining gap, a power system to supply the power energy, and a fluid-circulating system to supply the electrolyte and to flush the machining gap. The power system used in ECM is a DC power source with the voltage ranging from 8 to 30 V and a high current in the range of 50 to 50,000 A, depending on the specific design of the power system.

The *electrolyte* is the medium enabling the reaction of electrochemical dissolution occurring on the anode. The electrolyte can be classified into categories of aqueous and nonaqueous, organic and nonorganic, alkaline and neutral, mixed and nonmixed, and passivating and nonpassivating, and acidic. The electrolyte is selected according to the type of workpiece material, the desired accuracy, surface finish requirements, and the machining rate. Neutral salts are used in most cases. Acid electrolytes are used only for small hole drilling when the reaction products must be dissolved in the electrolyte.

Machining Parameters. ECM performance is mainly influenced by electrical parameters, electrolyte, and geometry of tool and workpiece. The electrical parameters include machining current, current density, and voltage. ECM systems with 50 to 50,000 A and 5 to 30 V DC are available, and the current density can be in the range of 10 to 500 A/cm^2. Key electrolyte parameters consist of flow rate, pressure, temperature, and concentration. Important parameters of tool and workpiece geometry are contour gradient, radii, flow path, and flow cross section. When the tool feed rate equals the machining rate, an equilibrium gap size is obtained; this is critical to maintaining shaping accuracy.

The electrolyte selection also plays an important role in ECM dimension control. In this regard, the sodium nitrate solution is preferable because the metal-removal rate at smaller gap size locations is higher than at other places. Therefore, the characteristics of current efficiency in ECM influence the uniformity of the gap size distribution. ECM accuracy has recently been shown to improve with pulsed voltage (instead of continuous voltage) and an appropriate set of pulse parameters (on-time, off-time, etc.).

Capabilities and Process Limitations. ECM is capable of machining any electrically conductive metallic material, and the process is generally unaffected by the hardness, strength, and thermal behaviors of materials. This process can be used to machine parts with low rigidity such as parts with thin walls. Machining rates of 2 to 2.5 cm^3/min/1000 A current, surface roughness of 4 to 50 μin. (0.1 to 1.2 μm), and accuracy of 400 μin. to 0.01 in. (10 to 300 μm) can be achieved with ECM. The available ECM equipment can machine a 0.04 to 80 in. (1 to 2000 mm) long workpiece. The typical energy consumption is 300 to 600 J/mm^3.

The machining rate and surface finish with ECM are much higher than that with EDM due to higher allowable current density and the molecular level of material removal. The machining accuracy, however, is substantially lower than EDM and is much more difficult to control.

The gap size distribution is influenced by many factors including the type of electrolyte, electrolyte flow rate and flow pattern, electrolyte temperature, current density of machining, etc. Therefore, the gap size distribution is not uniform in most cases and is difficult to determine analytically. The shape of machined surface by ECM will not be a perfect mirror image of the tool electrode. In order to achieve an acceptable accuracy, the tool shape must be modified by using trial-and-error method in test machining before machining of actual workpieces.

ECM cannot machine materials with electrical conductivity less than 10^3 W^{-1} cm^{-1}. ECM cannot produce very sharp corners (less than 800 μin., 0.02 mm, radius). The machining rate is limited by the electrolytic pressure (less than 5 MPa) and boiling point as well as the applied current (less than 50,000 A). The gap size which determines the final shape and accuracy is limited to 800 μin. to 0.004 in. (0.02 to 0.1 mm).

ECM generates a large amount of sludge and spent electrolyte. This waste requires significant processing before it can be safely disposed of.

Water and Abrasive Jet Machining

Description and Application. *Water jet machining* (WJM) and *abrasive water jet machining* (AWJM) are used in many applications. In the WJM process, relatively soft workpiece materials are cut by a high-velocity water jet; e.g., food, wood, paper, plastic, cloth, rubber, etc. The AWJM process uses the fine abrasive particles mixed in the water jet to machine harder workpiece materials. The AWJM is used for drilling, contour cutting, milling, and deburring operations on metal workpieces, as well as for producing cavities with controlled depths using multipass and non-through-cutting methods. The cutting path of the WJM and AWJM can be controlled by a CNC system according to a preprogrammed program.

Principle of Operation and Machine Structures. During WJM, the workpiece material is removed by the mechanical energy generated by the impact of a high-velocity water jet. In a water jet machine, a high-pressure pumping system increases the pressure of water in the pipe system, and the pressurized water is sprayed from a nozzle with a small diameter to generate a high-velocity water jet.

In the AWJM process, pressurized water is sprayed from an orifice in the nozzle body into a mixing chamber to generate a negative pressure that absorbs the abrasive particles (supplied by an abrasive feed hose) into the water jet. The water jet/abrasive grain mixture is then sprayed through a tungsten carbide nozzle. The abrasive grains in the high-velocity water jet provide small cutting edges that remove material. The relative distance between water jet nozzle and workpiece is controlled by a two- or three-dimensional CNC system. This process can be used to generate a complicated shape.

Process Parameters and Limitations. The key parameters are the water and/or abrasive flow velocity, abrasive grain size, and mixing tube (nozzle) length and diameter. The typical water flow velocity is in the range of 2000 to 3000 ft/sec (600 to 900 m/sec) as determined by the water pressure. The water pressure in WJM and AWJM is very high, up to 2.7×10^6 psi (400 MPa), and the nozzle diameter is in the range of 0.003 to 0.08 in. (0.8 to 2 mm). The abrasive flow rate is governed by the water flow rate and the mixing density, and can be controlled up to 10 g/sec. Abrasive particles are usually in the range of 60 to 150 mesh size. Increasing water jet flow velocity increases cutting depth. The taper error of cutting is determined by the traverse rate that describes the ratio between the material removal and the cutting depth and width. This parameter is influenced by the water and abrasive velocity and cutting speed. Proper selection of AWJ parameters is essential for the elimination of burrs, delaminations, and cracks.

Limitations of the process include stray cutting and surface waviness, high equipment costs, hazard from the rebounding abrasive, high noise levels, and short nozzle lifetimes due to wear and abrasion.

Ultrasonic Machining

Description and Applications. *Ultrasonic machining* (USM) is a process that uses the high velocity and alternating impact of abrasive particles on the workpiece to remove material. The abrasive particles are mixed in a slurry that fills the machining gap between the tool and workpiece. The alternating movement of abrasive particles is driven by the vibration of the frontal surface of tool at an ultrasonic frequency. The ultrasonic machining process can machine hard and brittle materials.

USM is often used for machining of cavities and drilling of holes on hard and brittle materials including hardened steels, glasses, ceramics, etc. Rotary ultrasonic machining (RUM) is a new application that uses a diamond grinding wheel as the tool for drilling, milling, and threading operations. During RUM, the tool is rotating at a high speed up to 5000 rpm and vibrating axially at ultrasonic frequency. This process is able to drill holes with diameter from 0.02 to 1.6 in. (0.5 to 40 mm) at depths up to 12 in. (300 mm). The material removal rate of 6 mm^3/sec can be obtained with the RUM process. The tolerance of ±300 μin. (±0.007 mm) can be easily achieved with both conventional and rotary ultrasonic processes.

Principle of Operation. In the USM process, the machining gap between tool and workpiece is filled with an abrasive slurry composed of an oil mixed with abrasive particles, with the frontal surface of the tool vibrating at ultrasonic frequency to provide the machining energy. The inverted shape of the tool is gradually generated on the workpiece. Material removal by the USM process is very complex. When the machining gap is small, the material may be removed as the frontal surface of the tool moves toward the workpiece, hitting an abrasive particle that impacts the workpiece surface. Material can also be removed by the impact of the abrasive particles when the machining gap is relatively large. In this case, the abrasive particles are accelerated by the pressure of slurry due to the ultrasonic vibration of the frontal surface of the tool. Also, ultrasonic-induced alternating pressure and cavitation in the slurry assist material removal.

Key System Components. The ultrasonic vibration in USM is generated by an *ultrasonic generator.* The ultrasonic generator consists of a signal generator, a transducer, and a concentrator. The signal generator produces an electrical signal whose voltage and/or current is changing at an ultrasonic frequency to drive the transducer. The frequency of the electrical signal can be adjusted in the range of 10 to 40 kHz.

The transducer converts the electrical voltage or current into the mechanical vibration. Two types of transducers are commonly used in USM: magnetostrictive and piezoelectric.

- The *magnetostrictive transducer* was extensively used prior to 1970. This transducer is constructed by surrounding a number of sheets of magnetostrictive material with a coil. When the strength of the electric current in the coil changes at an ultrasonic frequency, a mechanical ultrasonic vibration is generated in the magnetostrictive material. This transducer has a low energy conversion efficiency, usually less than 30%.

- The *piezoelectric ultrasonic transducer* is commonly used today. The geometrical dimensions of this transducer vary with the change in the applied electric field. A mechanical ultrasonic vibration is generated when the strength of the electric voltage applied across the transducer material changes at an ultrasonic frequency. This transducer has an extremely high energy conversion efficiency, up to 95%. The amplitude of the ultrasonic vibration generated directly by the transducer is very small, about 400 μin. (0.01 mm). A concentrator is used for amplifying the amplitude into a level that is acceptable for USM. The transducer is mounted on the larger end of the concentrator; the tool is mounted on the smaller end.

In the ultrasonic machine, the ultrasonic generator is held vertically on the ram that moves vertically, and the workpiece is mounted on an *x-y* table that determines the relative position between tool and workpiece. During machining, a force providing pressure between the tool and workpiece is added through the ram mechanism.

Process Parameters and Limitations. The material-removal rate during USM increases with an increase in the amplitude of ultrasonic vibration, grain size of the abrasive particles, and pressure between the tool and workpiece. The surface finish is essentially determined by the grain size for a given workpiece material; i.e., the smaller the grain size, the better surface finish. The abrasive grains used in USM are usually in the range of 100 to 900 mesh number.

The USM process is limited by the softness of the material. Workpiece materials softer than Rockwell C40 result in prohibitively long cycles. The best machining rate can be obtained on materials harder than Rockwell C60.

References

Benedict, G.F. 1987. *Nontraditional Manufacturing Processes*, Marcel Dekker, New York.

McGeough, J.A. 1988. *Advanced Methods of Machining*, Chapman and Hall, London.

Rajurakar, K.P. 1994. Nontraditional manufacturing processes, in *Handbook of Design, Manufacturing and Automation*, Dorf, R.C. and Kusiak, A., Eds., John Wiley & Sons, New York, 211–241.

Steen, W.M., 1991. *Laser Material Processing*, Springer-Verlag, New York.

Phase-Change Processes

Phase-change processes produce parts from materials originally in the liquid or vapor phase. These include processes such as metal casting and injection molding of polymers. The two most commonly used metal-casting processes are described below. The references listed at the end of the section contain detailed information on all phase-change unit processes.

Advantages and Applications of Metal Casting

Metal casting is one of the primary methods of producing bulk shapes. Very complex shapes can be cast from nearly every metal, making casting an extremely versatile process. Castings are made in sizes that range from fractions of an ounce to hundreds of tons.

The selection of the best molding and casting process for an application can be complex, and is governed by many factors which include casting size, variation in thickness of the casting sections, required mold strength, required surface finish and dimensional accuracy, production rates, environmental factors (e.g., reclamation of the sand and type of sand binder), and cost. Casting processes can be considered to fall into five categories (Kanicki, 1988):

- Conventional molding processes — green sand, shell, flaskless
- Precision molding and casting processes — investment, permanent mold, die casting
- Special molding and casting processes — vacuum molding, evaporative pattern casting, centrifugal casting

- Chemically bonded self-setting sand molding — no-bake, sodium silicate
- Innovative molding and casting processes — unbonded sand molding (Patz and Piwonka, 1988), rheocasting, squeeze casting, electroslag casting

This section discusses the most widely used conventional molding process, green sand, and the most widely used precision molding and casting process, investment casting. The reference list at the end of the section should be consulted for detailed information on all casting processes.

Casting Defects and Design Considerations

Properly designed and manufactured castings provide many advantages, and are competitive with other unit process methods. Success with casting processes, as with every process, requires design and process engineers knowledgeable regarding the advantages and limitations of casting processes so that appropriate design and process choices can be made that avoid or minimize the occurrence of defects.

Table 13.2.14 summarizes the most common defects that occur during casting and suggests design and process changes that can avoid or reduce the effect of the defects. However, the suggested mitigation strategies may introduce different casting defects or the evolution of other problems, so each change should be carefully evaluated with regard to the system as a whole.

Green Sand Casting Processes

Description and Applications. Sand-mold casting is adaptable to a very wide range of alloys, shapes, sizes, and production quantities. Hollow shapes can be produced in these castings through the use of cores. Sand-mold casting is by far the most common casting process used in industry; some estimates are that as many as 90% of industrial castings use the sand-mold casting process (O'Meara et al., 1988). Green sand molding is currently the most widely used of all sand casting methods, although dry sand methods are preferred for very large castings. "Green" sand refers to the fact that water is added to activate the clay binder. In dry sand molding, the moisture is removed prior to casting.

Green sand-mold casting involves mixing sand with a suitable clay binder (usually a bentonite clay) and other additives, and packing the sand mixture tightly around a pattern that is constructed from the part design, but the pattern is not an exact replica of the part since various dimensional allowances must be made to accommodate certain physical effects. After extracting the pattern from the sand mold, a cavity is left behind that corresponds to the shape of the pattern. Next, molten metal is poured into this cavity and solidifies into a cast replica of the desired part. After the casting cools, the sand is readily removed from the casting and may be reclaimed for further use.

Key System Components. A *mold pattern* is constructed from the casting design with suitable modifications. Depending on the complexity of the part, CAD/CAM programs are able to design patterns that require very little adjustment to achieve desired solidification control and dimensional accuracy in the resulting casting (Berry and Pehlke, 1988). The computation is complex and cannot be easily done for complex shapes. The principal adjustments that must be made to translate a part design to a mold design result from many considerations (ASM, 1988). One needs to

- Compensate for shrinkage of the sand during the drying/curing operations
- Compensate for expansion of the sand caused by the rapid introduction of the molten metal into the mold
- Compensate for contraction of the liquid metal during freezing
- Allow easy extraction of the pattern from the packed sand through a taper on the vertical sides of the pattern
- Add a gating network to allow molten metal to smoothly flow into the cavity
- Add risers (including size), as required in key locations to continue feeding molten metal into the solidifying casting
- Add provisions for core prints, as required, to anchor cores that produce internal cavities which could not be directly molded from the pattern

TABLE 13.2.14 Typical Casting Defects and Mitigation Strategy

Casting Defect	Description and Cause	Mitigation Strategy
Cold shuts	Appear as folds in the metal — occurs when two streams of cold molten metal meet and do not completely weld Possible causes: • Interruption in the pouring operation • Too slow a pouring rate • Improperly design gating	Pour as quickly as possible Design gating system to fill entire mold quickly without an interruption Preheat the mold Modify part design Avoid excessively long thin sections
Hot tears and cracks	Hot tears are cracklike defects that occur during solidification due to overstressing of the solidifying metal as thermal gradients develop Cracks occur during the cooldown of the casting after solidification is complete due to uneven contraction	Fill mold as quickly as possible Change gating system; e.g., use several smaller gates in place of one large gate Apply thermal management techniques within the mold (e.g., chills or exothermic material) to control solidification direction and rate Insulate the mold to reduce its cooling rate Modify casting design: • Avoid sharp transitions between thin and thick sections • Taper thin sections to facilitate establishment of appropriate solidification gradients • Strengthen the weak section with additional material, ribs, etc.
Inclusions	Presence of foreign material in the microstructure of the casting Typical sources: • Furnace slag • Mold and core material	Modify gating system to include a strainer core to filter out slag Avoid metal flow turbulence in the gating system that could cause erosion of the mold Improve hardness of the mold and core
Misruns	Incomplete filling of the mold cavity Causes: • Too low a pouring temperature • Too slow a pouring rate • Too low a mold temperature • High backpressure from gases combined with low mold permeability • Inadequate gating	Control mold and metal temperature Increase the pouring rate Increase the pouring pressure Modify gating system to direct metal to thinner and difficult-to-feed sections quicker
Porosity	Holes in the cast material Causes: • Dissolved or entrained gases in the liquid metal • Gas generation resulting from a reaction between molten metal and the mold material	Pour metal at lowest possible temperature Design gating system for rapid but uniform filling of the mold, providing an escape path for any gas that is generated Select a mold material with higher gas permeability
Microshrinkage	Liquid metal does not fill all the dendritic interstices, causing the appearance of solidification micro-shrinkage	Control direction of solidification • Design gating system to fill mold cavity so that solidification begins at the extremities and progresses toward the feed gate • Lower the mold temperature and increase the pouring temperature • Add risers, use exothermic toppings to maintain temperature longer • Control cooling rate using chills, insulators, etc. in selected portions of the mold

There are basically three types of molding methods which are categorized by the resulting hardness or density of the sand mold (O'Meara et al., 1988; Brown, 1988):

Low-density molding

- *Hand-ramming* is the oldest, slowest, and most variable method for packing sand around the pattern. It is rarely used for production, but can be employed for prototypes or very limited production runs.
- *Jolt* machines operate with the pattern mounted on a table which is attached to the top of an air piston. Typically, a flask is placed on the table with the pattern centered in the flask cavity. The flask is filled with sand. Compressed air lifts the piston, and then the air is released, allowing the entire assembly to fall with a sharp jolt. The sequence can be repeated. The sand is compacted by its own weight and is densest at the pattern plate.
- *Jolt-squeeze* molding machines employ the same pattern equipment as the jolt machines but after the jolting operation, the sand is squeezed by hydraulic pressure to improve the packing uniformity in the mold.

Medium-density molding

- *Rap-jolt machines* are improved versions of the low-pressure machines, capable of exerting higher pressure to compact the sand.
- *Sand slingers* direct sand into a mold from a rotating impeller. Sand-packing density is a function of the centrifugal velocity of the impeller. This method is particularly useful for large molds.

High-pressure molding

- *Pressure wave* methods allow the sand to gravity fill the mold. Then, the top of the mold is sealed and a high-pressure wave is created by a controlled explosion of a combustible gas or by the rapid release of air pressure.
- *Horizontal flaskless molding* utilizes a pattern carrier to support the top half (cope) and bottom half (drag) of the mold pattern. The cope and drag are spaced apart in the carrier, and the space is evacuated. The molding setup is depicted in Figure 13.2.10. Vents in the pattern cause sand to be drawn into the mold. When the mold is filled, it is tightly squeezed.

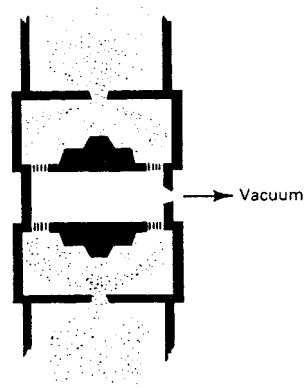


FIGURE 13.2.10 High-pressure vacuum-fill squeeze machine. (From *ASM Handbook, Machining*, Vol. 15, 9th ed., ASM International, Metals Park, OH, 1989, 343. With permission.)

The details of the *metal pouring* operation vary quite a bit depending on the metal, the material specification, the furnace type, and the foundry layout. A ladling method using gravity pressure is commonly used to transfer the molten metal, which may be alloyed, degassed, etc. immediately before it is discharged into the poring basin that feeds the gating system of the castings.

A *gating system* must allow complete fill of a mold cavity without causing flow turbulence that can entrap loose sand or slag and feed shrinkage as the liquid metal within the mold cools. The gating system should be designed to promote progressive solidification from the point most distant from the gate toward the gate.

A *riser* is a reservoir of molten metal that is attached to the casting. It feeds the voids that develop within the casting as the liquid metal cools and begins to solidify. A thin skin of frozen metal first forms in a shell around the outer part of the mold cavity immediately after the metal is poured. This rigid shell serves as a mold for the remainder of the casting. The volume lost by the shrinkage of the metal as it solidifies within this shell must be replaced from a liquid metal source, such as a riser or a feeding gate, to prevent internal porosity. Risers are subsequently removed from the casting.

Management of thermal gradients within the solidifying casting is essential to minimize shrinkage. The solidification of risers can be slowed by the use of an *exothermic material* placed on top of a riser. The heat generated by this material can allow the riser to continue to feed the casting until it is solidified. *Chills* can be used to reduce the local temperature at the mold-casting interface, and thus accelerate freezing in selected locations, thereby establishing the solidification direction. Chills are usually metal inserts strategically placed in the mold. The mold can also be *insulated* to reduce the overall cooling rate of the casting if necessary to reduce residual stresses, but the metallurgical effect must be carefully considered.

In most cases, *electric furnaces* are used for melting, and *pouring* is nonpressurized.

Influence of Process on Part Design. Properly designed gating systems and risers should produce completely solid castings. A number of design rules for the gating and risering system have been developed for casting various metals to achieve continuous feeding of the solidifying casting. These rules are embedded in computer programs that aid design of casting molds and layouts.

It may be less expensive and easier to change the design than to develop a complex thermal management system within the mold. Typical design changes to evaluate include

- Thickening thin sections that feed heavier, more remote sections
- Reducing the mass of the remote thick section
- Adding a riser to feed separately the remote thick section

Process Limitations. Each production step contributes to *dimensional variations* in a sand casting. For instance, dimensions are affected by the sand-packing density, by the process of withdrawing the pattern from the sand, by the moisture content of the sand, and by both the temperature of the molten metal and the speed with which it enters the mold cavity. The result is that sand casting is not inherently a precision process.

The surface finishes of sand castings are controllable only within rather wide limits (ASM, 1988). Normally, the maximum allowable surface roughness is specified, and any smoother surface is acceptable. For instance, casting steel in green sand can have a surface roughness varying from 500 to 2000 μin. (12 to 50 μm), and aluminum from 125 to 750 μin. (3 to 20 μm). These values can be improved in certain instances through the application of coatings on the sand mold.

Porosity cannot be prevented in all cases. Changes to the part design and postcasting processing, such as hot isostatic pressing (HIP), should be considered.

Investment Casting

Description and Applications. Investment castings are noted for their ability to reproduce extremely fine details and conform to very tight tolerances. As a result, these castings are used in critical, demanding structural applications, such as superalloy turbine airfoils and works of art.

The investment casting process employs a mold produced by enclosing an expendable pattern with a refractory slurry that is then hardened. The pattern, usually made from wax or plastic, is subsequently removed (e.g., by melting, dissolving, or burning), creating the desired mold cavity. The expendable

patterns are themselves cast in a permanent pattern die. Ceramic cores can create internal passages within the casting.

Key System Components. Shell investment and solid investment processes are used in the production of investment castings (Horton, 1988). The two processes differ in the method of mold preparation, not in pattern preparation or pattern assembly. In the *shell investment process*, the pattern assembly is precoated, dipped in a coating slurry, and covered with granulated refractory until the shell is built up to desired thickness, usually less than 0.5 in. (20 μmm); the thickness depends on the casting size and weight, cluster size, and type of ceramic and binder. As thin a shell as possible is specified to maximize mold permeability. Ceramic shell molds are used for the investment casting of carbon and alloy steels, stainless steels, heat-resistant alloys, and other alloys with melting points above 1100°C (2000°F).

The *ceramic material* used in shell investments is often silica, zircon (zirconium orthosilicate), an aluminum silicate, or alumina (Horton, 1988). *Silica glass* (fused silica) is desirable because it is readily available, but it has a high coefficient of thermal expansion and an abrupt phase transition, and it cannot be used in vacuum casting because the silica decomposes at low vapor pressures, leading to severe metal–mold reaction. *Zircon* is readily available, is resistant to wetting by molten metal, and has a high refractoriness. Its use is limited to prime coats though, because it is not available in large grain sizes. *Aluminum silicates*, such as mullite, can be manufactured to a range of pellet sizes and over a range of compositions. *Alumina* is more refractory than silica or mullite and is not very reactive with many metals.

The *binders* most often employed in shell investments are colloidal silica, ethyl silicate, and sodium silicate. *Colloidal silica* is an excellent general-purpose binder, and is the most widely used binder. Its primary disadvantage is that it is slow to dry. *Ethyl silicate* produces a bond between the refractory material that is very similar to that of colloidal silica. It dries much faster, but poses a fire hazard and is more expensive. Liquid *sodium silicate* forms a strong, glassy bond. The material is inexpensive but its refractoriness is poor, and the bond deteriorates in the presence of steam used to remove the wax pattern.

In the *solid investment process*, the pattern assembly is placed in a flask, which is filled with a refractory mold slurry. This slurry hardens in air, forming a solid mass in which the pattern assembly is encased. The types of bonding materials and refractories differ depending on the pouring temperature of the metal. For nonferrous alloys, pouring temperature is usually below 2000°F (1100°C). In these cases, *alpha gypsum* is commonly used as both the refractory and the binder, with other refractories such as silica added to improve mold permeability. The process is primarily used for making dental and jewelry castings.

For extremely limited production and for the development of production process parameters, investment mold patterns can be directly machined from an expendable material, such a plastic. For production-level investment casting, however, the patterns are produced by injecting wax or plastic into *permanent pattern molding dies*. The dimensional tolerance of these permanent dies must be closely controlled.

A mixture of paraffin and microcrystalline *wax* is widely used for making investment casting patterns (Horton, 1988). Waxes are strong, stiff, and provide adequate dimensional control during pattern making. They are easy to remove with pressurized saturated steam or elevated temperatures. *Plastic* patterns have several advantages compared with wax: higher strength, less subject to handling damage, can withstand automatic ejection from the pattern mold, and can reproduce thinner sections, finer definition, sharper corners, and better surface finish. But a major disadvantage is that certain plastics, such as polystyrene, expand during burnout and can crack ceramic shell molds.

Some investment castings require complex internal cavities (e.g., holes and air passages). For complicated shapes, the pattern-die *cores* are used to form portions of the pattern that cannot be withdrawn after the pattern is made. The cores must subsequently be dissolved or etched out from the casting.

As much *gating* as possible is included in the wax patterns. This allows use of standard methods of joining patterns together so that a number of investment castings can be produced during one pouring operation.

Investment casting is done in air and in vacuum. *Gravity pouring* fills the pouring basin from a ladle or directly from a furnace. This method requires low equipment investment, but highly skilled operators. In *pressure pouring* the molds are filled from the furnace with an assist from a pressurized gas to fill rapidly thin sections. With *vacuum-assisted pouring*, a vacuum pump evacuates air from a mold ahead of the stream of molten metal to minimize flow resistance. *Centrifugal casting* uses a spinning mold assembly to develop added pressure to fill the mold.

Capabilities and Process Limitations. Investment castings:

- Produce complex shapes that are difficult to make by other means
- Reproduce fine detail, high dimensional accuracy, and smooth surfaces requiring only minimal finishing
- Adapt to most metal alloys
- Allow control of metallurgical properties such as grain size and grain orientation

A tolerance of ±0.002 in. (0.05 mm) can be held on investment castings for each inch in its maximum dimension; however, a tolerance of ±0.005 in. (0.13 mm) is more typical (Horton, 1988). A surface finish of 125 min. (3 mm) can be readily achieved, and surface finishes as smooth as 30 to 40 min. (0.8 to 1.0 mm) can be produced with suitable process control.

The size and weight of castings that can be investment cast are usually limited by physical and economic considerations. Generally, the process can be applied cost-effectively to casting weighing up to 10 lb (4.5 kg); investment castings weighing 50 lb (22.5 kg) are not unusual, and castings as large as 1000 lb (450 kg) are feasible. The initial tooling costs of investment casting can be high.

References

ASM. 1988. in *Casting*, Vol. 15, *ASM Handbook*, 9th ed., ASM International, Metals Park, OH.

Berry, J.T. and Pehlke, R.D. 1988. Modeling of solidification heat transfer, in *Casting*, Vol. 15, *ASM Handbook*, 9th ed., ASM International, Metals Park, OH, 860ff.

Brown, R.B. 1988. Sand processing, in *Casting*, Vol. 15, *ASM Handbook*, 9th ed., ASM International, Metals Park, OH, 341–351.

Horton, R.A. 1988. Investment casting, in *Casting*, Vol. 15, *ASM Handbook*, 9th ed., ASM International, Metals Park, OH, 253–269.

Isayev, A.I. Ed., 1987. *Injection and Compressive Molding Fundamentals*, Marcel Dekker, New York.

Kanicki, D.P. 1988. Casting advantages, applications, and market size, in *Casting*, Vol. 15, *ASM Handbook*, 9th ed., ASM International, Metals Park, OH, 37–45.

O'Meara, P., Wile, L.E., Archibald, J.J., Smith, R.L., and Piwanka, T.S. 1988. Bonded sand molds, in *Casting*, Vol. 15, *ASM Handbook*, 9th ed., ASM International, Metals Park, OH, 222–230.

Patz, M. and Piwonka, T.S. 1988. Unbonded sand molds, in *Casting*, Vol. 15, *ASM Handbook*, 9th ed., ASM International, Metals Park, OH, 230–237.

Structure-Change Processes

Structure-change processes alter the microstructure of a workpiece. These changes can be achieved through thermal treatment involving heating and cooling (quenching) under controlled conditions, sometimes in combination with mechanical forces, in order to effect desired solid-state phase transformations. These processes include those that diffuse selected species into a surface layer to modify its composition or to create a thin layer of material that does not increase the dimensions of the workpiece.

The two structure-change processes described here are representative examples. *Normalization of steel* is a process that changes bulk properties of a workpiece. *Laser surface hardening* of steel only changes its surface properties; it does not affect the bulk properties. Even though both examples relate to ferrous

materials, structure-change processes are used to impart desired properties to many material systems; e.g., age hardening of aluminum alloys. The references listed at the end of this section contain detailed information on structure-change processes.

Normalizing Steel

Description and Applications. Normalizing is a heat-treating process that results in a relatively uniform steel microstructure. Essentially all the standard carbon steels can be normalized. The resulting phases and their size/distribution depend heavily on the carbon content of the steel. Normalization treatments are performed for a variety of reasons (Ruglic, 1991):

- Refine the dendritic grain structure remaining from casting
- Eliminate severe texture (and hence anisotropic properties) that results from forging and rolling operations
- Reduce residual stresses
- Improve the response of a steel to further processing, such as machining or surface hardening
- Improve mechanical properties by precipitating desirable phases

Principle of Operation. The steel workpiece must be heated sufficiently high to transform the entire structure to austenite, a face-centered cubic phase which essentially solutionizes all the carbon (at room temperature iron exists as ferrite, a body-centered cubic phase which has a very low carbon solubility). Diffusion-controlled solid-state phase transformations require time at temperature to occur. Therefore, the workpiece must be held at temperature long enough for austenite to dissolve the carbon.

The workpiece is then cooled slowly enough to avoid trapping the carbon in a supersaturated solution as the iron transforms back from austenite to ferrite. A time–temperature-transformation curve depicts which phase transformations will occur for different cooling gradients. A typical T-T-T curve is shown in Figure 13.2.11, which depicts the difference between a normalizing cool rate and that for annealing. For the normalization treatment to be successful, the regions in the microstructure with a carbon content of 0.8% carbon precipitate fine lamellae of ferrite and iron carbide (Fe_3C) on cooling, known as pearlite. Those areas low in carbon content should precipitate ferrite grains during the initial phase of the cooling cycle, followed by pearlite precipitation. The regions high in carbon should precipitate iron carbide in the austenite grain boundaries, followed by pearlite precipitation (Ruglic, 1990).

The end result of normalization is a microstructure, and hence the mechanical properties characteristic of the composition of the steel (primarily governed by carbon content) as opposed to a microstructure that was principally shaped by its previous thermomechanical processing.

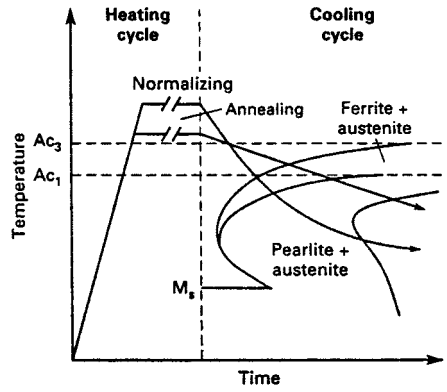


FIGURE 13.2.11 Time–temperature-transformation curve for normalizing compared to annealing. (From *ASM Handbook, Heat Treating*, Vol. 4, 10th ed., ASM International, Metals Park, OH, 1991, 35. With permission.)

Key System Components. Conventional heat-treating furnaces are used for normalization, such as batch or continuous furnaces (Ruglic, 1991). The rate of heating is not critical. The furnaces must be able to heat the component to about 100°F (55°C) into the austenitizing region — this temperature depends on the composition of the steel.

Control of the cooling rate is critical. In the usual case, a workpiece can be removed from the furnace and allowed to air cool uniformly until the diffusion-controlled phase transformations are completed.

Process Limitations. The ability to normalize the microstructure of steel is governed by thermodynamics. Phase stability and kinetics of the phase transformations are crucial. But heat transfer considerations also play a major role. For instance, it may not be possible to achieve the desired cooling gradient in the center of a thick section, and hence the properties of such sections will not be uniform.

For complex shapes or for those workpieces having a high degree of residual stresses, some distortion will occur during the normalization process. The extent of distortion can be reduced by appropriate fixturing, but it cannot be eliminated. Hence, some tolerance allowance must be made.

Laser Surface Hardening

Description and Applications. Laser surface hardening is used as an alternative to flame hardening and induction hardening ferrous materials. The rapid heating rate achievable by the laser minimizes part distortion and can impart surface hardness to low-carbon steels. The ability to locate the laser some distance from the workpiece can also be advantageous. The entire operation can be performed in air. This process is used to harden selected areas of machine components, such as gears, cylinders, bearings, and shafts.

Laser surface hardening imparts wear resistance and strength to the surface of a component without affecting its overall dimensions or changing its bulk properties. It is applied to selected areas which can be accessed by a laser beam. The process relies on rapid laser heating, followed by rapid quenching, to effect the necessary degree of hardening through phase transformation. The result is a very fine grain structure that is extremely hard. Typical case depth is a function of the composition of the ferrous material, but it will usually not exceed 0.1 in. (0.25 cm). For low- and medium-carbon steels, the case depth will range from 0.01 to 0.05 in. (0.03 to 0.13 cm), with the case depth increasing as the carbon content increases (Sandven, 1991).

Principle of Operation. An industrial laser rapidly heats a thin surface layer into the austenite phase region (austenite is a face-centered cubic allotropic phase of iron that has a high solubility for carbon). The interior of the workpiece is unaffected. When the laser beam is moved, the heated surface quickly cools. Consequently, the carbon does not have time to diffuse as the iron attempts to transform back to its ferrite (body-centered cubic) structure. The resulting microstructure is extremely hard since the trapped carbon atoms distort the iron crystal structure into a highly strained body-centered tetragonal form, known as martensite.

Key System Components. The majority of *industrial metalworking lasers* are either solid-state Nd:YAG or carbon dioxide type. Either pulsed or continuous mode can be used for surface treatment. The power output range for YAG lasers is 50 to 500 W. Carbon dioxide lasers are available in much higher power levels, up to 25 kW.

The surface to be hardened is usually *coated* to improve its ability to absorb laser radiation. A typical coating is manganese phosphate. *Paints* containing graphite, silicon, and carbon are also used. These coatings/paints can increase the absorption of laser energy to 80 to 90% (Sandven, 1991).

The output beam of the laser must be shaped and directed by an *optical system* to generate a laser spot of desired shape and size at the correct location on the workpiece surface. Reflective optical components are used since they are sturdy and easily adapted to an industrial environment.

Process Parameters. Many factors affect the end result of laser surface hardening. Important is the hardenability of the workpiece material which is affected by its composition and prior thermomechanical

history. For the laser process, the key parameters are beam power density, uniformity of the beam, and processing speed. Following are some general processing guidelines (Sandven, 1991).

- The range of usable power densities for laser surface hardening is 3200 W/in.^{2} (500 W/cm^2) to 32,000 W/in.^{2} (5000 W/cm^2) with beam dwell times ranging from 0.1 to 10 sec; higher power levels would melt the surface.
- Alloys with high hardenability can be processed at low speed with low power density to produce relatively thick cases.
- Alloys with low hardenability should be processed at high speed with high power density; the result is a shallow case.
- Beam configuration can be rectangular, square, or round; uniform energy density within the beam is very important.
- Maximum achievable surface temperature is proportional to the square root of the processing speed; thus doubling the beam power density requires the processing speed to be increased by a factor of four to maintain the equivalent maximum surface temperature.
- Smaller workpieces are not as effective a heat sink as larger workpieces, and hence self-quenching may have to be assisted by quenching media.

Process Limitations. The depth of hardness that can be practically achieved is limited by the surface melting point. Because of the high beam energy density, heat flow on complex-shaped surfaces, particularly those involving sharp corners or edges, can cause unexpected surface melting. Therefore, power density and process conditions must be carefully controlled.

It may be necessary to overlap passes of the laser beam, such as at the end of a complete pass around a cylinder. As a result, some tempering of the area already hardened occurs. The slower the processing speed, the greater the degree of tempering.

References

ASM. 1995. *Heat Treater's Guide: Practices and Procedures for Irons and Steels*, 2nd ed., ASM International, Metals Park, OH.

Boyer, H.E. 1982. *Practical Heat Treating*, ASM International, Metals Park, OH.

Brooks, C.R. 1982. *Heat Treatment, Structure and Properties of Nonferrous Alloys*, ASM International, Metals Park, OH.

Ruglic, T. 1991. Normalizing of steel, in *Heat Treating, ASM Handbook*, Vol. 4, 10th ed., ASM International, Metals Park, OH, 35–42.

Sandven, O.A. 1991. Laser surface hardening, in *Heat Treating, ASM Handbook*, Vol. 4, 10th ed., ASM International, Metals Park, OH, 286–296.

Sudarshan, T.S., Ed. 1989. *Surface Modification Technologies*, Marcel Dekker, New York.

Deformation Processes

Deformation processes change the shape of an object by forcing material to flow plastically from one shape into another shape without changing mass or composition (Table 13.2.15). The initial shape is usually simple. This shape is plastically deformed between tools or dies to obtain the final desired geometry, properties, and tolerances. A sequence of such processes is generally used to progressively form material. Deformation processes, along with casting and machining, have been the backbone of modern mass production.

In addition to shape change, forming processes alter the microstructure of the workpiece and can improve material properties. Deformation processes are normally considered when (Semiatin, 1988):

- Part geometry is moderately complex
- Component properties and structural integrity are important

TABLE 13.2.15 Significant Factors in Modeling a Deformation Process

Process Component	Characteristics
Input material	Flow stress; workability; surface condition
Output material	Geometry; mechanical properties; dimensional accuracy and tolerances; surface finish
Deformation zone	Deformation mechanics; stress state; temperature
Tooling	Material and geometry; surface conditions; temperature
Tool/material interface	Friction and lubrication; heat transfer
Process equipment	Speed and production rate; power range; precision

- Sufficient production volume can amortize tooling costs

Deformation processes can be classified (Semiatin, 1988) as bulk forming processes and sheet forming processes.

Bulk forming processes (e.g., rolling, extrusion, and forging) are characterized by

- Input material form is a billet, rod, or slab
- Workpiece undergoes a significant change in cross section during forming

Sheet metal forming processes (e.g., stretching, flanging, and drawing) are characterized by

- Input material is a sheet blank
- Workpiece is deformed into a complex three-dimensional form without appreciably changing the cross section

The key to attaining desired shape and properties is controlling metal deformation (Altan et al., 1983). The direction of the metal flow, the magnitude of the deformation, the rate of deformation, and the processing temperatures greatly affect the properties of the formed part. Design of the end product and the required deformation process consists of these steps:

- Predict metal flow by analyzing kinematic relationships (e.g., shape, velocities, strain rates, and strains) between the deformed and undeformed part configurations
- Establish producibility limits
- Select the process equipment and tooling capable of operating within the producibility limits

A bulk forming process (forging) and a sheet forming process (bending) are described below as representative examples of deformation processes. The references listed at the end of the section should be consulted for further information on these unit processes.

Die Forging

Talyan Altan

Description and Applications. Forging involves the controlled plastic deformation of metals into useful shapes (ASM, 1988c). Deformation may be accomplished by means of pressure, impact blows, or a combination. In order to reduce the flow stress, forging is usually accomplished at an elevated temperature. Forging refines the microstructure of a metal and can improve its mechanical properties, especially in preferred directions. Forging can also be used for other purposes, such as to consolidate powder preforms by welding grains, eliminate porosity in castings, break up long inclusions in forgings, and demolish the dendritic structure resulting from casting (Altan, 1988a).

Forgings are generally considered when strength, reliability, fracture toughness, and fatigue resistance are important. Forgings are used in critical, high-load applications, such as connecting rods, crankshafts, transmission shafts and gears, wheel spindles, and axles. Military and commercial aircraft are major users of forgings for numerous critical items, such as bulkheads, beams, shafts, landing gear cylinders

and struts, wheels, wing spars, and engine mounts. Similarly, jet engines depend on forgings for disks, blades, manifolds, and rings.

Types of Forging Processes. There are two broad categories of forging processes: closed-die forging and open-die forging. *Closed-die forging*, also known as *impression die forging*, employs precision-machined, matching die blocks to forge material to close dimensional tolerances. Large production runs are generally required to justify these expensive dies. During forging, the die cavity must be completely filled. In order to ensure this, a slight excess of material is forged. Consequently, as the dies close, the excess metal squirts out of the cavity in a thin ribbon of metal, called flashing, which must be trimmed.

Isothermal forging in heated superalloy dies minimizes the die quenching effect, preventing the rapid cooling of the workpiece in cold dies. This allows complete die fill and the achievement of close dimensional tolerances for difficult to process materials, such as superalloys.

Open-die forgings are the least refined in shape, being made with little or no tooling (Klare, 1988). These forgings are large, relatively simple shapes that are formed between simple dies in a large hydraulic press or power hammer. Examples are ship propeller shafts, rings, gun tubes, and pressure vessels. Since the workpiece is always larger than the tool, deformation is confined to a small portion of the workpiece at any point in time. The chief deformation mode is compression, accompanied by considerable spreading in the lateral directions.

Key System Components. There are two major classes of forging equipment as determined by their principle of operation: forging hammer, or drop hammer, which delivers rapid impact blows to the surface of the metal, and forging press, which subjects the metal to controlled compressive force. Each of these classes of forging equipment needs to be examined with respect to load and energy characteristics, its time-dependent characteristics, and its capability for producing parts to dimension with high accuracy.

Forging hammers generate force through a falling weight or ram (Altan, 1988b). These machines are energy restricted since the deformation results from dissipating the kinetic energy of the ram. The forging hammer is an inexpensive way to generate high forging loads. It also provides the shortest contact time under pressure, ranging from 1 to 10 msec. Hammers generally do not provide the forging accuracy obtainable in presses.

Forging presses are either mechanical or hydraulic (Altan, 1988c). *Mechanical forging presses* are stroke-restricted machines since the length of the press stroke and the available load at various positions of the stroke represent their capacity. Most mechanical presses utilize an eccentric crank to translate rotary motion into reciprocating linear motion of the press slide. The blow of the press is more like a squeeze than an impact of a hammer. Because of this, dies can be less massive and die life is longer than with a hammer. *Hydraulic presses* are load-restricted machines in which hydraulic pressure actuates a piston that squeezes the die blocks together. Full press load is available at any point during the stroke of the ram. A hydraulic press is relatively slow, resulting in longer process time; this may cause undesirable heat loss and die deterioration.

Design Considerations. Preform design is the most difficult and critical step in forging design. Proper preform design assures defect-free flow, complete die fill, and minimum flash loss. Although metal flow consists only of two basic types, extrusion (flow parallel to die motion) and upsetting (flow perpendicular to the direction of die motion), in most forgings both types of flow occur simultaneously, leading to a very complex flow field. An important step in understanding metal flow is to identify the neutral surfaces. Metal flows away from the neutral surface in a direction perpendicular to the die motion. Ideally, flow in the finishing step should be lateral toward the die cavity without additional shear at the die–workpiece interface. This type of flow minimizes forging load and die wear. A milestone in metalworking is the use of CAD in establishing the proper design for preforming and finishing dies in closed-die forging (Gegel and Malas, 1988). Figure 13.2.12 illustrates the relationships between forging process variables and those of a forging press that must be understood in order to estimate process performance for a hot-forging operation.

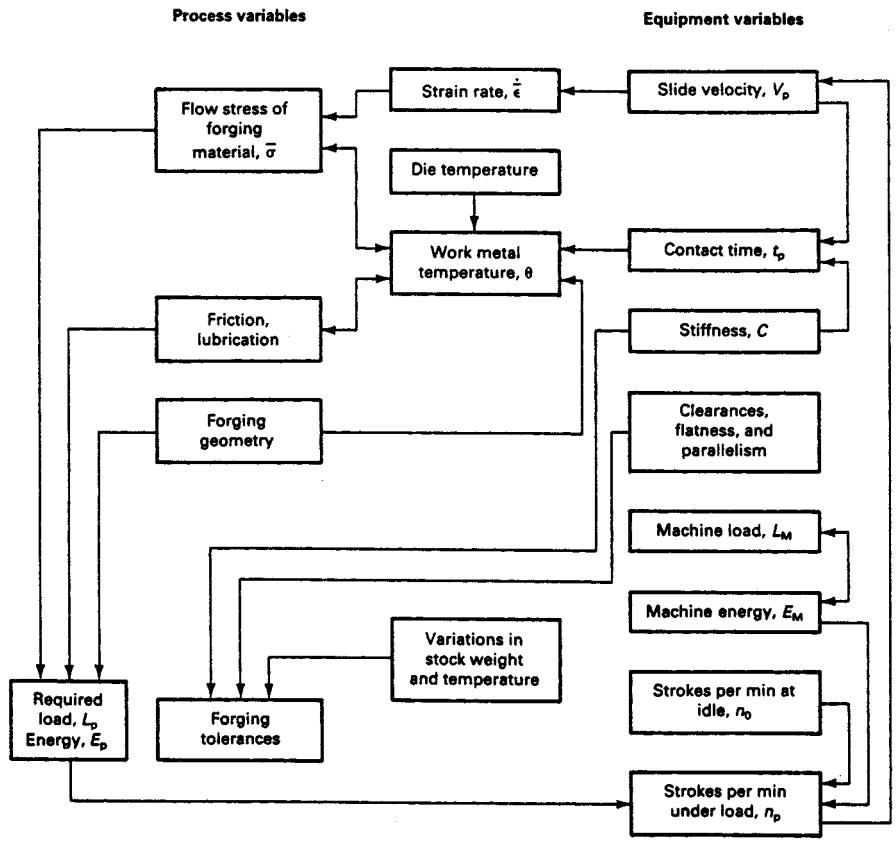


FIGURE 13.2.12 Relationships between process and machine variables in hot-forging processes conducted in presses. (From *ASM Handbook, Forming and Forging*, Vol. 14, 9th ed., ASM International, Metals Park, OH, 1988, 36. With permission.)

Designing a mechanical component that is to be made by forging together with the optimum geometry for the forging dies requires analysis of many factors (Altan, 1988a), including

- Design rules
- Workpiece material specification, and its critical temperatures
- Flow stress of the material at the process conditions (e.g., temperature, temperature gradient, strain rate, total strain)
- Workpiece volume and weight
- Frictional conditions in the die
- Flash dimensions
- Number of preforming steps and their configuration (flow field for the material)
- Load and energy requirements for each forging operation
- Equipment capability

In closed-die forging, it is particularly difficult to produce parts with sharp fillets, wide thin webs, and high ribs. Moreover, dies must be tapered to facilitate removal of the finished piece; draft allowance is approximately 5° for steel forgings (see Table 13.2.16).

TABLE 13.2.16 Typical Forging Defects and Mitigation Strategies

Defect	Description and Cause	Mitigation Strategy
Surface cracking	Fine cracks in the surface of the forging Possible causes: • Excessive working of the surface at too low a temperature • Brittle or low melting phases in the grain boundaries • Cracking at the die parting line	Increase the amount of preheating of the forging billet and forging die Change material specification Change furnace atmosphere to avoid diffusion of unwanted elements Increase the flash thickness Relocate the die parting line to a less critical location Stress relieve the forging prior to flash removal
Cold shut	Appears as a fold; occurs when two surfaces of metal fold against each other without welding Possible causes: • Poor metal flow in the die • Excessive chilling during forging • Poor die lubrication	Redesign forging die and/or forging preform to improve plastic flow of the metal during the forging operation Redesign the forging to avoid areas which are difficult to fill Increase the amount of preheating of the forging billet and forging die Improve die lubrication
Underfill	Incomplete forging in which all details are not produced Possible causes: • Debris residue in die • Scale on forging billet • Billet too small to completely fill die	Clean die thoroughly Completely descale the billet Redesign preform
Internal cracks	Cracks not visible from the surface, but detected during inspection and/or exposed during metal removal Possible causes: • Scale embedded in the internal structure of the forging • High residual tensile stresses	Completely descale the billet For open-die forgings, use concave dies Redesign for closed-die forging Use a hydraulic press and heated dies to avoid formation of excessive tensile stresses during forging

Press-Brake Forming

Description and Applications. Press-brake forming is a process used for bending sheet metal; the workpiece is placed over an open die and then pressed into the die by a punch that is actuated by a ram known as a press brake (ASM, 1988b). The main advantages of press brakes are versatility, ease and speed with which new setups can be made, and low tooling costs.

Press-brake forming is widely used for producing shapes from ferrous and nonferrous metal sheet and plate. Although sheet or plate 0.250 in. (10 mm) thick or less is commonly formed, metals up to 1 in. (25 mm) thick are regularly formed in a press brake. The length of a sheet is limited only by the size of the press brake. Forming can be done at room or elevated temperature. Low-carbon steels, high-strength low-alloy steels, stainless steels, aluminum alloys, and copper alloys are commonly formed in a press brake. Press-brake forming is applicable to any metal that can be formed by other methods, such as press forming and roll forming.

Press-brake forming is considered for bending sheet metal parts when the production quantities are small, dimensional control is not critical, or the parts are relatively long. In contrast, press forming would be considered when production quantities are large, tolerances are tight, or parts are relatively small. Contour roll forming would be another option for high-rate production applications (ASM, 1988c).

Principle of Operation. Bending is a method of forming sheet metal by stressing a material beyond its yield strength while remaining below its ultimate strength so that cracking is avoided. In press-brake forming the tooling and setup are relatively simple. A workpiece is placed over a die, typically having a V-shape. The bend angle is determined by the distance the workpiece is pressed into the die by the

punch. The width of the die opening (top of the V) affects the force needed to bend the workpiece. The minimum width is determined by the thickness of the workpiece and the radius of the punch nose.

Key System Components. A *press brake* is basically a slow-speed punch press that has a long, relatively narrow bed and a ram mounted between end housings. Rams are mechanically or hydraulically actuated. Figure 13.2.13 depicts a typical setup for press-brake forming.

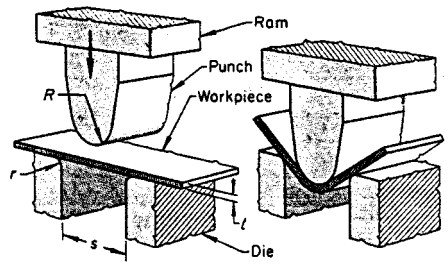


FIGURE 13.2.13 Typical setup for press-brake forming. (From *ASM Handbook, Forming and Forging*, Vol. 14, 9th ed., ASM International, Metals Park, OH, 1988, 533. With permission.)

V-bending dies and their corresponding punches are the tools most commonly used in press-brake forming. The width of the die opening is usually a minimum of eight times the sheet thickness.

Process Parameters. Capacities of commercial press brakes range from 8 to 2500 tons. Required capacity is governed by the size and bending characteristics of the work metal and by the type of bend to be made (ASM, 1988b).

The nose radius of the punch should not be less than the work-metal thickness for bending low-carbon steel, and must be increased as the formability of the workpiece material decreases. The radius of the V-bending die must be greater than the nose radius of the punch by an amount at least equal to the workpiece thickness to allow for the bottoming of the punch in the V.

It is preferable to orient a bend so that it is made across the rolling direction rather than parallel to it. Sharper bends can be made across the rolling direction without increasing the probability of cracking the material. If bends must be made in two or more directions, the workpiece should be oriented on the sheet layout such that none of the bends will be parallel to the rolling direction.

Springback after press-brake bending is considered only when close dimensional control is needed. It can be readily compensated for by overbending. Factors that affect springback include the mechanical properties of the work material, the ratio of the bend radius to stock thickness, the angle of bend, the method of bending, and the amount of compression in the bend zone. A greater amount of overbending is needed to correct for springback on small bend angles than on large bend angles.

Capabilities. The generally accepted tolerance for dimensions resulting from bending of metal sheet in the press brake is ±0.016 in. (±0.4 mm) up to and including 0.125 in. (3 mm) thickness (ASM, 1988b). For heavier gauges, the tolerance must be increased accordingly. Achievable tolerances are influenced by the part design, stock tolerances, sheet metal blank preparation, the condition of the machine and its tooling, and operator skill.

Design Factors. In press-brake forming, as in other forming processes, the metal on the inside portion of the bend is compressed or shrunk, and the metal on the outside portion is stretched. This results in a strain gradient across the thickness of the workpiece in the area of the bend with tensile strain on the outside and compressive strain inside. These residual strains (and resulting stresses) can lead to distortion of the part under loading conditions, heating, or cooling.

The formability of metals decreases as the yield strength approaches the ultimate strength. In press-brake forming, as the yield strength of the work metal increases, power requirements and springback problems also increase, and the degree of bending that is practical decreases.

There are several factors which will make it difficult to establish or maintain accurate placement of a bend line in a press brake. Corrective action may require a design change or change in processing sequence.

- Bends or holes that are located in close proximity to the required bend line can cause the position of the bend line to wander.
- Notches and cutouts located directly on the bend line make it difficult to maintain an accurate bend location.
- Offset bends will shift location unless the distance between bends in the offset is at least six times the thickness of the workpiece material.

If multiple bends must be made on a workpiece, it may not be possible to avoid a bend that is parallel to the rolling direction. Depending on the degree of texture in the sheet and the anisotropy of the material, a change to a higher-strength material may be necessary to achieve the desired geometry.

References

Altan, T. 1988a. Selection of forging equipment, in *Forming and Forging, ASM Handbook*, Vol. 14, 9th ed., ASM International, Metals Park, OH, 36–42.

Altan, T. 1988b. Hammers and presses for forging, in *Forming and Forging, ASM Handbook*, Vol. 14, 9th ed., ASM International, Metals Park, OH, 25–35.

Altan, T., Oh, S.I., and Gegel, H.L. 1983. *Metal Forming: Fundamentals and Applications*, ASM International, Metals Park, OH.

ASM. 1988a. Closed-die forging in hammers and presses, in *Forming and Forging, ASM Handbook*, Vol. 14, 9th ed., ASM International, Metals Park, OH, 75–80.

ASM. 1988b. Press-brake forming, in *Forming and Forging, ASM Handbook*, Vol. 14, 9th ed., ASM International, Metals Park, OH, 533–545.

ASM. 1988c. *Forming and Forging, ASM Handbook*, Vol. 14, 9th ed., ASM International, Metals Park, OH.

Gegel, H.L. and Malas, J.C. 1988. Introduction to computer-aided process design for bulk forming, in *Forming and Forging, ASM Handbook*, Vol. 14, 9th ed., ASM International, Metals Park, OH, 407ff.

Klare, A.K. 1988. Open-die forging, in *Forming and Forging, ASM Handbook*, Vol. 14, 9th ed., ASM International, Metals Park, OH, 61–74.

Kobayashi, S., Oh, S., and Altan, T. 1989. *Metalforming and Finite Element Methods*, Oxford Press, New York.

Lascoe, O.D. 1988. *Handbook of Fabrication Processes*, ASM International, Metals Park, OH.

Muccio, E.A. 1991. *Plastic Part Technology*, ASM International, Metals Park, OH.

Semiatin, S.L. 1988. Introduction to forming and forging processes, in *Forming and Forging, ASM Handbook*, Vol. 14, 9th ed., ASM International, Metals Park, OH, 17–21.

Consolidation Processes

Consolidation processes fuse smaller objects such as particles, filaments, or solid sections into a single solid part or component to achieve desired geometry, structure, and/or property. These processes use either mechanical, chemical, or thermal energy to bond the objects. Interaction between the material and the energy that produces the consolidation is a key feature of the process.

Consolidation processes are employed throughout manufacturing, from the initial production of the raw materials to final assembly. One group of processes involves the production of parts from powders of metals, ceramics, or composite mixtures. The resultant consolidated products are typically semifinished and require further processing. For instance, the consolidation of powders produces bar, rod, wire, plate, or sheet for upstream processes.

The consolidation of net shape composite structures (i.e., require minimal finishing work) is an increasingly important area. The design of the structural geometry, selection of material, and choice of consolidation processes all act together to provide the required level of performance. There are two types of matrix materials used: themosetting and thermoplastic. The consolidation process of each of these types of resins is different. A unit process described in this section addresses the consolidation of composites using polymeric thermosetting resins.

An important family of consolidation processes includes welding and joining processes used to permanently assemble subcomponents. Historically, welding and joining processes are developed empirically and quickly evaluated for benefit in manufacturing applications, driven by the promise of significant potential benefits. The need for welding and joining is substantial since only monolithic parts can be made without joining. The ideal joint would be indistinguishable from the base material and inexpensive to produce (Eagar, 1993). However, experience indicates that no universal joining process exists that can entirely satisfy the wide range of application needs, and thus design engineers must select the most appropriate joining methods that meet requirements. Shielded metal-arc welding, the most widely-used welding process, is described in this section.

The unit processes described here are a representative sample of the types most likely to be encountered. The references listed at the end of the section should be consulted for detailed information on these unit processes.

Polymer Composite Consolidation

Weiping Wang, Alan Ridilla, and Matthew Buczek

A composite material consists of two or more discrete materials whose combination results in enhanced properties. In its simplest form, it consists of a reinforcement phase, usually of high modulus and strength, surrounded by a matrix phase. The properties of the reinforcement, its arrangement, and volume fraction typically define the principal mechanical properties of a composite material. The matrix keeps the fibers in the correct orientation and transfers loads to the fibers.

Continuous-fiber-reinforced materials offer the highest specific strengths and moduli among engineering materials. For example, a carbon fiber/epoxy structural part in tensile loading has only about 20% of the weight of a steel structure of equal stiffness. Composite parts can integrate component piece parts, such as molded-in rib stiffeners, without the need for subsequent assembly operations and fasteners.

Description and Applications. There are many types of polymer composites in use. Composities are usually identified by their fiber material and matrix material. Fiber materials are necessarily strong, stable materials that can be processed into fiber formats. Typcial fibers are glass, graphite/carbon, aramid, and boron. *Glass* fibers represent the largest volume usage since they have excellent properties and are low cost. *Graphite/carbon* fibers are widely used for advanced composite applications in which stiffness and high performance are critical. These fibers are also expensive. *Aramid* fibers, a type of polyamide, have proved useful in applications where its performance in axial tension can be exploited without incurring too severe a penalty by the material's poor performance under compressive loading. *Boron* filaments have high strength and high modulus, but most applications requiring high performance, such as military aircraft structures, are now using carbon fibers.

Low-cost *polyester* resins are the most widely used matrix material for the general composite industry; the majority of these application use glass as the fiber. Many applications are found in the chemical process, construction, and marine industry. The most widely used polymers are *epoxy* resins, which are used with carbon, aramid, and boron fibers for many advanced applications, such as in aircraft structure and rocket motor fuel tanks. *Polyimides* are polymer resins that provide more temperature performance than is possible from the epoxy-based material, and these are used in advanced aircraft structures and jet engine components where heating of the structure will occur.

Principle of Operation. Each of the constituent materials in advanced composites acts synergistically to provide aggregate properties that are superior to the materials individually. The functional effectiveness

of composites is principally due to the anisotropy of the materials and the laminate concept, where materials are bonded together in multiple layers. This allows the properties to be tailored to the applied load so that the structure can be theoretically more efficient than if isotropic materials were used. The reinforcements come in a variety of formats. Unidirectional tapes with all fibers along a common axis, woven fabrics constructed with fibers along both axes in the *x-y* plane, and multidimensional architectures with reinforcements in more than one axial direction are just a few of the available formats.

Consolidation in composites can be considered to occur at two levels: the fibers are infiltrated with the matrix to form a lamina or ply, and the individual laminae are consolidated together to form the final structure. In the prepreg process, these two levels are distinctly separated, since the fiber/matrix consolidation process forms the prepreg, which is then laid up to form the laminate or final component. In other processes, such as resin transfer molding, fiber/matrix infiltration and the consolidation of the final part are done in a single stage. Single-stage consolidation processes are attractive because they eliminate the additional cost associated with prepreg production; however, two-stage consolidation processes have major advantages that often outweigh the benefits of single-stage consolidation. These include flexibility in part geometry, high fiber content, excellent fiber wet-out, and better control of fiber volume fraction distribution. Because of these advantages, prepreg processing is firmly entrenched in high-value products, such as aerospace applications, in spite of its high cost.

Fabrication Methods. Typical steps in manufacturing continuous-fiber composites involve *preform fabrication* and consolidation/curing (Advani, 1994). Preform fabrication creates the structure by positioning material close to the final part shape (Table 13.2.17). The material comes in either the *dry fiber* (without resin) form or with resin included, called prepreg. Dry fibers are used in filament winding, weaving, braiding, and pultrusion. The resin can be introduced in the operation or downstream molding. *Prepreg*, at a higher material cost, eliminates the step of resin addition and provides the adhesion to hold the material together. Consolidation/curing involves compacting the preform to remove entrapped air, volatiles, and excess resins while developing the structural properties by increasing the polymer chain length and cross-linking.

Process Parameters. Thermosetting polymeric materials will not soften and flow upon reheating after polymerization because of the formation of a cross-linked polymer network. Hence, thermosetting polymer matrices must be cured *in situ* with the fibers to form the composite structure. The goals of a successful cure are good consolidation with low porosity and high conversion of initial monomeric constituents to polymer.

The challenge of the cure process is to manage the interactions of temperature distribution, degree of cure, laminate thickness, and void content by manipulating the applied temperature, pressure (or displacement), and vacuum. Temperature must be controlled so that resin temperature stays within limits. Both the duration and the magnitude of pressure application are important as excessive resin flow results in a resin-starved laminate. Similarly, pressure application too soon in the process can entrap volatiles in the material. Materials can also exhibit lot-to-lot variability. The problem is further complicated when processing a complex-shaped part or multiple parts of different geometry.

Composite cure processes are typically performed in an autoclave or in a heated press. An *autoclave* is essentially a heated pressure vessel. Nitrogen gas is normally used for pressurization. The temperature, pressure, and vacuum are controlled vs. time to effect the cure. Recent advancements in *intelligent processing* use sensors to determine the state of cure in real time, and make appropriate control adjustments to optimize the cure cycle (NRC, 1995). Figure 13.2.14, which plots data from an actual implementation, depicts the potential that intelligent processing has in reducing autoclave processing time and cost.

Press molding uses a high-pressure press and matched metal tools to form a part. The main components of a press-molding system include the tools, the ram, and the heated platens. Again, release materials must be applied to the tools to avoid laminate adhesion. Advantages of press molding include improved surface finish (since both sides are tooled) and the elimination of the vacuum bagging systems. However, the pressure inside the mold is not necessarily uniform and volatiles are not easily removed. Hence,

TABLE 13.2.17 Methods of Composite Preform Fabrication

Method	Description	Application
Weaving	Process of interlacing yarns to form a stable fabric construction that is flexible Less frequent interlacing results in better composite strength	Closely conforms to surfaces with compound curvature
Braiding	Intertwines parallel strands of fiber A tubular braid consists of two sets of yarn which are intertwined in "maypole dance" fashion; produced with varying diameter or circumferential size	Sporting equipment Good torsional stability for composite shafts and couplings Geometric versatility and manufacturing simplicity
Pultrusion	Reinforcing fibers pulled from a series of creels through a resin impregnating tank; preformed to the shape of the profile to be produced; enters a heated die and is cured to final shape	Produce constant cross-section pieces at high production rates
Filament winding	Pulling roving (unlisted bundles of fibers) over a mandrel by rotating the mandrel about a spindle axis; can cure on the mandrel	Cylindrical parts Can be low cost
Tube rolling	Material cut from prepreg tapes and laid on a flat surface; plies of different orientations joined together; a cylindrical mandrel is rolled on the material; curing is typically done on the mandrel	Low-cost method for making tubular structures or tapered tubes, e.g., golf shafts
Manual layup	Plies of different fiber orientations are cut from flat sheets of the prepreg material and laid up on a tool; bagging materials are applied and sealed to the tool for subsequent consolidation Flexible in producing complicated features at low start-up/tooling cost; building a curved, variable-thickness part can be complicated	Most common method used in the aerospace industry
Automatic tape layup	Computer-controlled machine tools or robots with a material delivery system • Automated tape layup machine uses prepreg tape, typically 3 to 12 in. wide, suitable for large surface of gentle contours • Fiber placement employs multiple tows for more complex surfaces	Provides lower cost and variability over manual layup, but requires higher capital investment

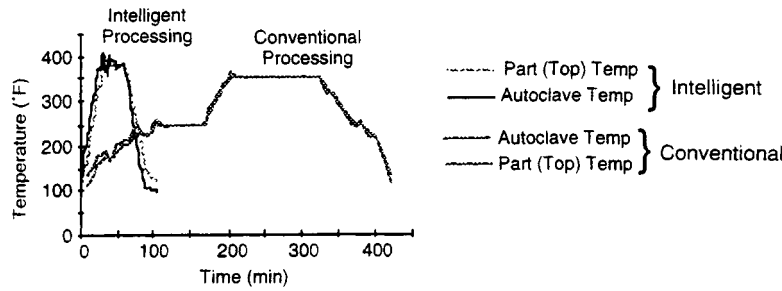


FIGURE 13.2.14 Intelligent processing of composites. (From NRC, *Expanding the Vision of Sensor Materials*, NMAB-470, National Academy Press, Washington, D.C., 1995, 39. With permission.)

press molding is generally suitable for composite systems which do not generate a significant amount of volatiles.

Shielded-Metal Arc Welding

Description and Applications. Bonding is achieved in fusion welding by interposing a liquid of substantially similar composition as the base metal between the surfaces to be joined (Eagar, 1993). The need for welding and joining is substantial since only monolithic parts can be made without joining. Traditional welding processes have unique advantages, which make them the processes of choice for a large number of applications. For example, in the fabrication of heavy structures, arc welding will dominate other assembly processes because of the inherent flexibility and economy of welding.

Principle of Operation. In the majority of arc-welding methods, the workpiece is made part of the electric welding circuit, which has as its power source a welding generator or transformer. To start a weld, an arc is struck by touching the workpiece with the tip of the electrode. The welder guides the electrode by hand in welding a joint, and controls its direction and traveling speed. The welder maintains arc voltage by controlling arc length (the distance between the end of the electrode and the work surface). Because an electric arc is one of the hottest sources of heat, melting occurs instantaneously as the arc touches the metal. Arc welding is a highly popular process because of its flexibility and relatively low cost (ASM, 1993).

In shielded-metal arc welding, an arc is struck between the workpiece and a covered (or coated) metal electrode. Filler metal is provided by the consumable electrode. Combustion and decomposition of the electrode covering from the heat of the welding arc produce a gaseous shield that excludes the oxygen and nitrogen in the atmosphere from the weld area; these gases would otherwise cause excessive porosity and poor ductility in the weld. Welds by this method are of very high quality (Juers, 1993). Figure 13.2.15 depicts the components of the shielded-metal arc welding process.

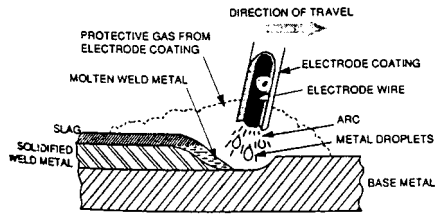


FIGURE 13.2.15 Shielded-metal arc-welding process. (From *ASM Handbook, Welding, Brazing, and Soldering*, Vol. 6, 10th ed., ASM International, Metals Park, OH, 1993, 175. With permission.)

Key System Components. For shielded-metal arc welding, the metallurgical properties of a weld depend greatly on the type of electrode and its covering. The *electrode coverings* contain shielding gas formers that exclude atmospheric gases from the weld area. Electrode coverings offer additional capabilities (Juers, 1993):

- Deoxidizers and nitrogen absorbers to purify the depositing metal
- Slag formers to protect the weld from oxidation
- Ionizing elements to stabilize the arc
- Alloying elements to produce higher-strength welds
- Iron powder to increase metal deposition rate

Selection of the proper electrode is based on many considerations (Juers, 1993):

- Base metal strength
- Base metal composition
- Welding attitude (position)
- Welding current
- Joint design and fit
- Base metal thickness and shape
- Service conditions
- Production efficiency and conditions

Many types and sizes of *power supplies* are used. Supplies can be either direct current (DC) or alternating current (AC) types; combination AC/DC power supplies are widely used. In general, power supplies are required that produce controllable levels of constant-current output. The rate of metal deposition is determined by the output current from the power supply.

Capabilities. Shielded-metal arc welding is the most widely used welding process for joining metal parts, principally because of its versatility. Also, the welding equipment is less complex, more portable, and less costly than for other arc-welding processes.

Shielded-metal arc welding is generally very useful in joining components of complex structural assemblies. Joints in virtually any position that can be reached with an electrode can be welded, even if directly overhead. Joints in blind areas can be welded using bent electrodes. Welding in positions other than flat require the use of manipulative techniques and electrodes that cause faster freezing of the molten metal to counteract gravity. Shielded-metal arc welding can be done indoors or outdoors.

Metals welded most easily by the shielded-metal arc process are carbon and low-alloy steels, stainless steels, and heat-resistant alloys. Cast iron and high-strength steels can also be welded but preheating and postheating may be required. Shielded-metal arc-welding electrode materials are available for matching the properties of most base metals; thus, the properties of a joint can match those of the metals joined.

Design Considerations. Joint design (shape and dimension) is determined by the design of the workpiece, metallurgical considerations, and established codes or specifications.

Welds should preferably be located away from areas of maximum stress. Poorly placed welds can result in undesirable, and unplanned, stress concentrations that can cause early failure of the joint.

Poor joint fit-up increases welding time and is often the cause of poor welds.

Process Limitations. Metals with a low melting point, such as zinc, lead, and tin, cannot be welded by electric arc methods.

Limitations of shielded-metal arc welding compared with other arc-welding methods are related to metal deposition rate and deposition efficiency. Consumable electrodes have a fixed length, usually 18 in. (460 mm), and hence welding must be stopped periodically to replace the electrode. Another limitation is the requirement to remove the slag covering that forms on the weld after each welding pass.

There is a minimum gauge of sheet that can be successfully welded without burn-through. Generally 0.060 in. (1.5 mm) is the minimum practical sheet thickness for low-carbon steel sheet that can be welded by a welder possessing average skill (Juers, 1993).

Special techniques are required when welding pieces of unequal thickness because of their different heat dissipation characteristics. Solutions include

- Placing a copper backing plate against the thinner section to match the heat dissipation from the thick section
- Redesigning the component so that the thick and thin sections taper at the joint to approximately the same size

Distortion is unavoidable in welding because of residual stresses that arise from nonuniform heating and cooling. Various procedures can be used to minimize distortion, such as clamping the workpieces. But straightening of the workpiece may be required to achieve the required dimensional accuracy.

References

Advani, S.G., Ed. 1994. *Flow and Rheology in Polymer Composites Manufacturing*, Composite Metals Series, Vol. 10, Elsevier, New York.

ASM. 1984. *Powder Metallurgy*, 1984. *ASM Handbook*, Vol. 7, 9th ed., ASM International, Metals Park, OH.

ASM. 1993. *Welding, Brazing, and Soldering*, ASM Handbook, Vol. 6, 10th ed., ASM International, Metals Park, OH.

David, S.A. and Vitek, J.M., Eds. 1993. *International Trends in Welding Science and Technology*, ASM International, Metals Park, OH.

Eagar, T.W. 1993. Energy sources used for fusion welding, in *Welding, Brazing, and Soldering, ASM Handbook*, Vol. 6, 10th ed., ASM International, Metals Park, OH, 2–6.

Froes, F. 1996, *Hot Isostatic Pressing*, ASM International, Metals Park, OH.

Humpston, G. and Jacobson, D.M. 1993. *Principles of Soldering and Brazing*, ASM International, Metals Park, OH.

Jenkins, I. and Wood, J. V., Eds. 1991. *Powder Metallurgy: An Overview*. The Institute of Metals, London.

Juers, R.H. 1993. Shielded metal arc welding, in *Welding, Brazing, and Soldering, ASM Handbook*, Vol. 6, 10th ed., ASM International, Metals Park, OH, 175–180.

Linnert, G., 1994. Fundamentals, in *Welding Metallurgy: Carbon and Alloy Steels*, Vol. 1, 14th ed., American Welding Society, Miami, FL.

MPIF. 1995. *Powder Metallurgy Design Manual*, 2nd ed., Metal Powder Industries Federation, Princeton, NJ.

National Research Council (NRC). 1995. Intelligent processing of advanced materials, in *Expanding the Vision of Sensor Materials*, NMAB-470, National Academy Press, Washington, D.C., 34–40.

Schwartz, M.M. 1994. *Joining of Composite Matrix Materials*, ASM International, Metals Park, OH.

Woishnis, W.A. 1993. *Engineering Plastics and Composites*, 2nd ed., ASM International, Metals Park, OH.

Mechanical Assembly

S. H. Cho

Total labor involved in the assembly processes in the U.S. varies from 20% (farm machinery) to almost 60% (telecommunications equipment). On average, assembly tasks occupy 53% of manufacturing time, and 10 to 30% of total production cost of most industrial products. Use of improved assembly methods and technologies is essential to reduce overall manufacturing costs.

Assembly Methods and Systems

Assembly systems are classified in several different ways (Table 13.2.18). Figure 13.2.16 indicates the types of automated systems that could be cost-effective based on assembly part count and production volume.

Generally, *automatic assembly systems* consist of three major components:

- *Transfer system* to move work carriers with in-process subassemblies between workstations;
- *Parts feeding device* to supply parts to be assembled into the appropriate position, where the parts are loaded by the handling/placing mechanism;
- *Parts handling/placing mechanisms* to pick parts and perform assigned assembly tasks such as placing, inserting, and screwing.

TABLE 13.2.18 Classification of Assembly Systems

Type of Assembly System	Classification Basis	Description
Manual	Level of automation	Assembly tasks completed manually
Automatic	Level of automation	Adopts mechanized devices or industrial robots with supplementary equipment for handling and assembling parts
Semiautomatic system	Level of automation	Manual workers and mechanized devices cooperate to complete assembly tasks
Cell-type system	Configuration	Very flexible integrated assembly workstation; assembly completed by various equipment, such as robots or pick-and-place units, parts feeders, parts tray, magazines, automatic tool changer, and auxiliary jig/fixtures
Line-type system	Configuration	Assembly tasks divided into subtasks which are completed at workstations connected by transfer systems; handles large parts, cycle-time variation, and gripper change
Dedicated system	Degree of flexibility	Not flexible — can assemble only one product of a single model; generally economical for large production volumes
Flexible system	Degree of flexibility	Accommodates different products; economical for medium-size and mixed-model production

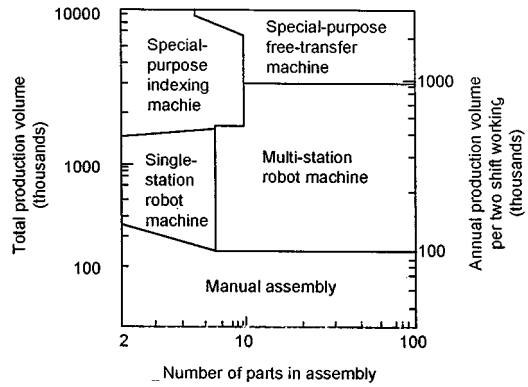


FIGURE 13.2.16 Conditions for economic application of various assemby systems. (Courtesy of S. H. Cho.)

The *transfer system* can be continuous or intermittent according to the transfer method. In the *intermittent transfer system*, assembly tasks in a workstation are performed during a stationary period of the work carriers, which are transferred to the subsequent workstation after completion of the assembly tasks. *Continuous transfer systems* have a problem of not being able to assure positioning accuracy between an in-process subassembly and the tools of a workstation; thus, the intermittent type is usually used (Boothroyd, 1992).

There are two modes of *intermittent transfer:* in-line and rotary. For the *in-line* mechanisms, the walking beam, the shunting work carrier, and the chain-driven work carrier are commonly used. The *rotary* type employs mechanisms such as rack and pinion, ratchet and pawl, Geneva mechanism, and mechanical cam drives. Currently, free-transfer conveyors with stopping and positioning mechanisms are widely used in flexible assembly lines.

Most *feeding systems* have devices to orient parts supplied by the following means (Yeong and Vries, 1994):

- *Bulk supply,* for parts that are easily separated, fed, and oriented automatically. Bulk supply usually adopts various part feeders, e.g., vibratory bowl feeders and various nonvibratory mechanical feeders for small parts. These feeders usually only handle one type of part and cannot be applied to assembly systems which require flexible part-feeding devices.

- *Organized supply* uses special pallets such as a kit or a magazine for parts that cannot easily be separated, fed, and oriented.

Parts-handling/placing mechanisms include pick-and-place units and various types of industrial robots. To load and assemble parts, the mechanism usually employs various assembly wrists: jaw-type gripper, vacuum suction pad, magnetic chuck, screwdriver, nut runner, and others. A number of different wrists are required when automatically assembling different parts in a workstation. To accommodate this assembly situation, a multifunctional gripper, a tool-changing system, and a universal gripper have been developed and widely used in robotic assembly systems.

Selection of Assembly Systems

The placement of a part in its assembled position and part mating impose tight constraints on the positioning mechanism and on part properties such as clearances and geometry. These constraints are more severe for assembly systems mating a variety of precision products that must adapt to frequent design changes. The assembly systems of this type must possess the adaptability to changing assembly environments, thus requiring *flexible automatic assembly* (Boothroyd, 1984).

A typical *flexible robotic assembly system* uses an industrial robot for part handling, part positioning, and part mating. These systems are limited by positioning and orientation misalignment caused by the low positioning accuracy of robots, uncertainty in part handling, and variation in the location of parts.

Various approaches are available that take into consideration uncertainty in orientation and parts properties variation. Figure 13.2.17 depicts various types of assembly wrists that are in use. In general, wrists can be classified into three basic configurations:

- Passive accommodation;
- Active accommodation;
- Passive-active accommodation.

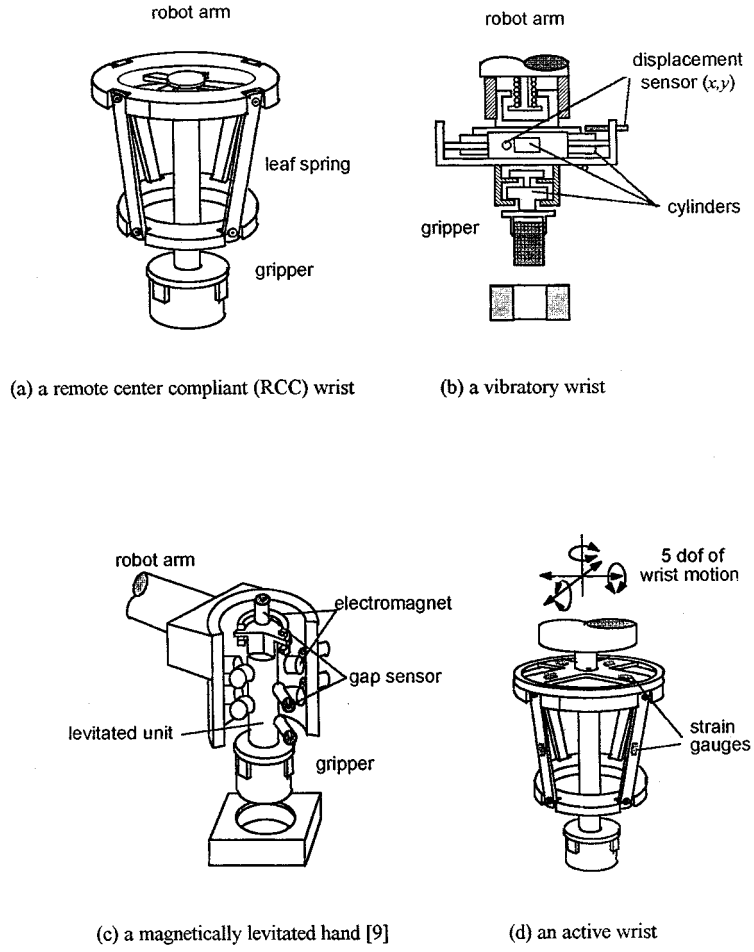
There are two types of *passive wrist methods:*

- Wrist accommodates misalignments by deforming its structure elastically under the influence of the contact forces generated during the assembly of the misaligned parts. A *mechanically compliant structure* is needed for either the robot wrist or the assembly worktable, which can be deformable according to the reaction force acting on the mating parts. Remote center compliant wrist is one of such typical wrists; this method usually requires part chamfering (Cho et al., 1987).
- Wrist corrects misalignment by *applying external forces* or torques to the misaligned parts in a prescribed manner or a random way. For instance, a vibratory wrist utilizes pneumatic actuators controlled by a pulse width modulation controller to generate desirable vibration; this method does not require part chamfering.

Active wrist methods employ sensor-controlled wrists and compensate misalignments by controlling the fine motion of the assembly wrist or the work table based upon sensory feedback. Advances have been made in the area of sensing technique, gripper and actuating mechanism design, and the related control algorithms. Sensors for these wrists are needed prior to contact (vision, range, displacement, proximity), during contact (touch, slip), and after contact (force, moment). Based upon the force sensor signals and the associated algorithms, the wrist motion can be corrected to reduce misalignment.

The *passive–active accommodation method* is achieved by combination of the “passive and active” techniques. The basic strategy is that the part mating is continued within some allowable forced moments, while beyond this the insertion method is switched from the passive to active to reduce the mating force by using sensors with compliant structures.

The *sensor-based assembly* is similar to the method employing the active wrist. Both rely on sensory information for fine-motion control. The principal difference is that the former utilizes robot motion, while the latter relies on wrist motion.



(a) a remote center compliant (RCC) wrist (b) a vibratory wrist

(c) a magnetically levitated hand [9] (d) an active wrist

FIGURE 13.2.17 Various assembly wrists. (Courtesy of S. H. Cho.)

Reaction forces (*force/torque information*) that occur during part mating can represent the status of the contact between mating parts. Among the approaches are pattern classifiers that determine the contact state at which the assembly parts are contacting each other, position error recovery via fuzzy logic, heuristic search with fuzzy pattern matching, and learning of nonlinear compliance using neural network.

Visual/optical information is critical to compensate for positioning and orientation error occurring due to misalignment. Visual information is often combined with other data, such as force/torque, pressure, and displacement, because a rather longer time is required for image processing, object recognition and error calculation and because visual information is sensitive to external environmental conditions such as illumination.

Assembly Line

An assembly line usually consists of a set of workstations that perform distinct tasks linked together by a transfer mechanism. Each task is an assembly operation, while each workstation represents a location along the line where the tasks are processed. A buffer storage is placed between workstations for reducing the effect of a workstation failure on the throughput.

Line balancing is essential for designing a cost-effective assembly system (Groover, 1980). The time required for the completion of a task is known as the process time, while the sum of the process times

of the tasks assigned to a station is the station time. The total task processing times for assembling all the parts is the work content. When *assembly sequences* are generated without considering line balancing, the sequences may not guarantee the minimum number of workstations. Therefore, line balancing must be concurrently considered in defining the assembly sequences.

Design for Assembly

Approximately 80% of manufacturing cost is determined at the conceptual design stage. Design for assembly (DFA) is crucial during early design. The objective of the DFA is to facilitate the manufacturing and assembly of a product. DFA applies to all the assembly operations, such as parts feeding, separating, orienting, handling, and insertion for automatic or manual assembly (Ghosh and Gagnon, 1989).

DFA is directed toward:

- Reducing the number of parts by modularization;
- Easing feeding and minimizing reorientation;
- Easing insertion by self-aligning, self-locating, elimination of part interference, and efficient fastening.

Axiomatic DFA uses design guidelines based on experience of product designs and assembly operations. *Procedural DFA* evaluates the design efficiency based upon the production cost.

Figure 13.2.18 shows results that were obtained using DFA rules to evaluate part designs in consideration of feeding and insertion. Such analysis has great potential in improving assembly operations.

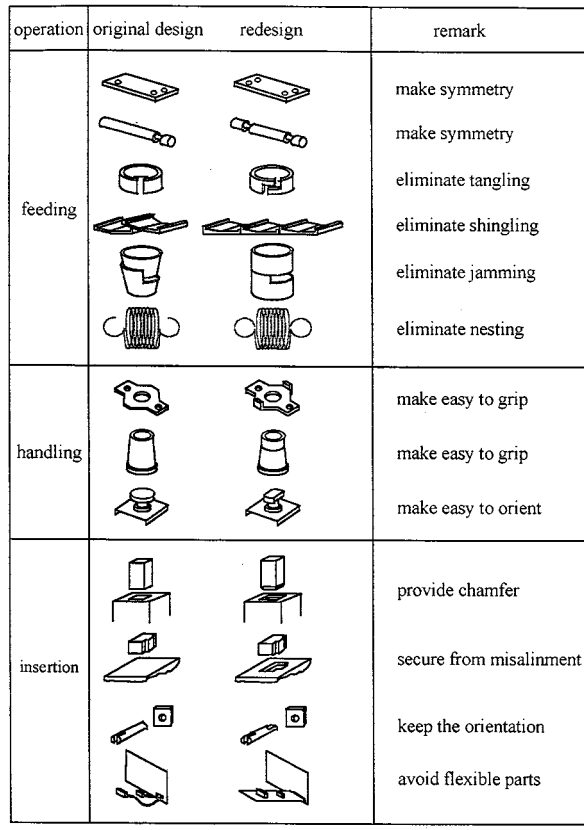


FIGURE 13.2.18 Results of DFA analysis. (Courtesy of S. H. Cho.)

References

Boothroyd, G. 1984. *Economics of General-Purpose Assembly Robots,* CIRP General Assembly, Madison, WI.

Boothroyd, G. 1992. *Assembly Automation and Product Design,* Marcel Dekker, New York.

Cho, H.S., Warnecke, H.J., and Gweon, D.G. 1987. Robotic assembly: a synthesizing overview, *Robotica,* 5, 153–165.

Ghosh, S. and Gagnon, R.J. 1989. A comprehensive literature review and analysis of the design, balancing and scheduling of assembly systems, *Int. J. Prod. Res.*, 27(4), 637–670.

Groover, M.P. 1980. *Automation, Production Systems, and Computer-Aided Manufacturing,* Prentice-Hall, Englewood Cliffs, NJ.

Yeong, M.Y. and Vries, W.R. 1994. A methodology for part feeder design, *Ann. CIRP,* 43(1), 19–22.

Material Handling

Ira Pence

Material handling provides the right amount of all the required materials at the right place and time to support manufacturing. Properly designed, the material-handling system provides for the acquisition, transportation, and delivery of material such that the minimum cost is incurred considering capital, labor, and expenses. It focuses on obtaining material and supplies, moving them between process steps, and delivering the finished product to the customer. Figure 13.2.19 depicts the elements of a modern materials-handling system which reaches beyond the factory floor, serving as an integrating force for production operations.

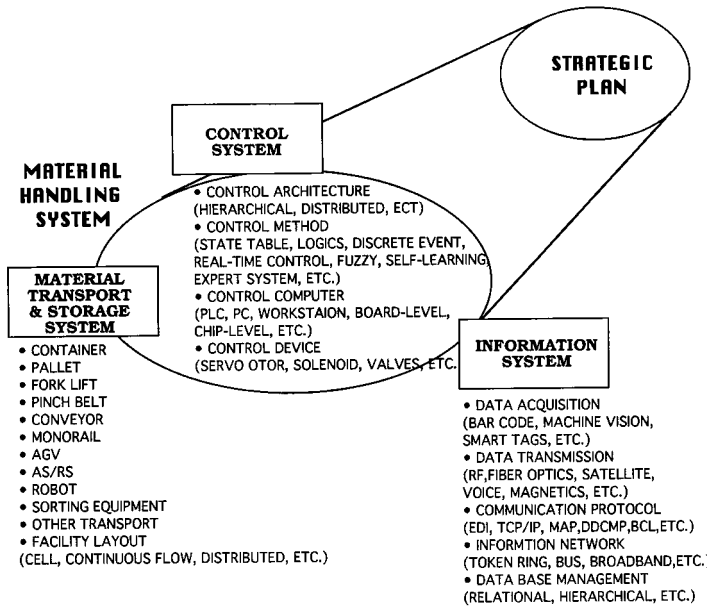


FIGURE 13.2.19 Modern material-handling system. (Courtesy of J. Lee.)

The material-handling system should be analyzed as a single entity so trade-offs in expense in various elements can be made to minimize the total cost. Often, however, the activities are broken into purchasing, transportation, warehousing, and distribution functions which operate independently. Such divisions are arbitrary because the material-handling system is an unbroken chain of activities that start when the suppliers complete their last value-added operation and extend until the product is delivered. Minimum-

cost material handling will not be achieved if the partitioning of material flow into subelements leads to optimization of each subelement at the expense of the entire system.

Planning is crucial to the design of a smoothly functioning materials operation. Planning should be consistent with the strategic plan for the entire manufacturing operation. In formulating the plan, a realistic assessment of the problems and opportunities associated with arranging, controlling, and implementing the material flow must be used. Specific tactics, such as just-in-time delivery, should support the strategic plan. Expensive automation should not be used unless the volume and stability of the product justify its use.

Logistics

The movement of material in a manufacturing enterprise is usually broken into two broad categories, inside the plant and external to it. External movement is generally referred to as "logistics," while the internal flow is known by many different names. Material movement considered logistics is generally marked by a wide geographic scope, diversity of equipment and technology, and some uncertainty in delivery time. But the basic goals are identical to in-plant movement and most of the analysis tools are applicable to both categories.

For large manufacturing enterprises the number of individual components involved in the production of the product, often referred to as SKUs (stock keeping units), can be very large. In the past, it was helpful to treat items of similar size, weight, storage requirements, and delivery times as a single commodity. This simplified manual analysis and was sometimes done by computer programs that analyze logistical information. However, current computer systems make it practical to treat each item individually.

Basic Elements

The basic activities associated with material handling are *moving*, *storing*, and *controlling* materiel. These activities are interrelated with production scheduling and information must flow both ways.

Movement may be over short distances, as from one machining center to another, or long, from one plant to another. In all cases the basic information that must accompany the move includes the part identity, timing, quantity, source, and destination. Each move should be planned and scheduled, taking into account the speed of the basic mechanism as well as allowance for loading, unloading, logging, counting, etc.

Movement can be continuous or in batches, synchronous or asynchronous, horizontal or vertical. Each move should be examined for the characteristics of urgency, safety, size, weight, etc. before selecting the technology to be used to perform the move. Singular instances should be accommodated in the most expeditious manner, and repetitive moves must be made in the most efficient and effective manner.

One of the basic tenets of material handling is to retain *control* of the material. Inventory control generally maintains up-to-date records on the quantity and location of material on hand. However, in a *virtual warehouse*, it is important to know what material has been ordered and the status. Control includes procedures and equipment to properly handle the material.

The *storage* of material should be minimized. Materials are typically stored to compensate for uncertainty in delivery systems and to allow ordering of economic quantities.

Occasionally unexpected changes in production will result in delaying the release of material to the factory floor and that material must be stored until needed. *Just-in-time delivery*, where parts are delivered directly from local suppliers to the assembly line several times a day, has been proved in several industries. Thus, the need for storage due to uncertainty of delivery has been reduced. At the same time, global sourcing has increased. As supply lines lengthen, the uncertainty of delivery increases. The more complex the supply system, the more likely the occurrence of an unexpected delay. In designing a storage system, the trade-off of higher transportation cost but less inventory vs. volume discounts on purchases and transportation, but with storage costs, must be evaluated.

Further Information

The Material Handling Industry produces two catalogs each year. One provides information on the publications available from the MHI, including all the educational material; standards and specifications; operating, maintenance, and safety manuals; and reference works. The other provides a directory of member companies and the products they manufacture. Both are available from the MHI Literature Department (704) 522-8644; FAX (704) 522-7826.

Case Study: Manufacturing and Inspection of Precision Recirculating Ballscrews

Toskiaki Yamaguchi, Yashitsugu Taketomi, and Carl J. Kempf

The precision and quality of the components of mechanical devices used in both industrial and consumer products must meet high performance and durability requirements. The manufacturing challenges that apply to ballscrews, in general, are representative of other components such as gears, shafts, and bearings. This case study on precision ballscrews illustrates how the different unit manufacturing processes presented in this chapter apply to a particular application, and how design and manufacturing engineering decisions are affected by quality and cost considerations.

Many of the processing steps and fundamental techniques discussed in this case apply to the manufacture of many other precision components. The case study also illustrates the rationale for continuously improving production processes and discusses strategies for improvement that benefit from past experience.

Overview of Ballscrew Design and Manufacturing Considerations

Ballscrews convert rotary motion into linear motion. Ballscrews have low friction compared with standard leadscrews, and have enabled precise control of mechanical systems at a relatively low cost. They are used extensively in production machinery, such as milling machines, and are being applied in other fields, such as robotics, inspection equipment, and office automation equipment.

The key components of a ballscrew system consist of a screw shaft with a spiral groove, a nut with a corresponding spiral groove that rides along the shaft, and balls which are captured between the shaft and the nut. A recircultation tube provides a return path for the balls from the end of the nut. Components in this assembly are typically made from steel alloys, chosen to provide a good combination of toughness, surface hardness, and ease of manufacture.

Ballscrews are applied in a variety of ways, but the most common configuration is one in which a rotational input of the shaft imparts a translational motion to the nut. The shaft is normally supported by rotating bearings at both ends, and the translating element attached to the nut is supported by linear guide bearings. Ballscrews range in size from very small units with a shaft diameter on the order of 0.08 in. (2 mm) and a length of about 4 in. (100 mm) to very large units with a shaft diameter on the order of 12 in. (300 mm) and a length of up to 50 ft (16 m). Key dimensional features which are controlled to a high degree of precision for an assembled ballscrew system are shown in Figure 13.2.20.

The main factors in ballscrew performance are accuracy and lifetime. *Accuracy* is determined primarily by the precision of the screw lead, i.e., the linear displacement of the nut which is produced by rotary displacement of the shaft. The measurements used to assess the *precision* of the lead are discussed below. To obtain a *long lifetime*, the shaft must have high surface hardness in order to withstand loads at the shaft/ball and nut/ball interfaces. Since ballscrews may be subject to a variety of loading conditions, the ballscrew must possess good impact strength and toughness.

Secondary design factors include low audible noise, low static and dynamic friction, low friction variation, minimal backlash, high mechanical stiffness, and resistance to dirt and contaminants. Ballscrews are often used in specialized applications, such as operation in vacuum or ultraclean environments, in corrosive or dirty environments, in a thermally controlled environment, and in environments where vibrations must be minimized. To meet these demands, basic ballscrew designs are adapted to

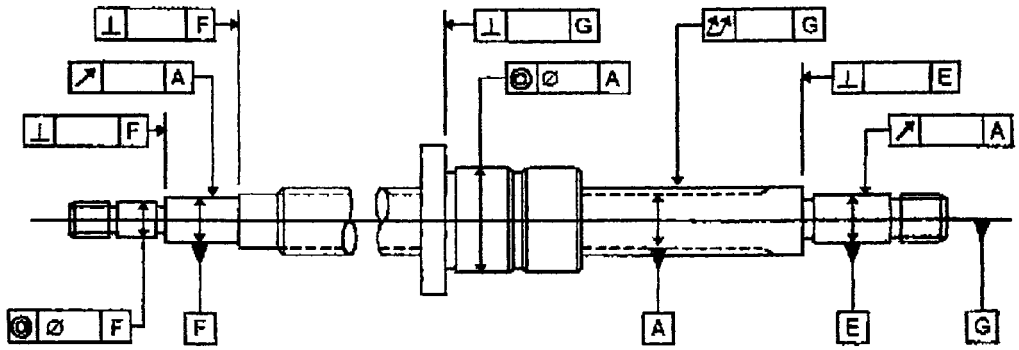


FIGURE 13.2.20 Key ballscrew dimensions. (Courtesy of T. Yamaguchi, Y. Taketomi, and C. J. Kempf.)

use special materials, surface treatments, and to provide features such as hollow shafts for coolant flow or damping materials.

Major steps in production and assembly of traditional ground ballscrews are summarized as a flowchart in Figure 13.2.21.

Initial Machining Operations

A lathe is used to establish the outer diameter of the shaft and cut the groove. In both operations, an allowance is made for material that will be removed in the final grinding operations. For short shafts, the support is provided from the shaft centers and a single cutting tool can be used with multiple passes to produce the desired outside diameter and groove. For long ballscrews, the lateral deflection of the shaft during cutting operations is significant and workpiece supports are necessary to prevent deflection of the shaft.

For long ballscrews, the total processing time can become extremely long using multiple passes with a single-point cutting tool. In such cases, a multiple-point tool that can remove more material on a single cutting pass is more efficient to use. But a multipoint cutter requires more time for setup and adjustment. Thus, there is an economic trade-off between single-point and multipoint cutting techniques.

Because of subsequent heat-treat operations, the shaft will undergo dimensional changes. The design must allow for these dimensional changes. This is an area in which production experience is critical. Statistical analysis of previous manufacturing results, together with modeling of material behavior, is critical to continued design and proceess improvement.

Surface Treatments

In order to withstand the loads at contact points, a very high surface hardness is necessary. A surface hardness of Rockwell C 58-62 with a depth on the order of 0.8 to 1.2 mm is required. For short ballscrews, *carburization* is used to develop the necessary surface hardness. For longer shafts, carburization is impractical because of the size of carburizing furnaces. In this case, an induction hardening process is used.

Electrically heated gas furnaces are used in the carburizing process. The immersion in the carburizing furnace is followed by quenching and tempering. During carburizing, key parameters such as temperature and gas concentrations are continuously monitored and adjusted by a process control system. Because high hardness is needed only in the ball groove, areas of the shaft and nut that will be subject to subsequent machining operations are coated before carburizing. This coating, which is applied like paint, prevents the diffusion of carbon into the material surface.

For longer shafts, a different steel is chosen and an induction hardening process is used. The general trend within the industry is toward induction hardening since it has the advantage of being a continuous process. When using induction hardening, single shafts can be processed immediately after machining. Induction hardening machines require less initial capital than for a carburizing system.

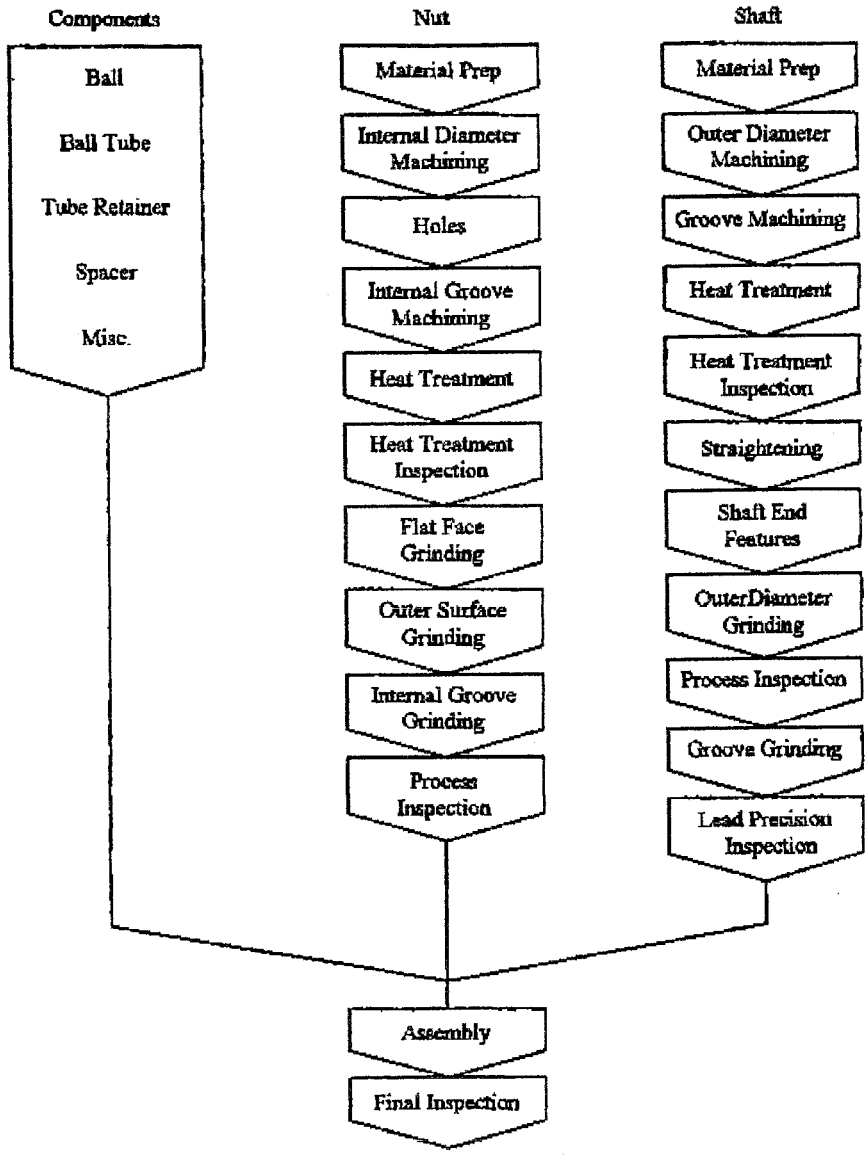


FIGURE 13.2.21 Ballscrew processing flowchart. (Courtesy of T. Yamaguchi, Y. Taketomi, and C. J. Kempf.)

The shaft undergoes dimensional changes as a result of the surface heating, and a *straightening* operation is necessary. Prior to straightening, the shaft is chilled in a carefully controlled manner to sub-zero temperatures to ensure that the solid-state transformations are completed; the result is a more stable microstructure. For straightening, a *specialized press* is used to measure automatically the deviation from perfect straightness; it automatically applies loads along the shaft as necessary to restore the shaft to near-perfect straightness.

Grinding and Finishing Operations

After surface treatment, chilling, and straightening, the outside diameter of the shaft is ground to the final dimensions. Except for very short ballscrews, the grinding cutting forces will cause lateral deflection of the shaft, and thus the centers in the shaft ends cannot be used as datums for final machining operations. Consequently, the outside diameter of the shaft is ground to precise tolerances. Although the outer

diameter of the shaft is not a functional part of the finished ballscrew, it is used as a datum surface when the final grinding of the shaft groove is performed.

Grinding of the shaft can be done on either centerless grinders or cylindrical grinders. Since centerless grinders take a long time to set up, they are generally used for large production runs of shafts having the same diameter. For the case of cylindrical grinders, the shaft is supported at the centers in the shaft ends as well as at workpiece supports along the length of the shaft. By adjusting the work rests, minor variations in taper and bending of the shafts can be corrected. This processing step requires considerable operator skill and experience in order to minimize variations in the outer diameter and residual bending of the shaft.

To maintain precision and finish of the outer diameter, automatic balancing of the grinding wheel is necessary. CBN grinding wheels are used although they are more expensive than traditional abrasive wheels; their long life and low wear rates make them economical.

After turning the external diameter to final size, shaft end features are produced. These features include bearing seats, keyways, flats, and locknut threads. The final finishing operation for the ballscrew generates the ball groove. This is the key processing step in assuring that the ballscrew possesses the required lead accuracy. If this process is not carefully controlled, there will be variation in the lead, depth of the groove, and smoothness of the groove, causing subsequent problems with accuracy, stiffness, running force and noise, and lifetime.

Removal of the outer layer of hardened material during outer diameter grinding can cause minor bending of the shaft as the net residual stresses in the material change. Prior to final cutting of the ball groove, the straightness of the shafts is checked and minor corrections made.

To maintain very high lead precision, the final groove grinding is done in a specially temperature-controlled environment of $68 \pm 2°F$ ($20 \pm 1°C$). Cutting oil is applied liberally to the shaft to minimize thermal effects due to cutting and deviations between the grinding machine and shaft temperature. As in the case of external diameter grinding, automatic balancing is used to minimize vibrations. To further reduce vibrations, each grinding machine is mounted on an individual base to minimize vibration coupling between various machines. The isolation properties of the machine bases are adjusted when a new machine is placed in service and undergo periodic inspection and adjustment during operation.

Since the size and lead of the groove varies from ballscrew to ballscrew, the grinding wheel must be matched to the groove shape. Thus, an inventory of several types of wheels is necessary; traditional abrasives are used since the costs of CBN wheels would quickly become prohibitive.

In general, accuracy in the groove-grinding process requires a combination of modeling and statistical analysis. As in the case of surface treatment, the collection and analysis of past production data allows continuous refinement of the manufacturing process.

Assembly and Inspection

The main factor in ballscrew accuracy is the lead. To facilitate a quantitative measurement of the lead, four fundamental parameters are used. To measure the lead error, precision measurements over a long travel range are made using computer-controlled laser interferometry.

Unacceptable variations in friction, increased running noise, and a reduction in life result from poorly formed grooves. Thus, the depth and cross-sectional profile of the groove must fall within allowable tolerances. Direct measurement of groove cross section is quickly performed on an optical profile projector. Precise measurement of groove depth can be made on selected samples. The screw shaft is supported between two centers and rotated at a fixed speed. A table carrying a contact probe moves along the screw shaft in the axial direction synchronously with the screw rotation. The probe is placed in contact with the groove to measure variations in the groove depth over the entire length of the shaft. By performing a frequency analysis on the groove depth errors and accounting for the rotational speed of the ballscrew during the measurement process, frequencies of unwanted vibrations occurring in the production machinery can be detected and the source of the anomaly eliminated to improve the production process.

Even with good control of the screw shaft and nut groove diameters during production, additional steps are necessary to obtain the desired amount of axial play or preload in the assembled ballscrew. By selecting slightly different ball sizes, required axial play or preload can be achieved. For a given nominal ball diameter, balls usually are grouped in steps of 20 to 40 μin. (0.5 or 1.0 μm). To increase production efficiency, assembly jigs are used.

The majority of precision ballscrews are preloaded in order to remove backlash and achieve the desired axial stiffnes. Two methods of achieving the desired preload are used. The first preloading method increases the ball size until the desired preload is achieved. The second method uses double-nut preloading in which a spacer is inserted between the two nuts to take up the axial play and achieve the desired preload.

The preloaded ballscrew has some running torque when the screw is rotated. The relation between this running torque and the preload has been determined based on both theoretical and experimental studies. Since preload cannot be measured directly, it is estimated based on measurements of the running torque in a specialized torque-measuring machine.

A special machine is used for direct measurement of axial stiffness. In this process, the screw shaft is clamped and an axial load is applied to the nut. A displacement sensor is fixed to the shaft close to the nut and the relative displacement of the nut can be measured when force is applied. The measurements of force and displacement can be plotted on an X-Y recorder to depict the stiffness characteristics of the ballscrew assembly.

References for Case Study

Oberg, E., Ed. 1971. *Machinery's Handbook*, 19th ed., Industrial Press, New York 2044–2068.

Yamaguchi, T., 1983. Ballscrew manufacturing, and inspection, *Tool Eng. Mag.*, June, 92–99 (in Japanese).

13.3 Essential Elements in Manufacturing Processes and Equipment

Sensors for Manufacturing

John Fildes

Introduction

A good modern definition for a sensor must capture the diversity of these devices. A sensor is a device that detects or measures the state or value of a physical or chemical variable and provides the result in a useful way. At the minimum, a sensor contains a transducer that converts the detected or measured quantity to another form of representation. For example, a very simple sensor is an indicator whose color changes upon reaction with a minimum amount of a chemical species. Nonetheless, the sensors that are normally encountered are more complex, containing a transducer, an output display, and possibly supporting electronics for signal conditioning, communications, and logic functions.

Sensor technology is undergoing rapid change because of three developments. One development is the emergence of integrated and smart sensors, wherein transducers have been miniaturized, usually through the use of silicon micromachining, and integrated with electronics for signal conditioning, communications, and logic functions. The second development is the ongoing adaptation of nondestructive evaluation (NDE) measurements and laboratory measurements for on-line use in supervisory and intelligent process control systems. These NDE and laboratory-type measurements require rather complex systems, with extensive signal conditioning and data analysis. The third development is also related to the emergence of supervisory control systems. The data from multiple sensors is comparatively analyzed in a process called data fusion to better identify the state of the system and the occurrence of process faults. Thus, the topic of sensors and sensory systems now encompasses transducers, integration with supporting electronics for communications and logic functions, data analysis techniques, and data fusion methods.

Classification of Sensors

A good taxonomy for sensors is provided by the requirements for different degrees of process control, which is shown in Figure 13.3.1. This taxonomy contains three classes: sensors used in regulatory feedback control loops, process analyzers, and product quality analyzers. In Figure 13.3.2, the basic element of control is the regulatory feedback loop that maintains controllable processing parameters at the desired values. These types of sensors, which usually provide a single value and are relatively generic in their applicability, are used for monitoring variables such as temperature, pressure, flow, level, displacement, proximity, and velocity. For use of sensors in regulatory control, sensitivity, selectivity, simplicity, speed, reliability, and low cost are the critical attributes. Historically, sensors were almost solely used in this function, but this is no longer the case. Sensors, or more properly sensory systems, are now also used as process analyzers and product quality monitors. Sensors of this type, which are used for feedback in supervisory control, are more complex and application specific, and their output tends to be a matrix of values (e.g., a video image or a spectrum). Sensors of this type provide a representation of the process or product that has greater information content, but is more abstract than the representation provided by sensors used in regulatory control. Thus, extensive computations and modeling are needed for process analyzers and product quality sensors, but there is also more ability to correct for deficiencies in sensitivity and selectivity through computational means. In this case, measurement speed is usually less demanding. These differences in the three types of sensors are summarized in Table 13.3.1.

The typical regulatory variables are temperature, pressure, flow, level, humidity, position, and motion. These variables form the basis for many process factors as summarized in Table 13.3.2 and the major sensor technologies for these variables are summarized in Table 13.3.3.

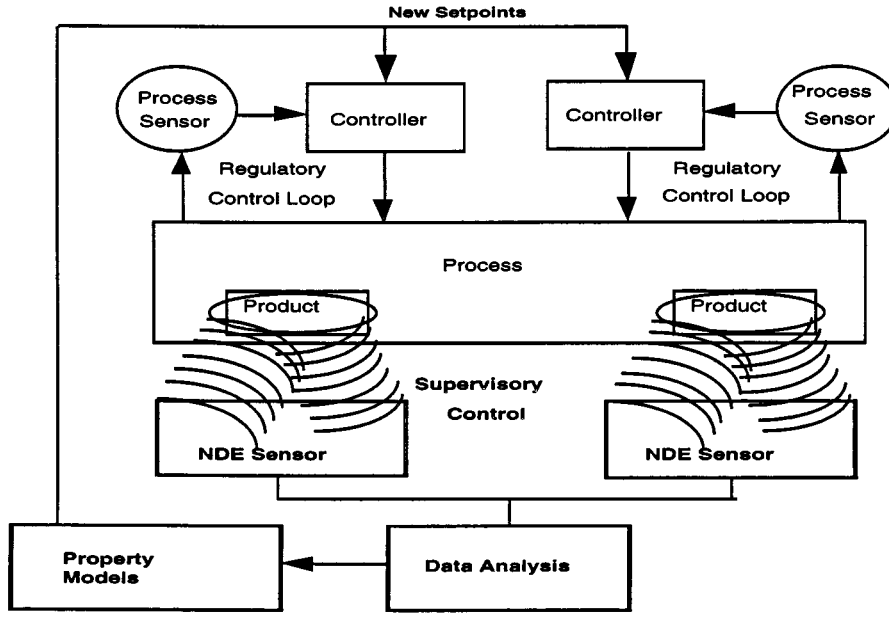


FIGURE 13.3.1 Regulatory and intelligent control.

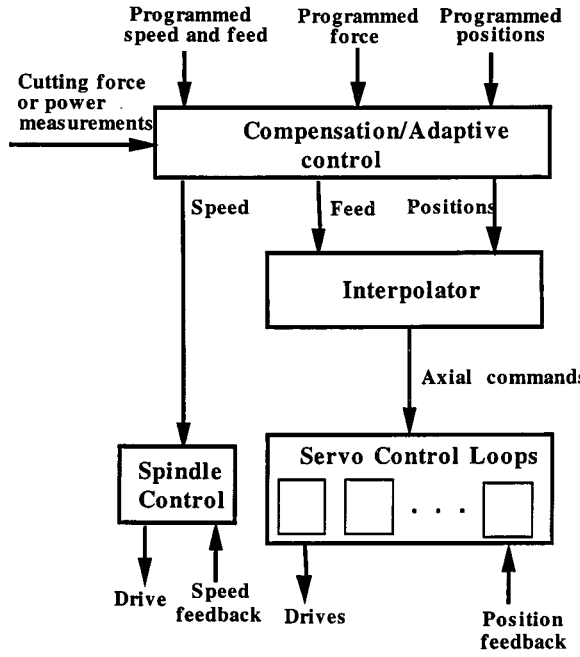


FIGURE 13.3.2 Typical CNC architecture.

Integration refers to inclusion of signal conditioning on the same substrate as the transducer. One of the major advantages of integration is that the integrated sensor can operate with a very small transducer output, which results in significantly smaller sensors. With the advent in telecommunications and the Microelectricalmechanical systems (MEMS) technologies, we can integrate software, hardware, power

TABLE 13.3.1 Characteristics of Regulatory and Intelligent Process Control Sensors

Characteristic	Regulatory Feedback Sensors	Process and Product Sensors
Speed of response	Very fast	Slow
Type of output	Single value	Matrix of values
Relationship of output to system parameter	Simple and direct	Abstract representation
Amount of computation	Little or none	Extensive
Sensitivity	Inherently high	Can be improved by computation and modeling
Selectivity	Inherently high	Can be improved by computation and modeling
Cost	Very low to moderate	High
Size	Small	Large
Applicability	Broad	Application specific

TABLE 13.3.2 Regulatory Processing Variables

Controllable Variable	Influenced Processing Factors
Temperature	Rates of chemical reactions; degree of cure of polymers; degree of softening and viscosity; cooking times; drying times; annealing times; degree of carborization
Pressure	Consolidation density; porosity; concentration of gaseous chemical reactants; forging; bending; injection times in molding
Flow	Mix ratios; quenching time; concentration of chemical reactants; color; pH; humidity
Level	Capacity; mix ratio; concentration of chemical reagents
Humidity	Drying of paper and grains; cooking
Position	Parts handling and placement; machining accuracy
Motion	Assembly; machining accuracy

TABLE 13.3.3 Sensor Technologies for Regulatory Control

Variable	Major Sensor Technologies
Temperature	Thermocouples, thermistors, resistance thermometers, infrared pyrometers, acoustic pyrometers
Pressure	Manometers, bellows and diaphragms, strain gauges, piezoresistive, piezoelectric, capacitive, thermocouple, ionization
Flow	Differential pressure (orifice plate, Venturi, pitot tube); velocity (turbine, vortex, electromagnetic, Doppler ultrasonic, time-of-flight ultrasonic); positive displacement (rotary sliding vane, gear, impeller); mass flow (thermal, coriolis); variable area (rotameters)
Level	Floats, pressure, radio frequency, ultrasonic, microwave, resistance tape, optical
Humidity	Dew point, length of hair, conductivity, capacitive, resonate
Position	Rotary encoder, linear variable differential transformer, potentiometers, magnetostrictive, magnetic, inductive proximity, magnetic, ultrasonic
Motion	Tachometers, pitot tube, anemometers

source, and communications on the same chip. Table 13.3.4 lists the advantages and disadvantages of integrated sensors.

TABLE 13.3.4 Characteristics of Integrated Microsensors

Characteristic	Advantage/Disadvantage
Batch fabrication	Excellent control, low cost, sensor arrays
Loss of modularity	Difficult to package and limitations on materials
On-chip amplification	Better signal-to-noise ratio
On-chip compensation	Better accuracy through compensation for interferences
On-chip feedback	Better linearity
On-chip scaling and conversion	Standardized output
On-chip multiplexing	Sensor arrays, bus addressable

Smart sensors are defined as those that also include logic functions, rather than just signal-conditioning electronics. Smart sensors may also include communications circuitry, diagnostics, and sometimes control outputs so that they can be directly connected to actuators, and documentation and trending functions. Conventional pressure sensors have rather broad accuracy limits, such as 0.25% of span. High-end smart pressure sensors have an accuracy of less than 0.1% of span, and mid-range smart pressure sensors approach this accuracy. Smart pressure transmitters are already available with built-in proportional-integral-derivative control functions, and fuzzy logic capabilities will soon be available. Multivariate measurement is not yet available, but will become so soon. Most likely, pressure and temperature will be the two measurements. Smart sensors are also available for presence detection, positioning, infrared photodetection with triangulation, and flowmeters, specifically magnetic meters, Coriolis mass flowmeters, and ultrasonic flowmeters. Integrated and smart sensor technology is also improving accelerometers, proximity detectors, and tactile sensors. The use of integrated and smart sensors will increase because of the overloading of shared resources and the use of distributed control schemes to solve this problem.

Use of Sensors in Supervisory and Intelligent Control Systems

As shown in Figure 13.3.1, supervisory control augments regulatory control by using process analyzers to better characterize the state of the process and sensors of product characteristics to assess the outcome of the process and to determine corrections for unacceptable deviations. These corrections are then implemented by the regulatory controllers. There are many ways in which supervisory control can be implemented. A common way is to measure important aspects of the process and the product performance, and to use a model to relate variations in processing parameters to variations in product performance.

Process analyzers are central to enabling supervisory control because they allow a more accurate, but abstract, representation of the state of the process. These devices augment regulatory control by monitoring other important variables in the process, but ones which are generally not directly controllable. Use of process analyzers provides improved reliability, flexibility, predictive diagnostics, ease of use, and central data collection with process documentation, trending, recipe handling, and statistical quality assurance. Process analyzers are now used for monitoring important process gases such as oxygen, carbon monoxide and dioxide, oxides of nitrogen and sulfur, and hydrocarbons. For liquids, process analyzers are available for pH, conductivity, redox potential, dissolved oxygen, ozone, turbidity, specific ions, and many organic compounds. The measurement techniques are summarized in Table 13.3.5.

TABLE 13.3.5 Process Analyzers

Process Analyzer Technology	Typical Uses
Electrochemical (potentiometric, amperometric)	Gases such as carbon monoxide and dioxide, oxides of nitrogen and sulfur, and hydrocarbons; species in liquids such as dissolved oxygen, pH, redox potential, specific ions, organic compounds, inorganic compounds
Chromatography (gas and liquid)	Gases such as CO and CO_2; species in liquids such as alcohols, flavors, lipids, polymers, and other organic compounds
Infrared spectroscopy	Near infrared — web processes for thickness, composition, solvent, coating Mid-infrared — chemical and petrochemical processes, polymers, food processes
Ultraviolet/visible	Gases and liquids including all elemental halogens, other inorganics, aromatics, carbonyls, many salts of transition metals

Traditionally, process analyzers have been stand-alone devices with a single sensor and a dedicated operator interface. The trend is toward modularity, multiple sensors, and digital communications so that process analyzers can be incorporated into distributed control systems in an open architecture control environment. Smart sensor technology is also turning conventional sensors into process analyzers. A good example is provided by smart infrared temperature sensors. These sensors provide sample and hold, correction for reflected radiation when the emissivity of the target is less than one and its temperature

is lower than ambient, analog outputs for control, digital outputs, trend analysis, and area sampling with line scanners. This functionality has made temperature mapping of surfaces in furnaces practical and is used in annealing of aluminum, steel reheating, and oven drying of webs such as paper.

Another trend is the use of sensor fusion, where information from multiple sensors is combined to improve the representation of the process. Sensor fusion techniques can use mechanistic models, statistical models, or artificial intelligence techniques, such as neural networks, fuzzy logic, and expert systems. Sensor fusion provides more reliability because the validity of the data of each sensor can be assessed from the other sensors, and if a sensor is faulty, its value can be predicted from the other sensors. Sensor fusion provides a better representation of the process because it captures the interrelationships that often exist between processing variables, which are treated independently without sensor fusion.

As shown in Figure 13.3.1, intelligent control involves the use of product quality sensors. Of the three classes of sensors discussed in this section, these sensors tend to be the most complex and provide the most complete, but abstract, representation of the state of the product. In some cases, the same technologies are used for product quality sensors as for process analyzers. An obvious example would be where chemicals are reacted without a change of state. In other cases, specialized product quality sensors have emerged. Table 13.3.6 lists some of the product quality sensors that are being used or are being experimentally evaluated.

TABLE 13.3.6 Product Quality Sensors

Sensor	Applications
Time-of-flight ultrasound	Either thickness or elastic constants, if the other is known
Electromagnetic acoustic transducer (EMAT)	Either temperature or elastic constants, if the other is known
Eddy currents	Electrical conductivity, magnetic permeability, thickness, temperature, presence of flaws
NMR (high resolution for liquids, low resolution for solids)	Composition of chemicals, petroleum, foods, polymers, fibers
Infrared and Raman spectroscopy	Degree of cure of polymers

Computer Control and Motion Control in Manufacturing

Yoram Koren and M. Tomizuka

Computerized Numerical Control Architecture

A typical architecture of a CNC system consists of three levels, as shown in Figure 13.3.2. At the lowest level are the axial servocontrol loops and the spindle controller. These servoloops are closed at high sampling rate. The interpolator that supplies the axial position commands to the control loops is at the intermediate level of this architecture. At the highest level are the compensation algorithms for the errors of mechanical hardware deficiencies, such as machine geometry errors and thermal deformation of the machine structure. This level also includes adaptive control algorithms that adapt the machine feed and speed to the cutting tool and workpiece material to maximize machine productivity at rough cutting and maintain precision at fine cutting.

CNC Part Programs

The CNC software consists of a control program and part programs. The numerical data which are required for producing a specific part by a CNC machine is called the *part program*. The part program is arranged in the form of *blocks* of information, where each block contains the numerical data required to produce one segment of the workpiece. Each block contains, in coded form, all the information needed for processing a segment of the workpiece: the segment shape and length, its cutting speed, feed, etc. Dimensional information (length, width, and radii of circles) and the contour shape (linear, circular, or other) are taken from an engineering drawing. In NC, dimensions are given separately for each axis of

motion (*X, Y,* etc.). Cutting conditions such as cutting speed, feed rate, and auxiliary functions (coolant on and off, spindle direction, clamp, gear changes, etc.) are programmed according to surface finish and tolerance requirements.

The part program contains the required positions of each axis, its direction of motion and velocity, and auxiliary control signals to relays. The controller generates an internal signal indicating that the previous segment is completed and that the new block of the part program should be read. The controller also operates the drives attached to the machine leadscrews and receives feedback signals on the actual position and velocity of each one of the axes.

In CNC systems the part dimensions are expressed in the part programs by integers. Each unit corresponds to the position resolution of the axes of motion and is referred to as the *basic length-unit* (BLU). The BLU is also known as the "increment size" or "bit-weight," and in practice it corresponds approximately to the accuracy of the system. To calculate the position command that the computer sends to the CNC machine, the actual length is divided by the BLU value. For example, in order to move 0.7 in. in the positive *X* direction in a system with BLU = 0.001 in., the position command is *X* + 700.

In the first generations of CNC systems, dimension were given in part programs by BLUs, as in NC. In new CNCs, however, dimensions, or desired cutter positions, are given in a normal way, as to a regular computer. The command *X* – 0.705, for example, will move the *X* axis the negative direction by 0.705 in. The resolution by which the dimension commands are given depends on the system BLU.

In addition to cutter positions, the part programmer must program the machining parameters such as tool diameter, cutting speed (*n*), feed (*s*), and depth of cut (*d*). The task of the part programmer is to convert the machining parameters *n* and *s* to NC control variables — spindle speed (*N*) and feed rate (*f*), as shown in Figure 13.3.3.

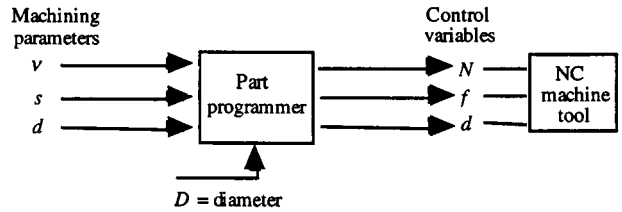


FIGURE 13.3.3 Conversion of machining parameters to control variables.

Point-to-Point and Contouring Axes of Motion

CNC systems consist of two functional types: point to point (e.g., drilling machine) and contouring or continuous path (e.g., milling machine).

Point-to-Point Systems. The simplest example of a point-to-point (PTP) CNC machine tool is a drilling machine. In a drilling machine the workpiece is moved along the axes of motion until the center of the hole to be drilled is exactly beneath the drill. Then the drill is automatically moved toward the workpiece (with a spindle speed and feed which can be controlled or fixed), the hole is drilled, and the drill moves out in a rapid traverse feed. The workpiece moves to a new point, and the above sequence of actions is repeated.

In a PTP system, this system requires only position counters for controlling the final position of the tool upon reaching the point to be drilled. The path from the starting point to the final position is not controlled. The data for each desired position is given by coordinate values. However, in high-speed drilling applications, such as the task of single-spindle drilling of an engine block, a control loop is needed to control the acceleration and deceleration of the motion. The digital signal processing technique has been used to control the settling time of the motion system for very fast PTP motion profile (1 to 4 g) so that the spindle could perform drilling operations without breaking the drill. A linear motor-based machine tool has been built and demonstrated by Anorad Corp. (Hauppauge, NY) for Ford Motor's engine machining line.

Contouring Systems. In contouring, or continuous-path, systems, the tool is cutting while the axes of motion are moving, as, for example, in a milling machine. All axes of motion might move simultaneously, each at a different velocity. When a nonlinear path is required, the axial velocity changes, even within the segment. For example, cutting a circular contour requires a sine-rate velocity change in one axis, while the velocity of the other axis is changed at a cosine rate.

In contouring machines, the position of the cutting tool at the end of each segment together with the ratio between the axial velocities determines the desired contour of the part, and at the same time the resultant feed also affects the surface finish. Since, in this case, a velocity error in one axis causes a cutter path position error, the system has to contain continuous-position control loops in addition to the end point position counters. Consequently, each axis of motion is equipped with both a position loop and a position counter. Dimensional information is given in the part program separately for each axis and is fed to the appropriate position counter. Then, an *interpolator*, in the controller, determines the proper velocity commands for each axis in order to obtain the desired tool feed rate.

Interpolators. In contouring systems the machining path is usually constructed from a combination of linear and circular segments. It is only necessary to specify in the part program the coordinates of the initial and final points of each segment and the feed rate. The operation of producing the required shape based on this information is termed *interpolation*, and the corresponding software algorithm in CNC is the *interpolator.* The interpolator coordinates the motion along the machine axes, which are separately driven, to generate the required machining path. The two most common types of interpolators are the linear and circular. Parabolic interpolators are also available in a few CNC systems which are used in the aircraft industry.

Linear interpolator: The ability to control the movement along a straight line between given initial and final coordinates is termed *linear interpolation.* Linear interpolation can be performed in a plane (two dimensional), using two axes of motion, or in space (three dimensional), where the combined motion of three axes is required. In this chapter only two-dimensional linear interpolators are discussed. To illustrate the interpolator function, consider a two-axis system, where a straight cut is to be made. Assume that the X axis must move p units at the same time that the Y axis moves q units. The contour formed by the axis movement has to be cut with a feed rate of V length-units per second (e.g., mm/sec). The numerical data of p, q, and V are contained in the part program and are fed into the interpolator. The interpolator then generates two velocity signals V_x and V_y, where

$$V_x = \frac{pV}{\sqrt{p^2 + q^2}}$$

and

$$V_y = \frac{qV}{\sqrt{p^2 + q^2}}$$

The position reference inputs to the axial control loops are

$$R_x = V_x t$$

and

$$R_y = V_y t$$

As seen, the two-dimensional linear interpolator supplies velocity commands simultaneously to the two machine axes and maintains the ratio between the required incremental distances.

Circular interpolator: The two most common interpolators in CNC systems are linear and circular. The *circular interpolator* eliminates the need to define many points along a circular arc. Only the initial and final points and the radius are required to generate the arc. The circular interpolator divides the arc into straight lines with a contour error smaller than one BLU. It operates on an iterative basis, where

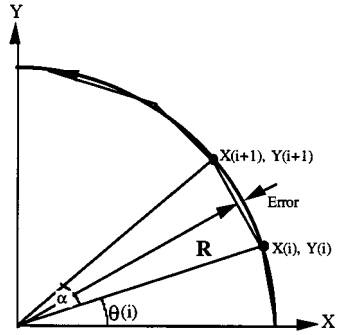


FIGURE 13.3.4 A circular interpolator divides the arc to straight lines.

for a given point, $x(i),Y(i)$, a small incremental angle is added to calculate the next point, as shown in Figure 13.3.4.

Both the linear and circular interpolators are based on generating incremental positions every T seconds, where T is the sampling period of the CNC system. Typically, CNC systems operate on one common sampling period for both the interpolator and the control loops. In these systems a typical T may be between 1 and 10 msec. Some other CNC systems utilize a separate computer for the interpolation and another one or several microprocessors for closing the control loops.

Motion Control Systems

Control Loops. A typical closed loop of a CNC machine is shown in Figure 13.3.5. The computer compares the command and the feedback signals and gives, by means of a digital-to-analog converter, a signal representing the position error of the system, which is used to drive the DC servomotor. The feedback device, which is an incremental encoder in Figure 13.3.4, is mounted on the leadscrew and supplies a pulsating output. The incremental encoder consists of a rotating disk divided into segments, which are alternately opaque and transparent. A photocell and a lamp are placed on both sides of the disk. When the disk rotates, each change in light intensity falling on the photocell provides an output pulse. The rate of pulses per minute provided by the encoder is proportional to the revolutions per minute of the leadscrew.

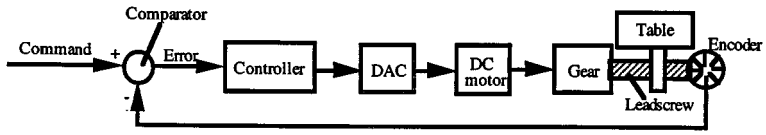


FIGURE 13.3.5 Closed loop of CNC.

Normally, the motor is mechanically coupled to the load via a drive mechanism. For example, robot links and positioning tables are loads. It is usually a good assumption to ignore the dynamics in the current feedback loop and regard the command current as the actual current, which sets the torque input to the mechanical portion of the system. Notice that the inertia and bearing friction of the motor must be considered as a part of the mechanical portion of the motion control system, which strongly influences the design of the velocity and position loop feedback controllers. These controllers can be implemented

in either analog or digital form. In recent years, digital implementation on microprocessors and/or DSPs has become popular, and digital velocity and position controls are often referred to as *digital servo*. The current feedback is usually built into the "drive" (power amplification system). While feedback signals for the velocity and position controllers are usually obtained from the motor, the velocity and position of the load, e.g., a positioning table, at the opposite end of the drive mechanism are the quantities of our ultimate concern. Closed-loop control schemes based on the motor velocity and position are sometimes called semi-closed-loop control schemes. The velocity and position of the load must be fed back for full closed-loop control. Typical drive mechanisms are ballscrews and various types of gears. Several manufacturers provide so-called direct drive (DD) motors, which are capable of delivering large torques but with a significantly reduced maximum speed. DD motors may eliminate drive mechanisms. However, they are heavy and may not always be the best solution depending on applications. Common sensors for positions are potentiometers and shaft encoders, and those for velocities are tachogenerators and frequency-to-voltage convertors (FVC), the input to which is encoder pulses. In digital servos, encoders are popular for measuring positions, and velocities are either estimated from encoder pulses or are obtained from FVCs.

Traditional velocity and position loop feedback controllers are of PID (proportional plus integral plus derivative) type. The output of a PID controller is

$$u(t) = k_p e(t) + k_i \int_0^t e(\tau)\, d\tau + k_d \frac{de(t)}{dt}$$

where e(t) is the error, $u(t)$ is the controlling input, and k_p, k_i, and k_d are, respectively, proportional, integral, and derivative control gains. The above equation represents the PID control law in the continuous-time (analog) form. In digital control, the PID control law is implemented in a discretized form. A typical discrete-time (digital) PID control law is

$$u(k) = k_p e(k) + k_i T \sum_{j=0}^{k} e(j) + \left(\frac{k_d}{T}\right)\left[e(k) - e(k-1)\right]$$

where k denotes the kth sampling instance and T is the sampling period. In position loop feedback control, e and u correspond to the positioning error and the velocity command, respectively. For the velocity loop controller, they are the velocity error and the current command. Another popular linear controller is the lead/lag compensator. Input-output (I/O) interfaces include analog-to-digital convertors (A/D), decoders for processing encoder pulses, and digital-to-analog convertors (D/A). Motion control systems can be built from components and programmed with custom software, they can be purchased as plug-in boards for various buses, or they can be purchased as packaged systems.

PID and lead/lag compensators are simple and utilized in many applications. However, they alone might not be adequate for problems where the performance requirements are stringent. Extreme care must be taken during the design of a closed-loop control system. By increasing the magnitude of the feedback signal (e.g., more pulses per one revolution of the leadscrew), the loop is made more sensitive. That is known as increasing the open-loop gain. Increasing the open-loop gain excessively may cause the closed-loop system to become unstable, which obviously should be avoided.

The basic nature of "feedback" control is that the control action is based on the error. When the reference input for the position loop is fixed, the integral action may assure zero error at the steady state. In tracking control, however, the position command is continuously varying, which combined with the dynamics of the closed-loop system makes tracking errors always remain. For example, in contouring of a circular arc in machining, the command position signal for each motion control axis is sinusoidal. Such an operation is essentially a test of the frequency response of the motion control axis. The controller is normally tuned so that the frequency response gain is close to but is not exactly 1 in the operating range. In high-speed contouring operations, corresponding frequencies are high, and the gain is normally

below unity. Then, the actual diameter is slightly smaller than the desired diameter. This consequence is often called the radial reduction error. Such errors may be reduced by applying the disturbance observer scheme to the position loop. However, the disturbance observer is still a feedback controller, and one sampling time delay in digital implementation further diminishes its effectiveness.

The following discussion describes some principles of motion control systems for PTP machines and contouring machines.

Control Loops for Point-to-Point Systems. The control loops of PTP systems are designed to control the position of the machine tool axes. Each axis is *separately* driven and should follow the command signal. The system design starts by selecting the type of control: open loop or closed loop, a decision which depends on the required specifications of the NC system and economy. Open-loop controls use stepping motors as the drive devices of the machine table. The drive units of the stepping motors are directly fed by the controller output pulses. The selection of the appropriate motor depends on the maximum torque, required velocity, and step size in the system. Stepping motors can be implemented on small-sized PTP systems in which the load torque is small and constant. Larger PTP machines and contouring systems utilize closed-loop control systems.

In PTP systems each axis is driven separately at the maximum allowable velocity. This velocity depends on the drive type and on the mechanical structure of the particular machined or manufacturing system. In order to avoid large overshoots the velocity is decelerated before the target point in which the tool starts to operate (e.g., to drill). Since the path between the points is insignificant, *the deceleration is accomplished in each axis separately.*

In practical systems the deceleration is accomplished by three stages. A typical three-stage deceleration diagram of one axis of the table is given in Figure 13.3.6. The table moves at rapid velocity V until reaching a distance L_1 from the target point, where the table is instructed to move at smaller velocity V_1. After a time delay, which depends on the system inertia, the table moves at a new velocity V_1 until reaching a distance of L_2 units from the target point, where again the velocity is reduced to V_2. When the table is at a distance of L_3 units before the target point, the velocity is reduced once more and the table "creeps" toward the final point at very low velocity V_3, and subsequently stops.

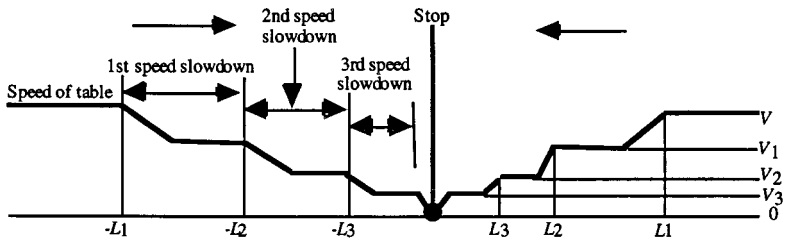


FIGURE 13.3.6 Deceleration procedure in a PTP system.

Control Loops for Contouring Systems. The control in CNC contouring systems operates in closed loops, which compare the command pulses from the interpolator with the feedback signal from the encoder or the resolver. The most sophisticated design applies to the closed-loop control of contouring systems. In the design of these loops, the transfer function of each element must first be determined; the system is then set up in block diagram form, and finally the loop gain is established based on performance analysis. The transfer function of each element is based upon its mathematical model. In establishing the mathematical model the engineer is faced with a compromise between *accuracy* and *complexity,* on one hand, and *approximation* and *simplicity,* on the other. In this section we shall discuss simple models, in which the principles of design can be readily illustrated.

The control loops of contouring systems are usually of the closed-loop type as shown in Figure 13.3.7. They use two feedback devices: a tachometer that measures the motor speed and is included in the drive unit and a position feedback transducer which is capable of also measuring the axis velocity (such as

an encoder, resolver, or inductosyn). In encoder-based systems the encoder is mounted on the leadscrew and emits pulses; each pulse indicates a motion of 1 BLU of axis travel. Therefore, the number of pulses represents position and the encoder pulse frequency is proportional to the axis velocity.

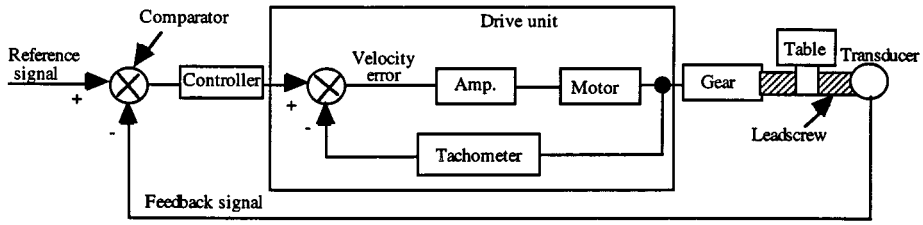


FIGURE 13.3.7 Control loop of contouring system.

Error Sources in CNC Systems

Despite the high precision of the CNC equipment, it still has position and contour errors. Error sources in CNC machines are classified into three categories:

1. Mechanical Hardware Deficiencies
   - Orthogonality of machine axes
   - Straightness of the machine axes
   - Thermal deformation of the machine structure
   - Backlash in the gears and leadscrew nuts
   - Uneven leadscrew pitch
   - Friction in moving components (leadscrews, guideways, etc.)
2. Cutting Process Effects
   - Tool deflection due to cutting forces
   - Large vibrations and chatter
   - Tool wear
   - Workpiece thermal deformations
   - Workpiece deformations due to cutting forces
3. Controller and Drive Dynamics
   - Disturbances due to cutting forces
   - Disturbance due to friction forces in the machine guideways
   - Contour errors caused by tracking errors of nonlinear contours (e.g., circles)
   - Contour errors caused by mismatch of equivalent parameters in the axial controllers
   - Corner errors in contouring operations

In general, errors of the first type, such as machine geometry errors and thermal deformation of the machine structure, can be compensated for by the machine controller. To further understand the measurement of errors, the next section will examine the metrology for precision engineering.

References

Ohnishi, K., Matsu, N., and Hori, Y. 1994. Estimation, identification, and sensorless control in motion control system, *Proc. IEEE*, 82(8), 1253–1265.

Tomizuka, M. 1993. On the design of digital tracking controllers, *ASME J. Dyn. Syst. Meas. Control*, 115(2), 412–418.

Metrology and Precision Engineering

Kam Lau

Introduction

Precision engineering in manufacturing generally refers to the engineering processes of achieving tighter tolerances or dimensional accuracy of a design. The process of precision engineering begins from the part *design*, part *fabrication*, and, finally, part *inspection*. This section addresses some of today's precision engineering considerations in fabrication and inspection.

Factors in Precision Engineering

The fabrication process is a material transformation process that determines the final part dimensions through material forming, removal, or insertion. Precision in the fabrication generally refers to the dimensional "repeatability and accuracy" of a part transformed under such a process. The factors to be considered to ensure good "precision" are (1) errors of the *machine* system(s), (2) the *environment* in which the machining is being performed, (3) the *manufacturing process*, and (4) the *instrumentation* to verify the performances of the factors.

Machine Errors. Machine errors can be classified into geometric, thermal-induced, dynamic, and structural errors. Geometric errors are errors related to the undesirable machine geometry caused by nonorthogonality (squareness error) of axes, linear positioning (or scale) error, reversal (or backlash) error, straightness error, pitch, yaw, and roll errors of a machine axis during a linear move. If the machine is equipped with a rotary table, geometric error would include axis wobble, rotational positioning error, eccentricity, and parallelism errors.

Geometric Errors: Geometric errors are much better understood in the machine industry than the thermal, dynamic, and structural errors. The techniques and instruments to check for the geometric error are very well established and are commonly applied in the machine tool industry. However, for small- to medium-size machines, geometric errors constitute only about 25% or less of the total manufacturing error. For larger machines, the percentage can go as high as 50%. As such, just knowing the geometric errors of a machine system may not be adequate in the realization of the total manufacturing error.

Traditional devices for measuring different types of geometric errors are the mechanical square, straightedge, autocollimator, electronic level, step gauge, optical polygon, and dial indicator. In some instances, the combined use of these devices is needed. Newer instruments that can offer faster and more precise measurements are the laser interferometer system, the 5-D laser interferometer system, and the telescopic ballbar.

Thermal Errors: Thermally induced machine error is considered one of the key factors affecting the accuracy of a machine tool. Thermally induced errors arise as a result of nonuniform heat generation within the machine such as in motors, bearings, and guideways; heat generated during the cutting process; and the coolant effect as well as the environmental effect resulting in the uneven growth of the machine structure. Thermal error can contribute as much as 50% of the total manufacturing error in small- to medium-size machines. However, this effect has been largely unrecognized or often ignored by the machine tool industry until recently.

There are basically two alternatives to monitor the thermally induced errors — intermittent and continuous monitorings. Intermittent monitoring generally involves the use of a touch probe or dial indicator(s) to periodically measure against one or multiple fixed positions as the machine is going through a thermal exercise. In the case of the dynamic spindle thermal study, an artifact such as a sphere or a rod can be mounted on the spindle while it is running at a certain speed. At some elapsed time (e.g., 1-min interval), the spindle will stop momentarily and the machine will reposition the artifact to the dial indicator(s) to check for the repeatability. Any deviations from the initial position (i.e., cold-start position) are considered as thermal growth due to the spindle warm-up. A typical spindle thermal test takes about 4 to 8 hr to complete. A similar procedure can be applied for the environmental and servomotion-related thermal growth measurements.

The continuous thermal growth monitoring generally involves the use of some noncontact sensors such as capacitance gauges, optical sensors, or inductance sensors arranged in the similar fashion as above. The benefits of continuous monitoring are that there is no interruption of the spindle dynamic and that the machine positioning repeatability does not necessarily interfere with the results. Furthermore, the noncontact nature generally allows the spindle to operate at any speeds, thus giving a much broader range of thermal assessment.

Laser interferometer systems are also used occasionally to monitor thermal growth effects. Because of its ability to identify the growth along the entire axis, the laser measurement is better suited for monitoring the growth of a linear-scale system caused by internal or external heat sources. Linear-scale error caused by axis movement is about one fifth that of the spindle thermal in medium to small machines. It is even less in larger machine since the heat dissipation effect in larger machine is much more effective.

Dynamic Errors: Dynamic errors here refer to those caused by the motions of a CNC machining center. These includes the servo-gain mismatch, servo-stick-slip, servo-oscillation, controller error, etc. Errors pertaining to tool chattering, structural deformation (caused by machine carriage acceleration and deceleration), machine/part deadweight, etc. are considered structural errors and are discussed in the later section.

Servo-gain mismatch is often caused by the electrical (or computer) gain setting of one axis not matching the other. The problem is not severe when the machine is primarily used for static positioning purposes, such as drill, boring, etc.; however, it can be a problem in precision contour milling when two or more axes are to be used in synchronization with each other. The magnified elliptical error is caused by one axis responding faster than the other in reaching the commanded positions. Servo-gain mismatch can easily be corrected in a routine machine maintenance.

Structural Errors: Structural errors include tool-chattering error and structural deformation error (due to acceleration and deceleration of the machine carriage, the deadweight distribution, the cutting force, etc.).

Tool chattering generally affects the machinability and surface finish of the workpiece, not the dimensional accuracy. Structural deformation caused by the acceleration and deceleration of the machine carriage and workpiece is rather insignificant for quasi-static positioning and slow-speed contouring. However, the error can be significant if the contouring speed is high. Another major contributor to the structural error is the cutting force. An excessive amount of cutting force can cause the spindle axis, the tool, the fixture, and the part to deform.

A good device to gauge the potential structural error is a compliance system. A basic compliance system consists of a load cell and a dial indicator, which are set up between the machine table and the spindle. The table is programmed to move in small increments (e.g., 5 mm) in either directions of the load cell. The readings from the load cell (force) and the indicator (actual displacement of the table) are then recorded at every increment. A compliance chart can then be obtained by plotting the forces (F) against the differences between the actual and commanded displacements (DD) of the table (i.e., F vs. DD).

Another type of structural error for large machines is from the machine foundation error. Most large machine bases are built in sections. These sections are then aligned, assembled, and anchored together to a common reinforced concrete foundation. As such, the foundation becomes part of the machine structure and the accuracy and repeatability of the machine are therefore heavily dependent on the stability of the foundation. It is not unusual to find the performance of the machine degraded as a result of floods and earthquakes, and loosening of anchor supports due to prolong use or lack of maintenance.

Errors Introduced by Environmental Effects. For the manufacturing plants that have little or no control of the plant environment, environmental effects can be very significant sources of errors. Two of the most dominant environmental effects are thermal and vibration effects.

Thermal Errors: For many nontemperature-controlled manufacturing plants, it is not unusual to observe a total temperature swing of 20°F throughout a day of operation. This wide fluctuation of environmental temperature can cause significant accuracy and repeatability problems to the machining

systems, as well as the workpieces. Even in a somewhat controlled environment, thermal errors can still be a major problem. For instance, if the machine is located in the vicinity of a frequently operated bay door, where there is a substantial temperature difference between the plant temperature and the outdoor temperature, or if it is placed next to a heat source such as a welder, an electric blower or exhaust, a hydraulic pump, or under direct sunlight, the heat source can still cause tremendous localized thermal distortion of the machining system.

In most cases, one can reduce the localized environmental thermal effect by isolating the heat sources with simple panel shielding or by relocating the machines. If an overall plant temperature control is unachievable or impractical for economic reasons, one may consider (1) applying localized thermal control of certain key machining systems or (2) implementing computer thermal compensation techniques.

Vibration Errors: Environmental vibration error is a result of one or more external vibration sources affecting the structural stability of the machine system. This type of error can be significant if the machining system is located next to a heavy punch press operation or where frequent forklifting operation is present. Using a spindle analyzer or a laser interferometer system to access the amount of the environmental vibration error is common.

Other Errors: Other types of environmental errors are the interference or instability of the electric power source, the factory-supplied pneumatic and hydraulic pressures, air pressure, and humidity, etc. These types of errors may be of lesser magnitude than the above; however, it is always good practice not to underestimate their potential effects in any precision manufacturing considerations.

Errors Introduced by the Manufacturing Process .

The magnitudes of these types of errors are very much dependent on the process control and manufacturing practices implemented by each individual plant. In general, the concerns in this area are the effects of the coolant on the workpiece and the machine structure, the pallet and the fixture, the repeatability of the pallet and tool changer, and the tool deflection in manufacturing.

Coolant Effects: Eighty percent of machining uses coolant, sometimes referred to as *wet machining.* In most cases, the coolant temperature is not controlled. For smaller machining systems, a small coolant tank can be found next to the machine. For larger machines, a large coolant tank can be found underneath the ground. The main purposes of the coolant in machining are (1) for lowering the cutting temperature and (2) for chip removal. As the coolant is being recycled during the machining process, its temperature gradually warms up. This significant increase in temperature affects the workpiece dimensions since the workpiece temperature before machining is generally at room temperature. Similarly, the pallet dimensions are also affected.

Tool and Fixture: The conditions of the pallets, tools, and fixtures often govern the repeatability of a manufacturing system. These components should be checked routinely for wear and chipping. Fixtures and clamping devices should be routinely checked to ensure that proper clamping forces can be applied and that contact surfaces are in good condition. Many of these checks can be accomplished visually or by performing dial indicator repeatability checks.

When excessive cutting force is expected, it is necessary to consider the maximum possible amount of the tool, workpiece, and fixture deflections resulting from the force. The amount, if it exceeds the manufacturing tolerance, should be reduced by either reducing the depth of cut or the feed rate or by strengthening the tool and fixture. A compliance system is a good qualifier to measure the tool deflection under load.

Instrumentation and Inspection in Precision Engineering

This section introduces the instrumentation often used in precision engineering.

CNC Machining Performance Evaluation

The American National Standards Institute developed *ANSI B5.54 Standards for CNC Machining Center Performance Evaluation* for the purpose of providing detailed guidelines for machine tool users and developers to evaluate and compare the performances of CNC machining systems. The *Standards* delineates procedures for the measurements of the machine and environmental errors as stated in the above, as well as choices of conventional and state-of-the-art instruments used in doing the measurements. Experience indicates that these measurements are useful not only for performance evaluation, but also for better understanding of the sources of errors. This information is crucial for preventive maintenance and for accuracy enhancement of CNC machine systems.

Some of the key techniques introduced by the *B5.54 Standards* are the telescopic ballbar measurement, spindle dynamic and thermal drift measurement, laser diagonal measurement, and the 1-day test.

Instrumentation and Metrology

Telescopic Ballbar. The telescopic ballbar test is gradually becoming one of the most powerful and convenient tests for CNC machining center evaluations. As is shown in Figure 13.3.8, a telescopic ballbar consists of a spring-suspended reed having two spheres attached to both ends. The spheres are allowed to move relative to each other with a limited travel, e.g., 3 to 4 mm. Inside the reed is a displacement sensor, known as an LVDT (linearly variable displacement transducer), that measures the relative move of the spheres. The output of the LVDT is connected to a computer.

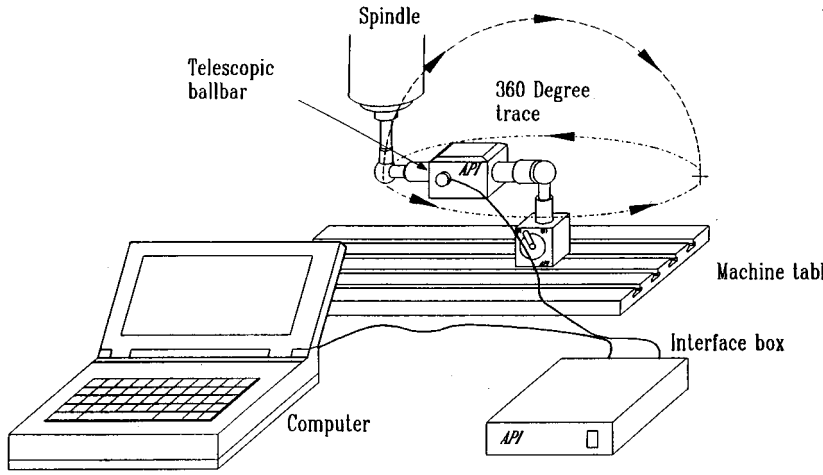


FIGURE 13.3.8 The telescopic ballbar.

To use the telescopic ballbar, one of spheres is magnetically attached to the magnetic socket which is installed on the machine spindle; the other is magnetically attached to a second magnetic socket mounted on the table. The machine is then programmed to contour in a circle in the *X-Y, X-Z,* or *Y-Z* plane, as shown in Figure 13.3.8. As the contouring proceeds, any errors of the machine in contouring will be translated into relative moves picked up by the LVDT and instantaneously recorded by the computer. After the machine has stopped, the computer then plots the errors on a chart for analysis. Errors such as backlash, axes nonsquareness, scale mismatch, servo-gain mismatch, cyclic motion, stick-slip, and certain straightness errors can be readily detected by using a telescopic ballbar.

Laser Interferometer. A laser interferometer system can produce measurements pertaining to the linear, angular, straightness, and squareness errors of a machine system. To many in the manufacturing industry, laser interferometer measurements are considered as the primary calibration standards. However, it is important that when using a laser interferometer, the proper procedures are followed. Precautions should

be taken when using a laser interferometer system: avoid setting up the laser in areas where large air turbulence is present; set up the laser to minimize the cosine error and Abbe error; beware of the environmental effects (i.e., temperature, pressure, and humidity) to the laser accuracy; and understand the material parameters (i.e., material temperature, coefficient of expansion, etc.) to the measuring accuracy. In the event that these effects are significant, it is necessary to compensate for them in the final results.

Setting up a laser interferometer system for machine measurement can be very tedious and sometimes frustrating. It is not unusual to take 2 to 3 days to measure all linear, straightness, angular, and squareness of a mid-size three-axis CNC machine (referred to as 21 parameters measurement). A new laser system known as the 5-D laser interferometer system, developed by Automated Precision, Inc. (Gaithersburg, MD), has the capability of measuring five degrees of freedom (i.e., *X*-, *dY, dZ*, pitch, and yaw) simultaneously in one single setup. In contrast, the 5-D system is able to cut down the measuring time to 3 to 4 hr.

A laser interferometer system should not be confused with an alignment laser system. A laser interferometer system works on the polarized light interference principle and can produce precision to within a tenth (or better) of a wavelength of light (i.e., 630 nm for an HeNe laser). By using the same measuring principle, a laser interferometer system can also generate angle measurements to within 0.1 arc-sec. An alignment laser works on beam-pointing effects and combines the measurements with a photodetector for straightness and angle measurements (not linear measurement). In this case, the laser beam is used as a straightedge. Unless the electro-optics elements are used carefully, the accuracy is much lower than that of a laser interferometer system. Although alignment lasers are commonly used in machine alignment, it should not be construed as a precision calibration standard.

Spindle Analyzer. A spindle analyzer can be used to measure several important parameters of a machining system. These include thermal growth of a machine resulting from environmental variations (i.e., temperature, vibration, etc.) and internal heat sources (i.e., motors, spindle bearings, hydraulic systems, etc.), spindle dynamic errors resulting from a worn or contaminated bearing, and machine axis repeatability.

For advanced applications, it is recommended that the sensors be of noncontact nature since contact sensors can create problem when measuring a high-speed spindle in motion. The noncontact sensors can be optical, capacitive, or inductive. They all have their advantages and disadvantages dependent on their applications. Users should consult with the manufacturers before making their selections.

Other Metrology Instruments. The above-mentioned metrology instruments represent some of the latest and most commonly used instruments in the manufacturing industry. Other instruments that are also used frequently are electronic autocollimators, electronic levels, force gauges, temperature sensors, vibration sensors, dial indicators, proximity sensors, mechanical straightedges, step gauges, precision-indexing tables, etc. One should not, however, overlook the benefits offered by some of the latest data acquisition and data analysis software packages such as the SPC (statistical process control). Of course, the proper selection and use of any instrument are keys to a better understanding of the factors in precision manufacturing.

Inspection System and Metrology

The direct-computer-controlled coordinate measuring machine (DCC-CMM) has been the dominating means for final workpiece dimensional inspection in the manufacturing industry since the early 1980s. Since then, several advanced dimensional measuring devices have also been developed and are gaining wide acceptance in industry. These are the high-speed laser tracking systems and the manually operated measuring robots. There is also a growing interest in the manufacturing industry to reduce the off-line inspection process (i.e., CMM-type applications) by performing some of the inspections on the CNC machining system. This is referred to as on-machine gauging. This section will discuss some of the key considerations related to the use of these advanced inspection systems.

Laser Tracking Interferometer System. The laser tracking interferometer system (LTS) was developed at the National Institute of Standards and Technology (NIST) in the mid-1980s for medium- to large-dimensional inspections. It was then commercialized and introduced to the industry in late 1989. Since then, the LTS has been gaining popularity and is becoming one of the dimensional measuring standards.

Through the combination of a precision dual-axis gimbal, a laser interferometer system and an optical target (i.e., a retroreflector), the laser beam is precisely directed to the target through the manipulation of the gimballed mirror via the control computer. As the beam is sent back to the laser by the target, it is partially deflected to a dual-axis photodetector. The beam position is then interpreted by the computer. As the target moves, the photodetector immediately generates an error signal to the computer which drives the mirror to ensure the beam stays locked onto the target. While it is tracking, the computer acquires the laser measurement and the two angle measurements (*a* and *b* angles) of the mirror and computes for the three-dimensional position of the target.

The advantages of the LTS are that (1) it is considered one of the most accurate large-dimensional coordinate measuring devices with an accuracy of better than 10 ppm (i.e., 100 mm at 10 m); (2) it has a very measurable envelope — 25 m × 360°; (3) it can track a target moving at a speed of 4 m/sec; (4) it can sample up to 1000 samples/sec; and (5) it is compact, portable, and fully automatic. It is well suited for rapid surface scanning, jigs and fixture alignment, replacing conventional CMMs for large structural measurements, and for large CMMs, robotic devices, or CNC machining center calibrations.

The disadvantages are that (1) when used in areas where significant air turbulence is present, the system becomes less reliable; (2) the accuracy may be reduced when used in areas where heavy forklifting activities are present (since the floor foundation is part of the measuring frame); and (3) the system needs zero referencing if the interferometer beam is interrupted during measurement.

On-Machine Gauging. On-machine gauging (or in-process gauging) refers to the dimensional inspection process implemented during or after a machining cycle right on the machining system. In other words, the machining system, retrofitted with a sensor (i.e., a touch probe), is used to serve as a coordinate measuring machine while the workpiece is still on the machine. This concept certainly has merit since most CNC machining systems have control, scale, and structural integrity equal to or better than many DCC-CMMs.

The advantages of performing dimensional inspections on the same machining system are obvious: (1) it eliminates the need of moving the workpiece to a CMM; (2) it reduces the inspection cycle time; and (3) it eliminates realignment error should reworking the workpiece become necessary. The disadvantages, however, are (1) since the same machine frame is used for machining and inspection, if the machining system has any inherent errors (such as geometric errors) which affect the workpiece accuracy, it is incapable of detecting those errors in the inspection and (2) the machine will experience a large degree of thermal distortion caused by the internal heats.

In order to implement on-machine inspection, it is therefore necessary to first perform geometric accuracy evaluation of the machining system according to B5.54 Standards. All efforts should be made to ensure that the accuracy is maintained. Second, a machine thermal growth analysis should be performed to ascertain the limitation of the thermal distortion in inspection. In either case, a minimum rule of thumb of four times the accuracy tolerance should be applied. New techniques of thermal and geometric modeling and compensation of the machining system can be considered in order to achieve the inspection goals.

Other Inspection Systems. Other emerging inspection machines, such as the manually operated robotic measuring device, stereotriangulation measuring systems, and photogrammetry systems, are also gaining popularity in manufacturing. Although they may not offer the same types of accuracy and versatility as the previously mentioned system, they feature a new trend of inspection requirements — portability, agility, shop-floor hardiness, high-speed data acquisition, low cost, and powerful software capability.

References

ANSI. 1992a. Performance Evaluation of Computer Numerically Controlled Machining Centers, ANSI/ASME B5.54-1992, ASME, New York.

ANSI. 1992b. Axes of Rotation, Methods for Specifying and Testing, ANSI/ASME B89.3.4M-1985 (R1995), ASME, New York.

ANSI. 1995. Temperature and Humidity Environment for Dimensional Measurement, ANSI/ASME B89.6.2-1973 (R1995), ASME, New York.

ANSI. 1997. Methods for Performance Evaluation of Coordinate Measuring Machines, ANSI/ASME B89.4.1-1997, ASME, New York.

Slocum, A. 1992. *Precision Machine Design*, Prentice-Hall, Englewood Cliffs, NJ.

Technical Manual, Automated Precision, Inc. (API), Gaithersburg, MD.

Mechatronics in Manufacturing

Tai-Ran Hsu

Introduction

Mechatronics can be defined as: "A technology that involves the design, manufacture and production of *intelligent* products or engineering systems involving *mechanical* and *electronic* functions." Mechatronics is a melding of two English words, *mechanical* and *electronics*. This terminology was first used by Yaskawa Electronic Corp. in Japan during the 1970s. The original notion of mechatronics involved the development and automated production of consumer products such as the Canon SLR autofocus camera (Gilbert, 1992). The application of this technology was soon extended to many other consumer electronic products that included video cassette recorders and the well-known Sony Walkman radios. The rapid advances of microprocessor and microcomputer technologies in the 1980s have broadened the applications of mechatronics to all smart products and systems, ranging from common consumer products to highly sophisticated space engineering equipment. In general, mechatronics systems engineering comprises a number of engineering disciplines. Figure 13.3.9 illustrates interactions of the mechatronics systems engineering with other engineering fields.

Elements of Mechatronic Systems Engineering

Most mechatronic products or systems consist of the following three modules:

1. *The sensing and control module:* This module involves hardware that includes electronic elements, electronic circuits, and the software that provides the commands and operation logics to all components in an automatic control systems. Basic hardware components include *sensors* (position, velocity, acceleration) including tactile, optical, and voice; *actuators* (linear, rotary, voice) driven by electrical, pneumatic, or hydraulic means; *drives* (DC/AC stepper or servomotors); *encoders* (optical or magnetic) and time counters. Other major components include signal processors, amplifiers, power suppliers, and A/D or D/A converters. Control software such as PLC, PID, and other application software are written in C, C^{++}, LISP, FORTH, and RPL programming languages.
2. *Mechanical systems and kinematic linkages module:* The mechanical components in a mechatronic system provide the means to achieve the desired objectives of the product or the system. Principal mechanical components include gears, cams, chains, levers, shafts, pulleys, couplers, bearings, joints, and fasteners. These components are synthesized to construct mechanisms that perform the desired motion of end effects with accurate paths, positions, velocity, and accelerations.
3. *The microprocessors and computer interface module:* Microprocessors act as the brain of the mechatronic product or system. Principal functions of a microprocessor are to manage various instructions to all components such as transducers and actuators in the system, as well as I/O

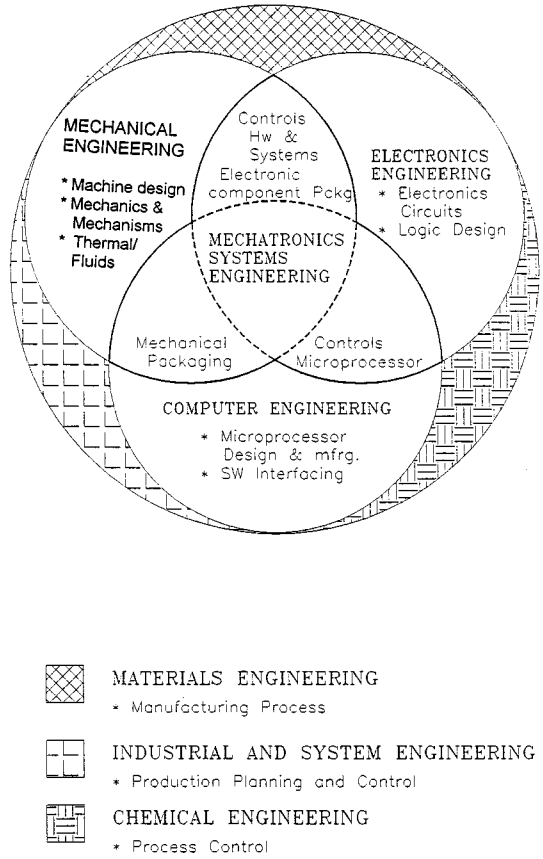


FIGURE 13.3.9 The relationship of various modules in a typical mechatronic device.

interfacing. These processors can also accept function commands from a microcomputer with specific control and applications software.

A block diagram that illustrates the relationship of various modules in a typical mechatronic system is presented in Figure 13.3.10. One may imagine that the microprocessor and microcontroller module functions as the brain, and the sensing/control module and the mechanical systems module each function like the respective sensor-nerve system and the body and limbs in the human anatomy.

References

Gilbert, M.M. 1992. Camera design is something to shoot for, *Mach. Des.*, March.
Sze, S.M. 1984. *VLSI Technology*, McGraw-Hill International Book Company, Singapore.
White, R.M. 1985. *Introduction to Magnetic Recording*, IEEE Press, New York, 72–79.
White, R. 1994. *How Computers Work*, Ziff-Davis Press, Emeryille, CA, 1994.

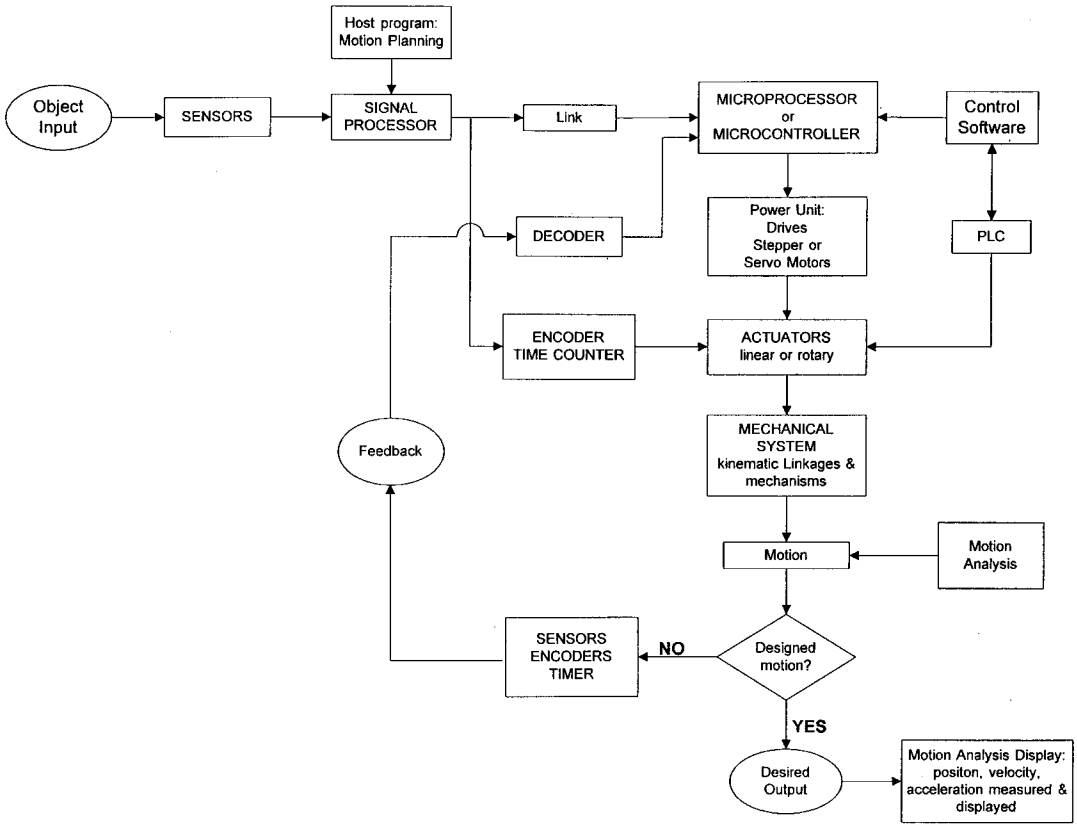


FIGURE 13.3.10 Relationship among modules of a typical mechatronic system.

13.4 Design and Analysis Tools in Manufacturing

Computer-Aided Design Tools for Manufacturing

David C. Anderson

Introduction

Computer-aided design (CAD) tools for manufacturing are computer programs that evaluate producibility of a product under development using computer models of the product and simulation models for manufacturing processes. Examples of such processes are assembly, casting, molding, stamping, forming, and inspection. The early stages of design are critical to the success of the final product since most of the final cost of a product is committed through decisions about the product's geometry and the materials that lead to the selection and planning of manufacturing processes and production facilities.

Concurrent engineering (CE) is a methodology for product development processes where the tasks are performed simultaneously based on "what if" decision-making processes. CE is a driving force behind the development of CAD tools for manufacturing.

CAD System and Manufacturing

One of the first and most prominent manufacturing applications of CAD system was the automated programming of numerically controlled (NC) machine tools. NC machines were developed in the 1950s, but their widespread use was hindered by the difficulty in writing the requisite NC programs, lists of coded cutting tool motions that directed the machine to cut the desired part. Computer programs that facilitated the generation of NC programs were among the first CAD tools for manufacturing. Early systems focused on profile cutting, generating cutter location data that described the 2D coordinates of the paths of cutting tools that would remove the material to create the final part. Three-dimensional NC programming capabilities are available in most CAD systems today. The user guides the generation of NC data by interactively selecting each surface to be machined and answering questions about tools, approach directions, and machining preferences.

Solid Models and Manufacturing

A key element of virtually all CAD tools for manufacturing is the need to interpret the CAD data according to manufacturing capabilities, requirements, and constraints. Generally, this interpretation involves determining the characteristic shapes in the CAD data related to the manufacturing process of interest and applying knowledge about the process to determine the manufacturing operations and parameters. Solid models are rigorous computer data structures that contain a complete, unambiguous representation of the nominal geometry of an object. Solid models and solid modeling operations, like the Boolean union, difference, and intersection of solids, enabled geometric computations about the design that were not possible with earlier CAD data. With solid modeling, the emphasis in CAD shifted from data for visual communication to product data models that could be used for more sophisticated computer analyses with more automation.

In solid models, geometric features can be described in a form more suitable for engineering and manufacturing applications. As a result, feature-recognition algorithms have become an important element of many CAD tools for manufacturing for translating CAD models into usable geometric data for design evaluations.

However, feature recognition has limitations because it is not possible to translate every solid model into a given set of features. Also, some applications of feature recognition require additional, nongeometric information. This information must be added to the CAD data before, during, or after the feature-recognition process. For example, part tolerance data are required for machining process planning but are not available in current CAD data. Feature-based design (FBD) was developed to overcome some of the limitations of feature recognition. Instead of deriving a features model from a solid model, FBD systems create the product model with features during the design process. Many FBD applications have

been demonstrated in the area of machining, and some commercial CAD vendors have incorporated these features into their systems.

FBD also has limitations. The features that are useful for one manufacturing process may not be useful for another. For example, a model made with cavity features, such as holes, slots, and pockets, provides ready-to-use data for machining planning. However, this model is inappropriate for sheet metal bending or welding processes that require an understanding of the bends and protrusions on a part, not its cavities.

Product Data Standards and Manufacturing

The International Standards Organization (ISO) has developed the Standard for the Exchange of Product model data, STEP. In the U.S., the Integrated Graphics Exchange Specification (IGES)/PDES Organization (IPO) developed the Product Data Exchange Specification, PDES. These efforts merged and PDES was renamed Product Data Exchange using STEP, which is now the American National Standard for STEP. STEP became the international standard (ISO 10303) in March 1994.

STEP is organized as a series of "parts" that are developed and published separately. The parts are organized into numerical series: description methods (parts 11 to 20), integrated resources (parts 41 to 200), application protocols (parts 201 to 1200), abstract test suites (parts 1201 to 2200), implementation methods (parts 21 to 30), and conformance testing (parts 31 to 40). The product information is specified in a formal specification language, EXPRESS, which is documented in Part 11. Part 1 is an overview of the entire standard.

The STEP application protocols (APs) are important to CAD tools for manufacturing. These are the implementable portion of the STEP standard. Each AP draws upon integrated resources and adds specific constraints, relationships, and attributes to meet the information requirements of a particular application. It is expected that several hundred APs may be developed for many different industrial applications.

Design for "X" Tools

Many CAD tools for manufacturing belong to a class of programs described as "design for x," where "x" signifies an application area, such as design for assembly or design for castability. The phrase "design for manufacturability" (DFM) can be considered as a specialization of design for x in which all the applications are manufacturing. Generally, DFM applications are computer programs that perform computations to analyze the producibility of a product with respect to a specific manufacturing process or set of processes. The format of the design data representing the product required by the program varies greatly. The program then provides an evaluation of the suitability of the design according to this domain. The form of this evaluation also varies with each program.

DFM programs act as "manufacturing experts" that provide qualitative, and perhaps quantitative, information about potential problems in a design based on predefined knowledge about a manufacturing process. The programs emulate the process done by human experts, examining the design data and reporting any problems based on experience. Many efforts are based on expert systems technology from the field of artificial intelligence (AI). The searching is performed through the facilities of an AI language, such as PROLOG or LISP, or using an expert system "shell," a preprogrammed generic expert system. Design data are first translated into a knowledge base, the AI version of a database, containing facts and rules. In a sense, the detailed design data are made into logical data that can be processed using AI methods. The program computationally searches the design data for data patterns that match problem conditions represented in the knowledge base of the program. The problem conditions are computer representations of design data that are known to cause manufacturing difficulties for a given process. For example, a "design for injection molding" program may report that an internal corner radius in a geometric model of a part is too small and may cause problems in the mold. In some cases, the program provides a quantitative evaluation of the design, or producibility index. This provides a convenient numerical comparison between two competing designs.

References

Amirouche, F.M.L., 1993. *Computer-Aided Design and Manufacturing*, Prentice-Hall, Englewood Cliffs, NJ.

Boothroyd, G. 1994. Product design for manufacture and assembly, *Comput. Aided Des.*, 26(7), 505–520.

Laurance, N. 1994. A high-level view of STEP, *Manuf. Rev.*, 7(1), 39–46.

The National Product Data Exchange Resource Center, U.S. Product Data Association (US PRO) National Institute of Standards and Technology, "http://elib.cme.nist.gov/nipde/" (world-wide web document), 1995.

Whitney, D.E., Nevins, J.L., and De Fazio, T.L. 1989. *Concurrent Design of Products and Processes: A Strategy for the Next Generation in Manufacturing*, McGraw-Hill, New York.

Zeid, I. 1991. *CAD/CAM Theory and Practice*, McGraw-Hill, New York.

Tools for Manufacturing Process Planning

Tien-Chien Chang

Introduction

Process planning prepares a production documentation which specifies the operations and operation sequence necessary to manufacture a product. Process planning is defined as an act that determines the manufacturing operations, operation sequence, and resources required to make a product. In the process domain, there are machining process planning, welding process planning, EDM process planning, forming process planning, etc. In the product domain, there are mechanical part process planning, mechanical assembly process planning, and electronics assembly process planning. The input to a process planning system (or a human process planner) can be an engineering drawing, a CAD model, or a three-dimensional solid model. While most human planners prefer an engineering drawing on paper or in electronic form, process planning systems usually use CAD models.

The result of the process-planning activity is a "process plan" (see Figure 13.4.1), also called route sheet, operation sheet, or operation planning summary. It can be as aggregate as a list of work center identification numbers or as elaborate as a 50-page document with setup drawings, tool specifications, operation time estimates, etc.

PROCESS PLAN — ACE Inc.

Part No. S0125-F — Material: steel 4340Si
Part Name: Housing
Original: S.D. Smart Date: 1/1/89 — Changes: _____ Date:_____
Checked: C.S. Good Date: 2/1/89 — Approved: T.C. Chang Date: 2/14/89

No.	Operation Description	Workstation	Setup	Tool	Time (Min)
10	Mill bottom surface1	MILL01	see attach#1 for illustration	Face mill 6 teeth/4" dia	3 setup 5 machining
20	Mill top surface	MILL01	see attach#1	Face mill 6 teeth/4" dia	2 setup 6 machining
30	Drill 4 holes	DRL02	set on surface1	twist drill 1/2" dia 2" long	2 setup 3 machining

FIGURE 13.4.1 A process plan.

Since the information on the process plan is used in scheduling the production and controlling the machine, the production efficiency and the product quality are affected.

Manual Process Planning

Process planning involves several or all of the following activities:

- Selection of machining operations
- Sequencing of machining operations
- Selection of cutting tools
- Selection of machine tools
- Determination of setup requirements
- Calculations of cutting parameters
- Planning tool path and generation of NC part programs
- Design of jigs and fixtures

The details incorporated in a typical process plan usually vary from industry to industry. It depends on the type of parts, production methods, and documentation needs. A process plan for a tool room–type manufacturing environment typically relies on the experience of the machinist and does not have to be written in any great detail. In fact, the instruction "make as per part print" may suffice. In typical mass-production-type industries, the process-planning activity is embodied in the transfer and flow lines used for manufacturing component parts and assembly. For metal-forming-type manufacturing activities, such as forging, stamping, die casting, sand casting, injection molding, etc., the process-planning requirements are embedded directly into the design of the die/mold used, where most process-planning activity is fairly simple. A process planner must

- Be able to understand and analyze part requirements
- Have extensive knowledge of machine tools, cutting tools, and their capabilities
- Understand the interactions between the part, manufacturing, quality, and cost
- Possess analytical capabilities

Tolerance Charting

During process planning it is important to ensure that the setup and operation sequence will yield a satisfactory part. Tolerancing charting (Figure 13.4.2) has been used to help in allocating process tolerances and verifying the operation sequence. A tolerance chart analyzes one dimension at a time. In a tolerance chart, the top is the part drawing. Dimensions and tolerances are presented with the geometry. Dashed lines show the stock boundary. From the features of the part and the stock, extension lines are drawn to the body of the chart. The section below the drawing shows the critical dimensions and tolerances. These dimensions and tolerances must be satisfied after the processes are complete. Following the process sequence, each operation is listed in the third section of the chart. A line is drawn from the reference surface of a setup to the cut surface. For example, in operation 10, the raw stock boundary at the left is the reference surface. The second surface from the right-hand side is created by this operation. From the chart, one can calculate the resultant tolerances. The results are compared with the blueprint tolerance.

Although traditionally a tolerance chart is implemented on paper and through a fixed procedure, it can also be implemented in a computer. The process tolerance stack-up may be used to verify the design specification and select the appropriate processes and sequences.

Computer-Aided Process Planning

There are two basic approaches to computer-aided process planning — variant and generative. The variant approach is signified by the terminology used by the computer to retrieve plans for similar

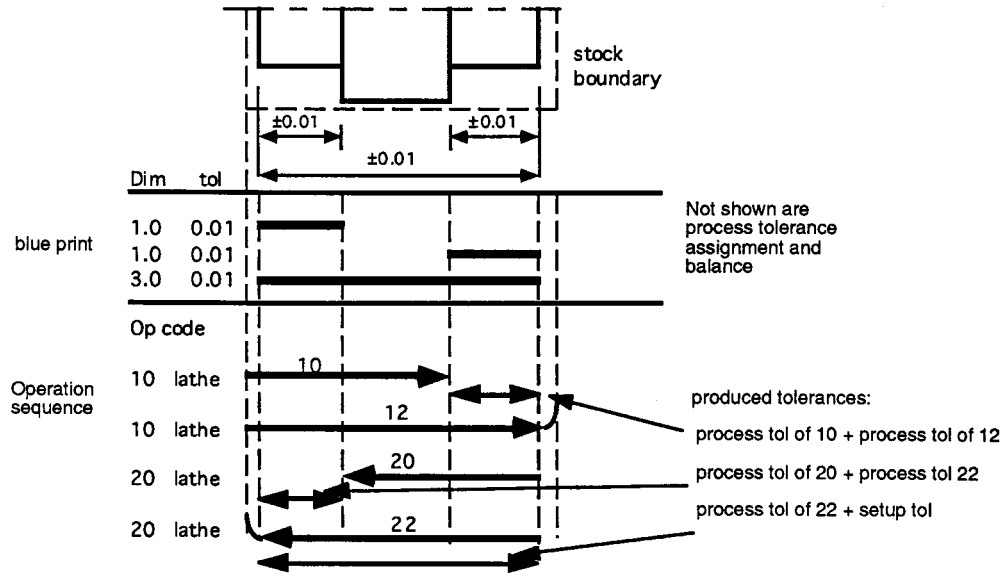


FIGURE 13.4.2 Tolerance chart.

components using table lookup procedures. The human process planner then edits the plan to create a "variant" to suit the specific requirements of the component being planned. Creation and modification of standard plans are the process planner's responsibility. The generative approach generates a plan for each component without referring to existing plans. Generative-type systems can perform many functions in a generative manner, while the remaining functions are performed with the use of humans in the planning loop.

Variant Process Planning

The variant approach to process planning was the first approach used to computerize the planning techniques. It is based on the concept that similar parts will have similar process plans. The computer can be used as a tool to assist in the identification of similar plans, retrieving them and editing the plans to suit the requirements for specific parts.

In order to implement such a concept, part coding and classification based on group technology is used as a foundation. Individual parts are coded based upon several characteristics and attributes. Part families are created of "like" parts having sufficiently common attributes to group them into a family. This family formation is determined by analyzing the codes of the part spectrum. A "standard" plan consisting of a process plan to manufacture the entire family is created and stored for each part family. The development of a variant-process-planning system has two stages: the preparatory stage and the production stage (Figure 13.4.3).

During the preparatory stage, existing components are coded, classified, and later grouped into families. The part family formation can be performed in several ways. Families can be formed based on geometric shapes or process similarities. Several methods can be used to form these groupings. A simple approach would be to compare the similarity of the part code with other part codes. Since similar parts will have similar code characteristics, a logic which compares part of the code or the entire code can be used to determine similarity between parts.

Families can often be described by a set of family matrices. Each family has a binary matrix with a column for each digit in the code and a row for each value a code digit can have. A nonzero entry in the matrix indicates that the particular digit can have the value of that row, e.g., entry (3,2) equals one

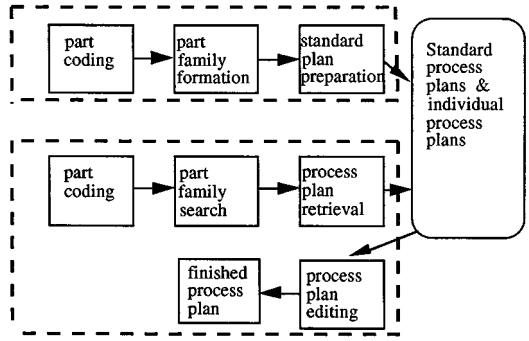


FIGURE 13.4.3 Variant-process-planning approach.

implies that a code x3xxx can be a member of the family. Since the processes of all family members are similar, a standard plan can be assigned to the family. The standard plan is structured and stored in a coded manner using operation codes (OP-codes). An OP-code represents a series of operations on one machine/workstation. For example, an OP-code DRL10 may represent the sequence center drill, change drill, drill hole, change to reamer, and ream hole. A series of OP-codes constitute the representation of the standard process plan.

Before the system can be of any use, coding, classification, family formation, and standard plan preparation must be completed. The effectiveness and performance of the variant-process-planning system depend to a very large extent on the effort put forth at this stage. The preparatory stage is a very time-consuming process.

The production stage occurs when the system is ready for production. New components can be planned in this stage. An incoming component is first coded. The code is then sent to a part family search routine to find the family to which it belongs. Since the standard plan is indexed by family number, the standard plan can be easily retrieved from the database. The standard plan is designed for the entire family rather than for a specific component; thus, editing the plan is unavoidable.

Variant-process-planning systems are relatively easy to build. However, several problems are associated with them., e.g.,

1. The components to be planned are limited to previously planned similar components.
2. Experienced process planners are still required to modify the standard plan for the specific component.
3. Details of the plan cannot be generated.
4. Variant planning cannot be used in an entirely automated manufacturing system, without additional process planning.

Despite these problems, the variant approach is an effective method, especially when the primary objective is to improve the current practice of process planning. In most batch-manufacturing industries, where similar components are produced repetitively, a variant system can improve the planning efficiency dramatically. Some other advantages of variant process planning are

1. Once a standard plan has been written, a variety of components can be planned.
2. Programming and installation are comparatively simple.
3. The system is understandable, and the planner has control of the final plan.
4. It is easy to learn and easy to use.

Generative Approach

Generative process planning is the second type of computer-aided process planning. It can be concisely defined as a system which automatically synthesizes a process plan for a new component. The generative approach envisions the creation of a process plan from information available in a manufacturing database

without human intervention. Upon receiving the design model, the system is able to generate the required operations and operation sequence for the component.

Knowledge of manufacturing has to be captured and encoded into computer programs. By applying decision logic, a process planner's decision-making process can be imitated. Other planning functions, such as machine selection, tool selection, process optimization, etc., can also be automated using generative planning techniques.

A generative-process-planning system comprises three main components:

1. Part description
2. Manufacturing databases
3. Decision-making logic and algorithms

The definition of generative process planning used in industry today is somewhat relaxed. Thus, systems which contain some decision-making capability on process selection are called generative systems. Some of the so-called generative systems use a decision tree to retrieve a standard plan. Generative process planning is regarded as more advanced than variant process planning. Ideally, a generative-process-planning system is a turnkey system with all the decision logic built in. However, due to the differences among manufacturing shops, decision logics have to be customized for each shop.

The generative-process-planning approach has the following advantages:

1. Consistent process plans can be generated rapidly.
2. New components can be planned as easily as existing components.
3. It has potential for integrating with an automated manufacturing facility to provide detailed control information.

There is no fixed representation or procedure that can be identified with generative process planning. The general trend is to use a solid model CAD-based input and expert system or an object-oriented planner construct. Most of the research systems are of this type. A few commercial products can also be classified as generative.

Conclusions

Process planning is a critical function in design and manufacturing. The quality of the product and the cost of production are affected by the process plan. A process plan incorporates information on the shop capability, resource requirement, best routing, etc. In order to produce a good process plan, a planner must be knowledgeable in both the manufacturing practices and the current shop status and capabilities. At this moment most computerized planners are still based on data retrieval and database lookup. As the information technology is further developed and we have better understanding of process capabilities, integrated planning systems will evolve.

References

Chang, T.C. and Wysk, R.A. 1995. *An Introduction to Computer-Aided Process Planning Systems*, Prentice-Hall, Englewood Cliffs, NJ.

Curtis, M.A. 1988. *Process Planning*, John Wiley & Sons, New York.

Halevi, G. and Weill, R.D. 1995. *Principles of Process Planning*, Chapman & Hall, New York.

Kambhampati, S., Cutkosky, M.R., Tennenbaum, J.M., and Lee, S. 1993. Integrating general purpose planners and special reasoners: case study of a hybrid planning architecture, *IEEE Trans. Syst., Man Cybernetics*, 23(6), 1503–1518.

van 't Erve, A.H., 1992. Generative Computer Aided Process Planning for Part Manufacturing, An Expert System Approach, Ph.D. Thesis, Department of Mechanical Engineering, University of Twente, Netherlands.

van Houten, F.J.A.M. 1991. PART: A Computer Aided Process Planning System, Ph.D. Thesis, Department of Mechanical Engineering, University of Twente.

Wang, H.-P. and Li, J.K. 1991. *Computer-Aided Process Planning*, Elsevier, New York.
Zhang, H.-C. and Alting, L. 1994. *Computerized Manufacturing Process Planning Systems*, Chapman and Hall, New York.

Simulation Tools for Manufacturing

Hank Grant

Introduction

Digital simulation uses a mathematical model to represent a real or hypothetical physical system. A computer simulation model of a physical system provides a laboratory in which alternative designs can be explored and analyzed. The model, executed on the computer, is a software replica of the manufacturing system and is controlled so that the behavior of the system can be studied and analyzed. Decisions can be made concerning production alternatives. For example, adding a new lathe can be considered without disrupting the actual physical system.

Simulation depends on describing a system in terms acceptable to the computer language used. To do this, it is necessary to have a *system-state description*, which is typically characterized by a set of state variables included in the computer program that make up the simulation model.

Types of Simulation Models

Simulation models are typically classified into three types: discrete event, continuous, and combined. Simulation software has been designed to address each of these types of models.

Discrete Event. Discrete event simulation is used to model systems when there are specific events in time when the variables of the system may change in values. The mechanics of those changes must be well known and easily characterized. The behavior of the system is represented by the behavior of individual objects of interest called *entities*. The simulation model characterizes the behavior of these entities as they move through the system in simulated time.

Discrete events are points in time where the characteristics of an entity may change and where the state variables of the system may change. For example, when a customer arrives for service, the state of the system may change (number in the system, status of the server, etc.). The modeling of systems using this approach consists of developing descriptions of the events and how they cause the state variables to change and the entities to be manipulated.

The individual events may not always be predictable, and stochastic elements may be present in the operation of the system. For example, the time between arrivals of customers to the system may be a random variable. Simulation languages have many tools to support random variation in models.

A special kind of discrete event model is a *network model*. Network models use a standard set of symbols to represent the flow of entities in the system. They are graphical in nature, and are very useful communication vehicles as well as very powerful in building simulation models quickly and easily. There are several languages available that include network modeling capabilities and they are described below.

Continuous. Continuous simulation is an approach that is popular among engineers and economists. The main building blocks of this approach are as follows (Pidd, 1994).

Aggregated variables: Instead of a concern with individual entities, the main concern is with the aggregated behavior of populations. For example, the changing sales of a product through time.

Smooth changes in continuous time rather than focusing on individual events, where the stress is on the gradual changes which happen as time progresses. Thus, just as the graph of a variable might be smooth, the aim is to model the smooth changes of the variable by developing the suitable continuous equations.

Differential or difference equations: The model consists mainly of a set of equations which define how behavior varies through time; thus, these tend to be differential equations or, in simpler cases such as system dynamics, difference equations.

Nature does not present itself labeled neatly as discrete or continuous; both elements occur in reality. Modeling, however, as mentioned above, involves approximation, and the modeler must decide which of these approaches is most useful in achieving the desired aim of the simulation.

Combined Discrete Event/Continuous. In some cases, both approaches are needed and the result is a mixed discrete-continuous simulation. An example of this might be a factory in which there is a cooking process controlled by known physics which is modeled by continuous equations. Also in the factory is a packing line from which discrete pallets of products emerge. To model the factory will require a mixed approach.

Modeling Languages

Specifically designed computer simulation languages provide many features for managing the updating of the state variables and advancing time. They also provide features for recording system performance statistics and for generating random numbers to introduce system randomness.

The lowest level of computer language typically used is FORTRAN or BASIC. This requires that the entire simulation model be coded, which is labor-intensive. High-level languages, such as SLAM, SIMSCRIPT, and GPSS, facilitate simulation because they provide subroutines for time advancement, entity maintenance, and statistic collections. Higher-level simulation languages are designed for special purposes; MAP/1, SPEED, and MAST are three designed for the simulation of manufacturing systems.

Some simulation languages can produce animations. This permits the simulation to be illustrated graphically on a computer terminal so that the analyst can see the system in action and observe its interactions and behavior, a visual function beyond the scope of standard reporting technique. For example, TESS (a software program) provides animation, as well as model-building and output analysis capabilities, for the SLAM simulation language.

The following discussion by Banks (1994) provides an overview of the primary languages available.

Applications of simulation exist in many arenas such as manufacturing, material handling, health services, military decision support, natural resources, public services, transportation, and communications, to mention a few.

These simulation applications are usually accomplished with the use of specially developed simulation software. This tutorial describes the software in two categories. The first of these is software for general purposes. This type of software can solve almost any discrete simulation problem. In this section, five products, GPSS/H™, GPSS/World™, SIMAN V′, SIMSCRIPT II.5′, and SLAMSYSTEM′, will be discussed to provide a feel for this type of software.

GPSS/H. GPSS/H is a product of Wolverine Software Corporation, Annandale, VA (Smith and Crain, 1993). It is a flexible, yet powerful tool for simulation. It provides improvements over GPSS V that had been released by IBM many years earlier. These enhancements include built-in file and screen I/O, use of an arithmetic expression as a block operand, interactive debugger, faster execution, expanded control statement availability, and ampervariables that allow the arithmetic combinations of values used in the simulation. The latest release of GPSS/H is version 2.0. It added a floating point clock, built-in math functions, and built-in random variate generators. Options available include Student GPSS/H, Personal GPSS/H within the 640K memory limit, and GPSS/H 386 providing unlimited model size.

GPSS World. GPSS World, from Minuteman Software, is a complete redesign of GPSS/PC™ (Cox, 1992). It is designed as a high-power environment for simulation professionals. It includes both discrete and continuous simulation. Its features include interactivity, visualizability, and configuration flexibility. It utilizes 32-bit computing, virtual memory, preemptive multitasking, symmetric multiprocessing, and distributed simulation. Highlights include drag-and-drop model building, 512 megabytes of virtual

memory for models, point-and-shoot debugging, an embedded programming language, built-in probability distributions, multiple data types, and many other improvements to GPSS/PC.

The GPSS World family is a set of three software products including:

1. GPSS World is the center of the family. This self-contained modeling environment includes local Simulation Server™ capabilities.
2. Simulation Server provides simulation services on a remote networked computer. It does not include a model-building user network.
3. Simulation Studio provides hierarchical modeling and user-drawn simulation capabilities.

There is an enhanced memory version of GPSS/PC that is also available. It allows access of up to 32 megabytes of memory.

SIMSCRIPT II.5. SIMSCRIPT II.5 from CACI Products Company is a language that allows models to be constructed that are either process oriented or event oriented (Russell, 1993). The microcomputer and workstation version include the SIMGRAPHICS animation and graphics package. SIMSCRIPT can be used to produce both dynamic and static presentation-quality graphics such as histograms, pie charts, bar charts, levels of meters and dials, and time plots of variables. Animation of the simulation output is also constructed using SIMGRAPHICS. SIMGRAPHICS can be used also to produce interactive graphical front ends or forms for entering model input data. An input form may include such graphical elements as menu bars with pull-down menus, text or data boxes, and buttons that are clicked on with a mouse to select an alternative. The graphical model front end allows for a certain set of modifications to the model to be made without programming, facilitating model use by those who are not programmers.

SIMAN V. SIMAN V from Systems Modeling Corporation is a general-purpose program for modeling discrete and/or continuous systems (Glavach and Sturrock, 1993; Banks et al., 1995). The program distinguishes between the system model and the experiment frame. The system model defines components of the environment such as machines, queues, and transporters and their interrelationships. The experiment frame describes the conditions under which the simulation is conducted, including machine capacities and speeds and types of statistics to be collected. “What-if” questions can usually be asked through changing the experiment frame rather than by changing the model definition. Some important aspects of SIMAN V are as follows:

1. Special features that are useful in modeling manufacturing systems include the ability to describe environments as work centers (stations) and the ability to define a sequence for moving entities through the system.
2. Constructs that enable the modeling of material-handling systems including accumulating and nonaccumulating conveyors, transporters, and guided vehicles.
3. An interactive run controller that permits break points, watches, and other execution control procedures.
4. The ARENA environment that includes menu-driven point-and-click procedures for constructing the SIMAN V model and experiment, animation of the model using Cinema, the input processor that assists in fitting distributions to data, and the output processor that can be used to obtain confidence intervals, histograms, correlograms, and so on. (More aspects of the ARENA environment are discussed later.)
5. Portability of the model to all types of computers.

SLAMSYSTEM. SLAMSYSTEM, from Pritsker Corporation, is an integrated simulation system for PCs based on Microsoft Windows™ (Pritsker, 1986; O’Reilly, 1993). All features are accessible through pull-down menus and dialog boxes and are selected from the SLAMSYSTEM Executive Window. A SLAMSYSTEM project consists of one or more scenarios, each of which represents an alternative system configuration. A project maintainer examines the components of the current scenario to determine if any of them have been modified, indicates whether or not tasks such as model translation should be performed,

and allows the user to accomplish these tasks before the next function is requested. SLAMSYSTEM allows multiple tasks to be performed in parallel while the simulation is operating in the background.

Some of the features of SLAMSYSTEM are as follows:

1. Models may be built using a graphical network builder and a forms-oriented control builder, or text editor. When using the first method, a network symbol is selected with the mouse, then a form is completed specifying the parameters for that symbol. The clipboard allows many other operations such as grouping one or more symbols and placing them elsewhere on the network.
2. Output analysis includes a "report browser" that allows alternative text outputs to be compared side by side. Output may be viewed in the form of bar charts, histograms, pie charts, and plots. Output from multiple scenarios can be displayed at the same time in bar chart form. By using the Windows environment, multiple output windows can be opened simultaneously.
3. Animations are created under Windows using the facility builder to design the static background and the script builder to specify which animation actions should occur when a particular simulation event occurs. Animations can be performed either concurrently or in a postprocessing mode. Two screens can be updated simultaneously and up to 225 screens can be swapped into memory during an animation.
4. SLAMSYSTEM was designed to be used in an integrated manner. For example, historic data may be read to drive the simulation. CAD drawings may be loaded. Output charts and plots created by SLAMSYSTEM may be exported via the clipboard to other applications.

The newest release of SLAMSYSTEM is version 4.0. Some of its unique features include the following:

1. Multiple networks in a single scenario: Networks can be constructed in sections and combined at run time. The sections can be reused in future models.
2. New output graphics: These graphics support three-dimensional X-Y grids and displaying of point plot data.
3. Direct interface to SimStat (product of MC^2 Analysis Systems): These files may be loaded for advanced statistical analysis.
4. OS/2 metafiles for graphics: The OS/2 metafile format can be read for animation backgrounds or icons.

Conclusion

Simulation is a powerful approach to modeling manufacturing systems in that many complex and diverse systems can be represented. Simulation can predict system performance measures that are difficult to assess without a model. It is a proven, successful tool and has been in use since the 1950s. The current languages take advantage of the capabilities of today's microprocessors and provide the user with the needed on-line support for model development, management, and analysis.

References

Banks, J. 1994. Simulation software, paper presented as 1994 Winter Simulation Conference, Atlanta.

Banks, J., Burnette, B., Rose, J.D., and H. Kozloski. 1995. *SIMAN V and CINEMA V,* John Wiley & Sons, New York.

Cox, S.W. 1992. Simulation Studio™, in *Proceedings of the 1992 Winter Simulation Conference,* J. J. Swain, D. Goldman, R.C. Crain, and J.R. Wilson, Eds., Association for Computing Machinery, New York, 347–351.

Glavach, M.A. and Sturrock, D.T. 1993. Introduction to SIMAN/Cinema, in *Proceedings of the 1993 Winter Simulation Conference,* G.W. Evans, M. Mollaghasemi, E.C. Russell, and W.E. Biles, Eds., Association for Computing Machinery, New York, 190–192.

O'Reilly, J.J. 1993. Introduction to SLAM II and SLAMSYSTEM, in *Proceedings of the 1993 Winter Simulation Conference*, G.W. Evans, M. Mollaghasemi, E.C. Russell, and W.E. Biles, Eds., Association for Computing Machinery, New York, 179–183.

Pidd, M. 1994. An introduction to computer simulation, 1994 Winter Simulative Conference, The Management School, Lancaster University, U.K.

Pritsker, A.B. 1986. *Introduction to Simulation and SLAM II*, 3rd ed., John Wiley & Sons, New York.

Russell, E.C. 1993. SIMSCRIPT II.5 and SIMGRAPHICS tutorial, in *Proceedings of the 1993 Winter Simulation Conference*, G.W. Evans, M. Mollaghasemi, E.C. Russell, and W.E. Biles, Eds., Association for Computing Machinery, New York, 223–227.

Smith, D.S. and Crain, R.C. 1993. Industrial strength simulation using GPSS/H, in *Proceedings of the 1993 Winter Simulation Conference*, G.W. Evans, M. Mollaghasemi, E.C. Russell, and W.E. Biles, Eds., Association for Computing Machinery, New York, 218–222.

Tools for Intelligent Manufacturing Processes and Systems: Neural Networks, Fuzzy Logic, and Expert Systems

Tien-I. Liu

Introduction

Starting in the 1980s, researchers and practitioners became increasingly interested in intelligent machines and intelligent manufacturing. The goal is to model the skills and expertise of professionals so that machines and manufacturing systems can possess some of the characteristics of human intelligence. Three techniques, neural networks, fuzzy logic, and expert systems, have been widely used in manufacturing. This section describes the principles and functions of these tools. Examples in applying these tools to manufacturing applications are highlighted as well.

Neural Networks

Neural networks consist of a set of nodes which are nonlinear computational elements. The pattern of connectivity between nodes, known as weights, can be modified according to some preset learning rule. The knowledge of the networks is stored in their interconnections (Kohonen, 1986).

Since neural networks are parallel distributed processing, they have the following advantages:

1. They are adaptive and can learn from experience.
2. The network can be refined at any time with the addition of new training data.
3. Various model architectures can be used.
4. They can compute very quickly and thus they are very suitable for real-time applications.
5. They can be used for analyzing large amounts of data to determine patterns that may predict certain types of behavior.
6. They can capture the complexities of the process, including nonlinearities, even if the dynamics of the process is unknown.
7. They can make decisions based upon incomplete and noisy information.
8. They degrade gracefully even when parts of the structure have been destroyed.

Neural networks are best at performing the types of tasks which need human perception. These tasks do not have exact answers, e.g., classification and trend analysis. A typical example of such problems is machine diagnosis. An experienced mechanic can point out what is wrong with an automobile by standing beside the car and listening to the sound of the running engine. Another example is character recognition. Any person who is familiar with alphabets can easily identify the letter "A" in any of various typefaces or handwritten scripts. In both cases it is practically impossible to develop a set of if–then rules to let a computer to do the job. These tasks are also difficult to program using standard computer techniques.

Applications of Neural Networks. Many electronics and computer companies have put considerable effort into neural network development. IBM has announced a neural network development package for its computer; Intel Corporation has developed a microchip that supports this technology. Japanese companies such as Fujitsu, Hitachi, Mitsubishi, and Sumitomo Heavy Industries are also working in this field.

A bomb-detection machine which uses neural network technology to detect plastic explosives hidden in baggage has been developed. This machine has been installed in several airports. Neural networks have also been applied to ensure the operation of an industrial power distribution substation. It has replaced a conventional mechanical system and improved the performance by greater than an order of magnitude.

Integrating sensors with neural networks for monitoring and diagnostic purposes can enhance production reliability, prevent potential problems caused by abnormal conditions, and maintain high product quality in the factory (Liu and Iyer, 1993). The applications of neural networks for monitoring and diagnostic purposes have been used with ball-and-roller bearings, turning processes, milling processes, drilling processes, tapping processes, glass furnaces, etc. The results are very successful (Liu and Anatharaman, 1994).

Neural networks have also been used in the image processing for computer vision and speech-recognition systems (Badal, 1993). They also are used for the control of machines and processes. The neural network controller is capable of on-line learning of the system dynamics and then taking adequate action to achieve the predetermined goal. They also can be used to tune the gain of control systems.

Fuzzy Logic

The theory of fuzzy sets has been developed as a methodology for the formulation and solution of problems which are too complex or too ill defined to be solved by traditional techniques. In fuzzy logic, the membership in a set is not either 0 or 1; instead it is a value between 0 and 1. Membership functions span some problem domain, such as length or weight, and show the membership for each value of the problem domain. Membership functions are subjective evaluations and can be represented by many kinds of curve. However, the membership function cannot be assigned arbitrarily. The formulation of the membership function should be based upon the professional feeling and physical understanding of the problem. Let $S = \{s\}$ represent a space of objects. Then a fuzzy set X in S is a set of ordered pairs

$$X = \left\{s, f_x(s)\right\}, \quad s \_ S(1)$$

where $f_x(s)$ is the grade of membership of s in X and $f_x(s)$ is a number in the interval (0,1).

Fuzzy mathematics, which consists of precise rules to combine vague expressions, such as “very high” and “somewhat heavy,” has been developed (Kandel, 1986). Generally speaking, fuzzy logic systems have the following advantages:

1. They are inherently flexible.
2. They are robust to noisy or missing data, unexpected disturbances, and errors in problem modeling.
3. They are suitable to deal with problems for which knowledge is approximate or problems which are so complex that it is difficult to develop an adequate mathematical model.
4. They usually are energy efficient.

Applications of Fuzzy Logic. Fuzzy mathematical techniques are very suitable for the control of machine tools, robots, and electronic systems (Mamdani, 1993). They are also applicable to image understanding for computer vision and pattern classification for the monitoring and diagnosis of manufacturing processes (Du et al., 1992).

The very first application of fuzzy logic was the control of the fuel-intake rate and gas flow of a rotating kiln used to produce cement. Since then, it has been used to control many automated manufacturing processes. Fuzzy logic has resulted in significant improvements to many commercial products,

such as cameras and air conditioners. Although fuzzy logic was developed in the U.S., most of the action is in Japan. The most impressive application is a subway system operated by a fuzzy computer. It was installed in the 1980s by Hitachi at Sendai, about 200 mi north of Tokyo, Japan. This system is more than 10% energy efficient and is so smooth that passengers do not need to hang onto straps.

At AT&T Bell Laboratories, Dr. M. Togai and Dr. H. Watanabe developed the very first fuzzy logic processing chip in 1985. NASA is developing fuzzy controllers to help astronauts pilot the space shuttle in earth orbit. In the U.S. the interest in applying fuzzy logic is growing continuously.

Expert Systems

An expert system is a computer system which possesses the capability of a human expert. An expert system has the following five essential parts (Figure 13.4.4):

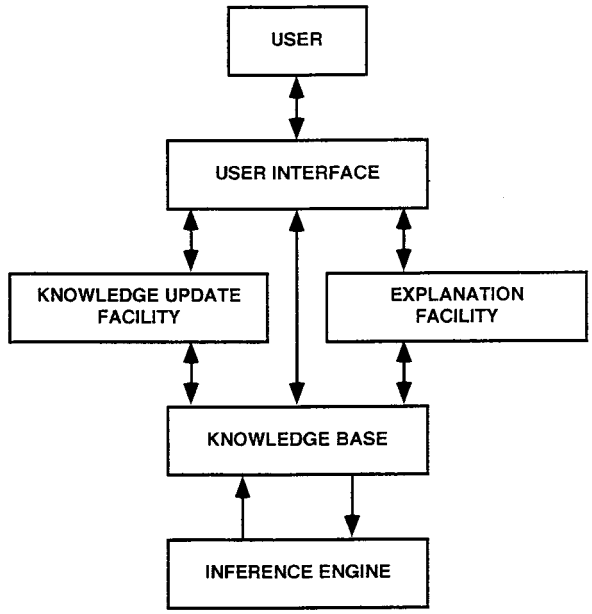


FIGURE 13.4.4 Typical expert system architecture.

1. *User Interface.* This module is the interface between the user and other parts of the expert system.
2. *Knowledge Base.* The knowledge about solving specific problems is stored in the knowledge base of an expert system. The process of building the knowledge base is called knowledge engineering and is done by a knowledge engineer (Figure 13.4.5). Knowledge engineering refers to the acquisition of knowledge from a human expert or other source and the coding of it into the expert system. The knowledge base usually consists of rules and facts. Rules are made up of English-like sentences or clauses. Rules are often defined using an if–then syntax that logically connects one or more antecedent clauses with one or more consequent clauses as follows:

   IF antecedent THEN consequent

   A rule says that if the antecedent is true, then the consequent is also true. The antecedent and consequent of rules refer to a specific fact that describes the state of the world. On the other hand, each fact is a single sentence that describes some aspect of the state of the world.
3. *Inference Engine.* The inference engine can infer new knowledge from existing knowledge stored in the knowledge base. Two general inference approaches are commonly used: forward chaining and backward chaining. Forward chaining is the reasoning from facts to conclusions resulting

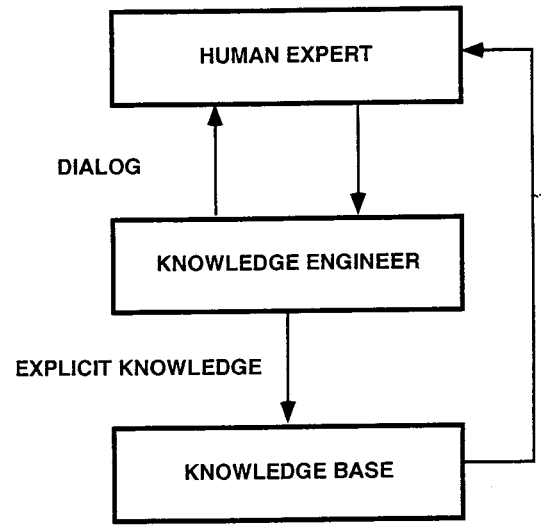


FIGURE 13.4.5 Knowledge engineering.

from those facts. Backward chaining involves reasoning in reverse from a hypothesis to the facts which support the hypothesis.

4. *Knowledge Update Facility.* The knowledge in many fields, including engineering and manufacturing, changes with time. The expert system can be updated through this facility.
5. *Explanation Facility.* Just like a human expert can explain how a specific conclusion has been drawn, the explanation facility can explain its reasoning to enhance the credibility of an expert system.

Expert systems are quite different from conventional programs because the problems usually have no algorithmic solution and rely on inferences to achieve a reasonable solution. In other words, the expert system is very suitable to solve problems which require heuristic rules. These heuristic rules can be stored in the knowledge base.

Expert systems have the following advantages as compared with the human expert:

1. They are steady and unemotional and have high reliability. Therefore, errors are reduced.
2. They have high availability. Users can get answers any time.
3. The expert system can be installed and used at multiple sites.
4. They can explain how a conclusion has been reached. Thus, the users feel comfortable working with expert systems.

Expert systems can help people solve problems. They can free the human expert from the routine job to do other work. Hence, they can increase efficiency and reduce cost.

Applications of Expert Systems. Expert systems have been used in many fields. There are many commercialized expert systems running on different computer platforms which help professionals in various fields, enhancing efficiency and productivity greatly.

Manufacturing, assembly, quality, reliability, and cost need to be taken into consideration in the early stage of product design. Expert systems have been developed for DFM, design for assembly (DFA), design for quality and reliability, product cost estimation, etc. These expert systems can be integrated with the existing CAD/CAM systems (Liu et al., 1995). Many expert systems have also been built and used in the areas of facility design, production planning and control, computer-aided process planning, material handling, quality control, equipment maintenance and repair, and real-time control (Alto et al., 1994). In simple words, design and manufacturing work can be upgraded from an experience-based to a science-based function by using expert systems.

Conclusion

Neural networks, fuzzy logic, and expert systems are the way of the future. They can make machines and manufacturing processes much smarter. Applying these techniques can lead to the realization of a fully automated factory in the future.

References

Alto, A., Dassisti, M., and Galantucci, L.M. 1994. An expert system for reliable tool-replacement policies in metal cutting, *ASME J. Eng. Ind.*, 116(3), 405–406.

Badal, D.Z. 1993. Neural network based object recognition in images, in *Proc. IEEE Int. Conf. Neural Networks*, San Francisco, CA, 1283–1288.

Du, R.X., Elbestawi, M.A., and Li, S. 1992. Tool condition monitoring in turning using fuzzy set theory, *Int. J. Mach. Tools Manuf.*, 32(6), 781–796.

Kandel, A. 1986. *Fuzzy Mathematical Techniques with Applications*, Addison-Wesley, Reading, MA.

Kohonen, T. 1986. An introduction to neural computing, *Neural Networks*, 1, 3–16.

Liu, T.I. and Anatharaman, K.S. 1994. Intelligent classification and measurement of drill wear, *ASME J. Eng. Ind.*, 116(3), 392–397.

Liu, T.I. and Iyer, N.R. 1993. Diagnosis of roller bearings using neural networks, *Int. J. Adv. Manuf. Technol.*, 8(2), 210–215.

Liu, T.I., Yang, X.M., and Kalambur, G.J. 1995. Design for machining using expert system and fuzzy logic approach, *ASM J. Mater. Eng. Performance*, 4(5), 599–609.

Mamdani, E.H. 1993. Twenty years of fuzzy control: experiences gained and lessons learnt, *Proc. IEEE 2nd Int. Conf. Fuzzy Systems*, San Francisco, CA, 339–344.

Tools for Manufacturing Facilities Planning

J. M. A. Tanchoco, Andrew C. Lee, and Su-Hsia Yang

Introduction

The main function of facility planning is the design of efficient flow of products from raw material to finished goods. It is one of the most important determinants of operating efficiency and production cost. Traditionally, the facility-planning problem is divided into three areas, namely, group technology (GT), material handling, and facility layout (see Figure 13.4.6). GT, which is closely related to cellular manufacturing, is usually defined as the grouping of dissimilar machines in close vicinity. Each group or cell is dedicated to the production of one or more parts families. The parts in the family are similar in their processing requirements (Wemmerlov and Hyer, 1989). Two of the most fundamental elements in facility planning are the facility layout and the material-handling system.

Facility layout positions the workstations around the fixed product based on the processing sequence. In a product layout, machines are arranged according to the processing sequence of the product, e.g., the assembly of automobiles, certain electronic products, etc. The machines are located so as to provide smooth and logical flow of material. In a group layout, also referred to as a cellular layout, products are grouped into logical product families. All machines which perform similar operations are grouped together in the process layout. A process layout is characterized by a high degree of flexibility and machine utilization. Regardless of the type of facility, a detailed layout should not be designed without giving serious consideration to material-handling requirements. The choice of material-handling methods and equipment is an integral part of the layout design. The facility layout design component performs two basic functions. The first function is to decide how to locate cells with respect to each other. The objective is to minimize either the total material flow distance or transportation time. The second function is to resolve the machine location and orientation relative to each other within the cell. The constraints in layout design or facility structure could prohibit the placement of the cells or machines in some

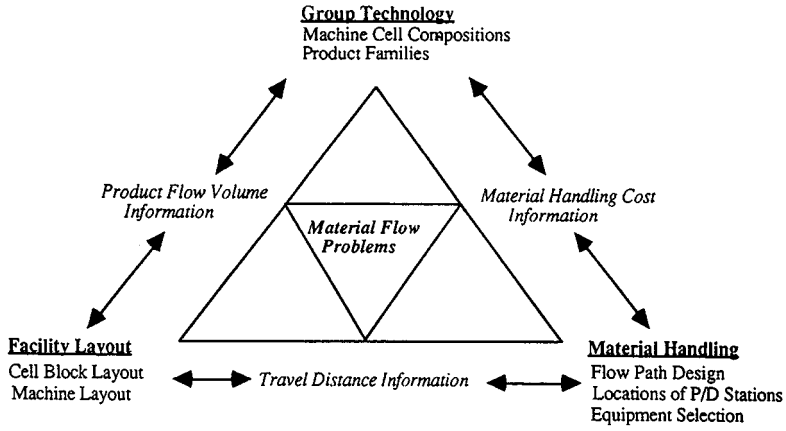


FIGURE 13.4.6 Facilities-planning framework components.

locations, thus changing the machine compositions of the cells. The resolution from this design problem has a significant impact on the distance that material has to be moved.

Material handling performs a critical function in modern dynamic manufacturing systems. The transportation of materials within a production system is often accomplished with limited resources, e.g., conveyors, forklift trucks, or automated guided vehicles (AGVs). Such transfer mechanisms not only deliver material in a timely fashion, but also provide temporary storage capacity. These limited resources require a capital investment, which increases the overall production costs. Excess transportation capability represents expenditures that are not accompanied by increased value. Insufficient capacity, on the other hand, can adversely affect production by delaying the delivery of material, thus reducing the total production volume. Material handling is generally considered as a non-value-adding activitiy. However, one can argue that material handling adds time and space value by making the material available and ready for processing. The material handling has three functions. First, the flow path design selects the route that material transfers from cell to cell or from machine to machine. Limited by the cost of remodeling, the existing plant configuration (such as the aisle width) might not be able to accommodate the traffic flow. Therefore, considerations must be given to the overall facility layout and building structure. The second function is to locate the pickup and delivery stations along the flow path. Along with the facility layout and flow path design, it is one of the most important determinants of the operating cost of material handling. Finally, the best combination of material-handling equipment is determined. Because of the complex, ill-structured, and experience-based nature of the problem, the decisions for choosing material-handling equipment tend not to be based on rigorous criteria. However, the designers must recognize that there is an appropriate level of technology for every application that will meet the combined need for maximum handling efficiency and acceptable cost.

Decision Factors for Facilities Planning

The effectiveness of a facility-planning system depends on the careful integration of GT, facility layout, and the material-handling system.

Group Technology. Mitrofanov (1959) is recognized as the first person to introduce the concept of GT and machine grouping. Since then, a number of researchers have developed techniques for machine grouping and part family formation. In general, these techniques can be categorized as follows:

1. *Machine-Component Matrix* (McAuley, 1972; King, 1980; McCormick et al., 1972): The main idea behind this approach is to delineate cells by grouping binary entries of a machine–component incidence matrix into fuzzy blocks along the matrix diagonal.

2. *Mathematical Programming Formulation* (Kusiak et al. 1986; Askin and Standridge, 1993): In this approach, a large, combinatorial mixed integer programming formulation is usually required to model the problem.
3. *Graph-Based Formulation/Partitioning* (Rajagopalan and Batra, 1975; Faber and Carter 1986): In this approach, the machine–part matrix is represented by a bipartite, transition, or boundary graph. Then, it is solved using graph theoretic clustering methodology.

Facility Layout. The facility layout problem has been formulated as a quadratic assignment problem (QAP). The objective function for this type of formulation is to minimize the total material-handling cost. Given the complexity of the QAP formulation, the size of problems which could be solved by optimal methods is very limited. Thus, heuristic algorithms are more suitable. These heuristics can be classified into two major groups. The first group is called construction algorithms (Seehof and Evans, 1967; Lee and Moore, 1967; Foulds and Giffin, 1985; Hassan and Hogg, 1991). The main idea behind these methods is to build the layout by adding one more block (a cell or a department) to a partial block layout until all blocks have been located. The methodology requires two steps: a selection step and a placement step. The selection step determines the order by which the blocks enter the layout, while the placement step selects the location of the new block to enter the layout relative to the blocks that are already in the layout. The objective is to maximize some kinds of performance criteria, e.g., total closeness rating. Graph theoretic methods have been applied as a solution methodology for this approach.

The second group of the layout heuristics is called improvement algorithms. This approach starts from an initial layout, and improvements are made by successive pairwise interchanges of blocks. The general form of an improvement algorithm consists of the following steps: (1) select a pair of activities, (2) estimate the cost of interchange, (3) exchange if the total cost is reduced, and (4) repeat until no more improvement can be made. This category of heuristics includes the computerized relative allocation of facilities technique (CRAFT) of Armour and Buffa (1963) and Buffa et al. (1964) and the methods of Hillier (1963), Fortenberry and Cox (1985), and Co et al. (1989).

Recognizing the weaknesses of both construction and improvement algorithms when applied separately, Golany and Rosenblatt (1989) proposed a hybrid method that takes advantages of both construction and improvement algorithms. They use the layout resulting from the construction algorithm as the initial layout and improve upon it using an improvement algorithm. Since most of the heuristics for facility layout are based on the "greedy approach," the solution is very sensitive to the initial layout. The final layout given by the algorithm may not be the best.

Material Handling. Material handling involves moving, storing, and controlling the flow of materials. There are several components, such as the flow path design, the locations of pickup and delivery stations, and the material-handling equipment selected, that have significant effects on the overall effectiveness of the material-handling system. The objective of flow path design is to determine the best "street network" that transporters pass through when parts are moved from one machine to the next. It is one of the major determinants in the calculations of travel times, operating expenses, and installation costs of the material-handling system. There are numerous types of flow path network configurations. The most widely adopted flow path design is the conventional flow network. It is usually a unidirectional flow network where any cell boundary is used as part of the flow path. In a unidirectional network, one has to determine the flow direction for each aisle. Its flexibility, reliability, and efficiency have made this type of network a popular choice especially when AGVs are used. In comparison, the potential of using a bidirectional flow network could make the system more efficient (Egbelu and Tanchoco, 1986). The same authors also developed guidelines for the use of a single-lane bidirectional flow network. However, the advanced hardware requirements and complicated system controllers are viewed negatively.

A single-loop flow path network can be found in many flexible manufacturing systems. The entire flow path design is made up of a single loop. This type of network can potentially minimize some of the problems associated with a conventional flow network. Congestion is inherently low and the operating rules are simple. It also has relatively low initial investment and maintenance costs.

Recently, a new flow path network configuration, a segmented flow topology (SFT), was developed by Sinriech and Tanchoco (1994). It is comprised of one or more zones, each of which is separated into nonoverlapping segments. Each segment is serviced by a single bidirectional material-handling device. Transfer buffers are located at both ends of each segment. The flow structure in each zone is determined by the logical flow requirements and by the existing aisle network. The SFT provides a simple flow structure and control system. The research also suggested that it can achieve a higher throughput capability compared to other material flow path network configurations.

The locations of pickup and delivery stations have generally been considered as a secondary issue in the design phase of facility layout and a material-handling system. Yet it can have detrimental effects on the costs of material handling and the a machine layout configuration. In a study by Warnecke et al. (1985), they confirmed that the actual pickup and delivery station flow distance is much more representative than taking the rectilinear distance between the centers of machine blocks. Since then, several design procedures have been proposed to find the optimal location of material-transfer stations. Montreuil and Ratliff (1988) suggested a systematic methodology for locating pickup and delivery stations within a facility layout using multifacility location theory. The objective function is to minimize the sum of the rectilinear distance traveled by all intercellular flows, given the boundary regions on station location. Luxhoj (1991) developed a two-phase design procedure that is suitable for the spine layout where the active flow lines are well defined.

In terms of material-handling equipment selection, there is a large variety of equipment types available with their own special functions and characteristics. Each equipment type has its own capability and limitations. Some of these characteristics are difficult to quantify. The integrated nature of the manufacturing systems complicates the material-handling selection problem. The problem was first addressed by Webster and Reed (1971). The procedure they suggested initially assigns material-handling equipment to departmental moves based on cost alone. Then, move assignments are interchanged to seek improvement in equipment utilization and total cost. Hassan et al. (1985) reformulated Webster and Reed's model as an integer programming model with the objective of minimizing the total operating and capital costs of the selected equipment. Due to the combinatorial nature of the problem, it is solved using a construction heuristic that exploits some similarities to both knapsack and the loading problems. Material Handling Equipment Selection System (MATHES) was developed by Fisher et al. (1988). MATHES is a rule-based system for the selection from 24 different types of material-handling equipment.

Conclusion

The framework discussed in this section provides an alternative perspective from the material flow viewpoint. It integrates all of the important design factors associated with facilities planning. It incorporates most of the desired properties with respect to the overall plant operations. At the same time, the framework also summarizes the difficulties and complexities which confront the facility designer. The description of the framework is intended as a general exposition of the fundamental concepts and a direction for future research in facilities planning.

References

Armour, G.C. and Buffa, E.S., 1963. A heuristic algorithm and similation approach to relative location of facilities, *Manage. Sci*., 9(2), 294–300.

Askin, G. and Standridge, R. 1993. *Modeling and Analysis of Manufacturing Systems*, John Wiley, New York.

Buffa, E.S., Armour, G.C., and Vollman, T.E. 1964, Allocating facilities with CRAFT, *Harvard Business Rev.*, 42(2), 136–159.

Co, H., Wu, A., and Reisman, A., 1989. A throughput-maximizing facility planning and layout model, *Int. J. Prod. Res.*, 27(1), 1–12.

Egbelu, P. and Tanchoco, J.M.A. 1986. Potentials for bi-directional guide path for automated guided vehicle base systems, *Int. J. Prod. Res.*, 24(5), 1075–1097.

Faber, Z. and Carter, M.W. 1986. A new graph theory approach to forming machine cells in cellular production systems, *Flexible Manufacturing Systems: Methods and Studies*, North-Holland, Amsterdam, 301–318.

Fisher, E.L., Farber, J.B., and Kay, M.G. 1988. MATHES: an expert system for material handling equipment selection, *Eng. Costs Prod. Econ.* 14(4), 297–310.

Fortenberry, J.C. and Cox, J.F. 1985. Multiple criteria approach to the facilities layout problem, *Int. J. Prod. Res.*, 23(4)), 773–782.

Foulds, L.R. and Giffin, J.W. 1985. A graph-theoretic heuristic for minimzing total transport cost in facility layout, *Int. J. Prod. Res.*, 23(6), 1247–1257.

Golany, B. and Rosenblatt, M.J. 1989. A heuristic algorithm for the quadratic assignment formulation to the plant layout problem, *Int. J. Prod. Res.*, 27(2), 293–308.

Hassan, M.M.D. and Hogg, G.L. 1991. On constructing a block layout by graph theory, *Int. J. Prod. Res.*, 29(6). 1263–1278.

Hassan, M.M.D., Hogg, G., and Simth, D. 1985. Construction algorithm for the selection and assignment of materials handling equipment *Int. J. Prod. Res.*, (23(2), 381–392.

Hillier, F.S. 1963. Quantitative tools for plant layout analysis, *J. Ind. Eng.*, 14(1), 33–40.

Irani, S.A., Cavalier, T.M., and Cohen, P.H. 1993. Virtual manufacturing cells: exploring layout design and intercell flows for the machine sharing problem, *Int. J. Prod. Res.*, 31(4), 791–810.

King, J.R. 1980. Machine-component grouping in production flow analysis: an approach using rank order clustering algorithm, *Int. J. Prod. Res.*, 18(2), 213–232.

Kumar, K.R., Kisiak, A., and Vanelli, A. 1986. Grouping of parts and components in flexible manufacturing systems, *Eur. J. Operat. Res.*, 24, 387–397.

Kusiak, A., Vannelli, A., and Kummar, K.R. 1986. Clustering analysis: models and algorithms, *Control Cybernetics*, 15(2), 139–154.

Lee, R.C. and Moore, J.M. 1967. CORELAP: COmputerized RElationship LAyout Planning, *J. Ind. Eng.*, 18(1), 195–200.

Luxhoj, J.T. 1991. A methodology for the location of facility ingress/engress points, *Int. J. Oper. Prod. Manage.*, 11(5), 6–21.

McAuley, 1972. Machine grouping for efficient production, *The Prod. Eng.*, pp. 53–57.

McCormick, W.T., Schweitzer, P.J., and White, T.E. 1972. Problem decomposition and data reorganization by a clustering technique, *Operations Res.*, 20, 993–1009.

Mitrofanov, S.P. 1959. Nauchniye Osnovi Grouppovoi Technologi, Lenizdaz, Leningrad; translated into English, 1966, *Scientific Principle of Group Technology*, National Lending Library, England.

Rajagopalan R. and Batra, J.L. 1975. Design of cellular production system — a graph theoretic approach, *Int. J. Prod. Res.*, 13, 567–579.

Seehof, J.M. and Evans, W.O. 1967. ALDEP: Automated Layout DEsign Program, *J. Ind. Eng.*, 18(2), 690–695.

Sinriech, D. and Tanchoco, J.M.A. 1994. SFT — Segmented flow topology, *Material Flow Systems in Manufacturing*, J.M.A. Tanchoco, Ed., Chapman & Hall, London, 200–235.

Tompkins, J.A. 1993. *World Class Manufacturing*, IEEE, New York.

Warnecke, H.J., Dangelmier, W., and Kuhnle, H. 1985. Computer-aided layout planning, *Material Flow*, 1, 35–48.

Webster, D.B. and Reed, R. Jr. 1971. A material handling system selection model, *AIIE Trans.*, 3(1), 13–21.

Wemmerlov, U. and Hyer, N.L. 1989. Cellular manufacturing in the US industry: a survey of users, *Int. J. Prod. Res.*, 27(9), 1511–1530.

13.5 Rapid Prototyping

Takeo Nakagawa

Manufacturing Processes in Parts Production

The rapid progress of CAD technology for the design of machine parts has now made it easy to store three-dimensional shape data on computers. The application of this three-dimensional data has realized NC programming by CAD/CAM, resulting in the remarkable advance of automated production. The increase in highly functional machine parts and advanced designs has led to the design of more and more complicated surfaces using CAD. To reduce the lead time and costs for the development of new industrial products, "rapid prototyping" has been recognized as a unique, layered manufacturing technique for making prototypes.

With this rapid prototyping, shapes of machine parts are created by building up layers and layers of materials, unlike the material-removal technique which shapes by gradual machining using a cutting tool. In this sense, rapid prototyping resembles the joining technique of small particles or thin layers. Various different types of rapid prototyping methods have been born over the last couple of years. A common feature of these methods is that parts are directly shaped fully automatically according to CAD data. Specifically, in all of these methods, the three-dimensional CAD data is taken as composed of thin layers of two-dimensional data. The thin layers are formed using the two-dimensional data and built up to form an actual three-dimensional solid product. With the nature of the processing steps, this rapid prototyping technique is called *layered manufacturing.* As the method can also be used in other applications than prototyping, it is also referred to as *free-form fabrication.* Because three-dimensional objects can be made from three-dimensional CAD data, this new rapid-prototyping method is also called three-dimensional plotting. At the present time, laser stereolithography is the most widespread rapid-prototyping method.

Rapid Prototyping by Laser Stereolithography

Laser stereolithography involves the use of a liquid photocurable resin which cures instantaneously when scanned with a laser as a result of polymerization. As shown in Figure 13.5.1, the laser is scanned over this resin repeatedly to form thin layers of cured resin, which eventually build up to form a three-dimensional solid product. Specifically, the process involves first slicing a model based on three-dimensional CAD data stored on computer horizontally into equal thickness. Based on this slice data, the laser scans the thin liquid resin layer to form the first solid layer. Liquid resin is then poured over this cured layer of resin and again scanned by laser to form the next layer according to the next slice data. To ensure that the liquid resin on the cured resin is even, its surface is often swept by a blade. So, by repeating this process and forming layer over layer of cured resin, a solid object is formed.

Although the surface accuracy problems are resolved by making each layer extremely thin, a small amount of roughness of the surface eventually remains and subsequent polishing is often performed to achieve a smooth finish. For shapes which cannot be formed by building the layers upward, the resin is cured on a support. This support is also made of the same photocurable resin and is removed after the product has been formed. For some types of resins which do not cure completely with a laser alone, the whole product is cured by exposure to an ultraviolet lamp after it has been formed.

The photocurable resin is composed of photopolymerizing oligomer, reactive diluent, and photoinitiator. When a laser is irradiated onto this resin, the monomer undergoes a series of reactions to form a solid polymer which has a three-dimensional network structure. The resins used for laser stereolithography are the radical polymerization type, the cation polymerization type, or the hybrid type, which is a combination of the first two types. The curing properties and mechanical properties of the cured resin are important, because these affect the applications of the cured product formed by laser stereolithography. These resins are prepared minutely by adjusting the mixture rate of the resin components and the additives to suit laser stereolithography.

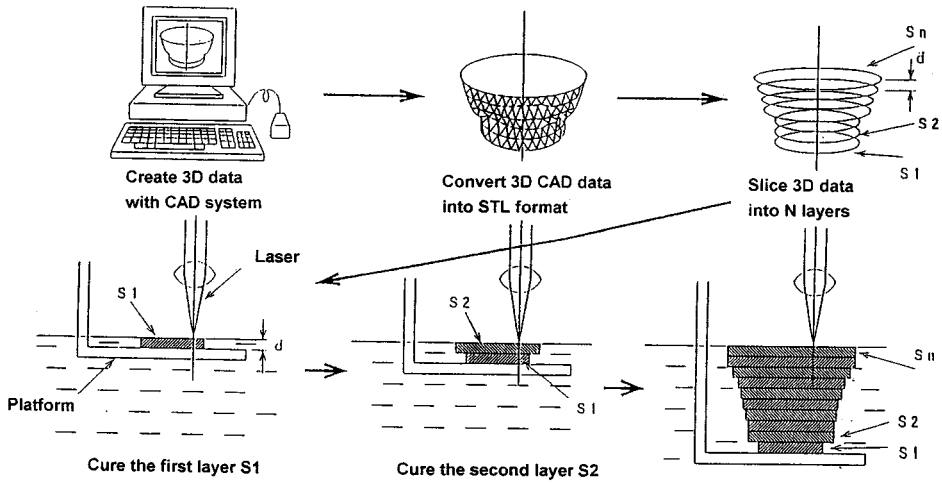


FIGURE 13.5.1 Principle of photocurable resin process.

In most cases, the He-Cd laser with 325-mm wavelength or Ar laser with 364-mm wavelength is used as the light source. Higher-power lasers like Ar perform higher-speed beam scanning, resulting in the increase of modeling speed. As shown in Figure 13.5.2, these lasers scan at very high speed in the same way as the laser printer by rotating galvanomirrors. In some special machines, the laser beam is exposed from the bottom as shown in Figure 13.5.3 by an *XY* plotter.

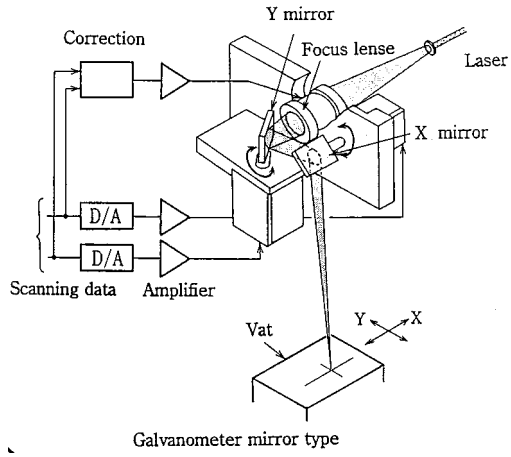


FIGURE 13.5.2 Scanning system of UV laser beam.

Laser stereolithography requires three-dimensional CAD data composed of surface or solid data in order to create solid models. Many types of CAD systems are now available on the market, and most of the CAD data they provide can be transmitted to laser stereolithography systems. At the same time, the laser stereolithography system is equipped with a scanning program, support and reinforcing rib design software, magnification/contraction functions, functions to determine the scanning and operating conditions, and simulation functions. Figure 13.5.4 shows a software flowchart in laser stereolithography.

The following shows the advantages of rapid-prototyping over material-removing processes such as machining.

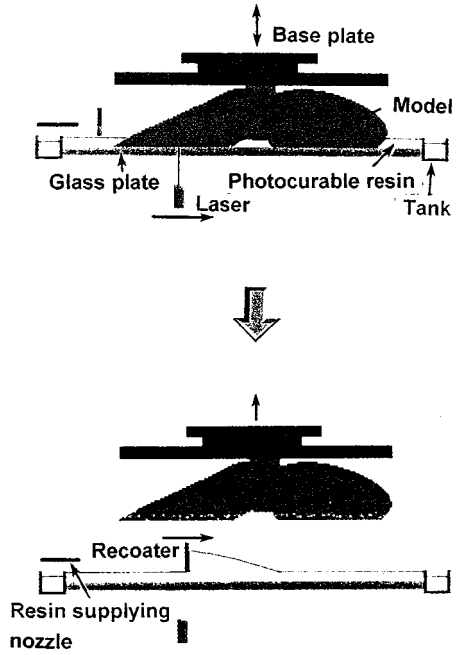


FIGURE 13.5.3 Laser beam radiation from below (Denken).

1. Deep holes and structures with complicated internal shapes which cannot be machined simply by cutting tools can be formed in a single process. Moreover, one rapid-prototyping machine is usually capable of fabricating any type of shape.
2. Rapid prototyping requires no complicated control programs such as tool path and repositioning of the workpiece. With the three-dimensional CAD data, there is no need for special knowledge of the cutting process, and operations from data input to actual fabrication are simple and short.
3. The rapid-prototyping systems produce no machining wastes. Because they do not vibrate and are silent, they can be used in offices like OA business machines. They can also be operated fully automatically even at night since there is no need for the management of tooling.

The major shortcomings of laser stereolithography are that only photocurable resins can be used and the material strength of these materials is slightly worse than the common polymer. In addition, metal products cannot be manufactured directly by laser stereolithography.

Other Rapid-Prototyping Methods

The laser stereolithography method was developed in an early stage and is currently applied extensively. Besides laser stereolithography, many different types of new rapid-prototyping methods have also emerged. As shown in Figure 13.5.5, rapid prototyping can broadly be classified into photopolymer, powder sintering, ink jetting, fused deposition, and sheet cutting. Figure 13.5.6 shows the history of these rapid prototyping systems. Most of the methods were developed in the U.S., but the photopolymer process and sheet lamination were first proposed in Japan.

Another photopolymer process is the mask pattern-curing method shown in Figure 13.5.7. Similar to the photocopying process, a master pattern based on slice data is created, this pattern on the glass sheet is placed over a photocurable resin layer, and this layer is exposed to ultraviolet light. Although the machine is large, the exposing speed is faster than the above-mentioned laser beam method, and the thickness of the product is very precise because each surface formed is cut by milling to obtain precise thin layers.

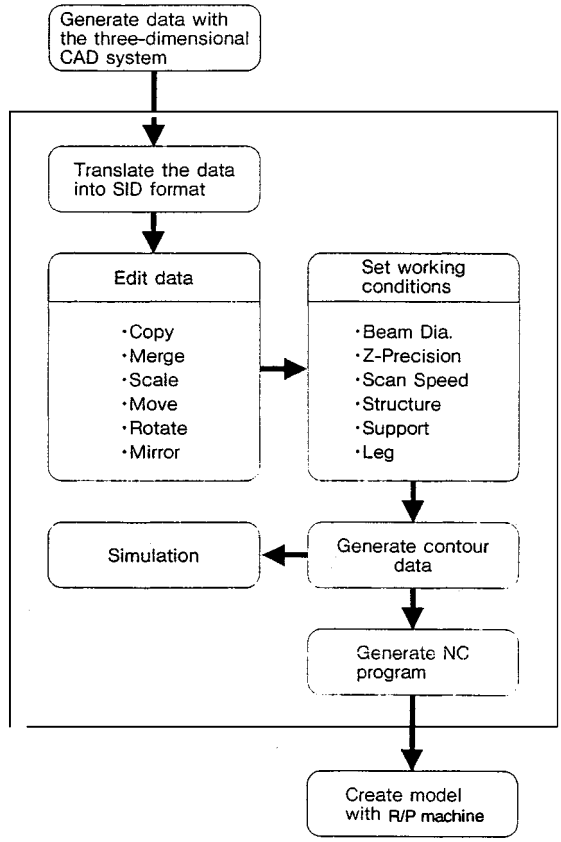


FIGURE 13.5.4 Software for R/P system (CMET).

Three-dimensional objects can also be formed by powder sintering. In the process shown in Figure 13.5.8, powder is used instead of liquid photocurable resin. The powder is evened out using a roller, a CO_2 laser is beamed, and the powder is bonded by heat fusion. In this case, powder is heated up beforehand to the temperature just below the melting point in the antioxidation environment using N_2 gas. It is possible to create high-density polymer solid models as well as porous models.

Porous polycarbonate models are quite suitable for the investment casting model. With this method, metal and ceramic powders can also be used. Metal and ceramic powders used are coated by resin and each metal or ceramic powder is bonded by the coated resin. Sintered porous ceramic molds can be used for casting molds. Powder binding can be performed by spraying binding material on the loose powder layer through the ink jet nozzle as shown in Figure 13.5.9. This is also used for making the sand mold for casting. When wax or resin is sprayed from the jet nozzle, wax or resin models can be fabricated as shown in Figure 13.5.10. In this case, the surface of the sprayed thin layer should be machined smoothly and flatly in order to obtain vertical accuracy.

Figure 13.5.11 shows the fused deposition method. In this method, a fine nozzle deposits a layer of resin or wax. Wax is normally used to form lost wax models. Fused deposition systems, in which material is supplied by the pellet or wire, enable materials to be formed very similarly to general injection mold materials like ABS and nylon. One of the two nozzles is used for making the support, where the support material is usually wax with lower melting points.

Figure 13.5.12 shows two methods which cut thin sheets according to slice data and laminate them to form three-dimentional objects. One method uses a laser to cut sheets applied with adhesive which are then laminated by hot toll pressing, while the other uses a knife to cut the sheets. In the latter case,

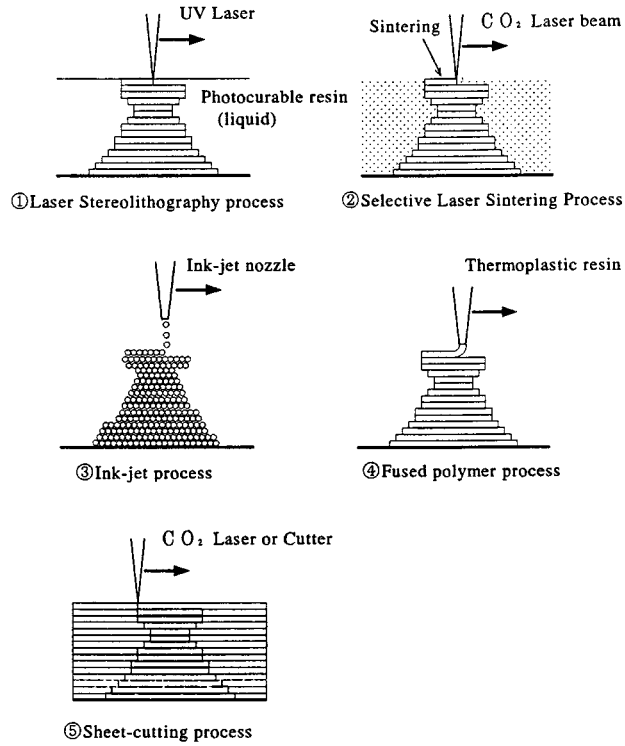


FIGURE 13.5.5 Schematic of various layer-additive fabrication processes.

adhesive is applied to sheets of paper according to the desired shape by spraying the toner using a dry Xerox-type copy machine. Due to the use of paper in these sheet lamination methods, the model formed should be immediately coated to prevent the absorption of moisture. Although there is a limit to the shapes that can be made, the method is nevertheless used for making wood models for casting, because it is inexpensive and enables large-sized models to be made and the model material is similar to wood.

The common feature of all of these methods is that slice data is obtained from three-dimentional CAD data and this slice data is used to laminate thin layers of material, which means that the same software can be used for all of these methods. Another feature is that all rapid-prototyping machines use the modified printing technology.

While many of these rapid-prototyping machines tend to be costly, inexpensive models are also now available. Machine cost reduction has been achieved by proper utilization of key parts which are used for the printer. New and improved methods should continue to be developed with the introduction of printing technology.

Application of Rapid Prototyping

Figure 13.5.13 shows the applications of three-dimentional models made by rapid prototyping. They are mainly intended for verifying CAD data, checking the designs, functional checks of prototypes, wax models for investment casting, master models for die and model making, mold making for prototype manufacturing, casting models, and medical use CT and MRI data. Although dimensional accuracy was given little importance in the verification stage of CAD data and design, high dimensional accuracy is now demanded of the functional check of prototypes. Because photocurable resins contract in the solidification process, slight distortions are generated in the fabricated product. Among the various rapid-prototyping machines, the photocurable resin process is most suitable for making very complicated shapes and obtaining the highest accuracy.

TOPOGRAPHY		PHOTOSCULPTURE	
Blanther patent filed	1890	1860	Willeme photosculpture
Perera patent filed	1937	1902	Baese patent filed
Zang patent filed	1962	1922	Monteah patent filed
Gaskin patent filed	1971	1933	Morioka patent filed
Matsubara patent filed	1972	1940	Moriola patent filed
DiMatteo patent filed	1974	1951	Munz patent filed
Nakagawa laminated fabrication of tools	1979		

1968	**Swainson patent filed**
1972	**Ciraud disclosure**
1979	**Housholder patent filed**
1981	**Kodama publication**
1982	**Herbert publication**
1984	**Marutani patent filed, Masters patent filed, Andre patent filed, Hull patent filed**
1985	**Helisys founded Denken venture started**
1986	**Pomerantz patent filed, Feygin patent filed, Deckard patent filed, 3D founded, Light sculpting started**
1987	**Fudim patent filed, Arcella patent filed, Cubital founded DTM founded, Dupont Somos venture started**
1988	**1st shipment by 3D, CMET founded, Stratasys founded**
1989	**Crump patent filed, Helinski patent filed, Marcus patent filed, Sachs patent filed, EOS founded, BPM founded**
1990	**Levent patent filed, Quadrax founded, DMEC founded**
1991	**Teijin Seiki venture started, Foeckele & Schwarze founded, Soligen founded Meiko founded, Mitsui venture started**
1992	**Penn patent filed, Quadrax acquired by 3D Kira venture started**
1994	**Sanders Prototype started**
1995	**Aaroflex venture started**

FIGURE 13.5.6 History of R/P (Joseph Beaman).

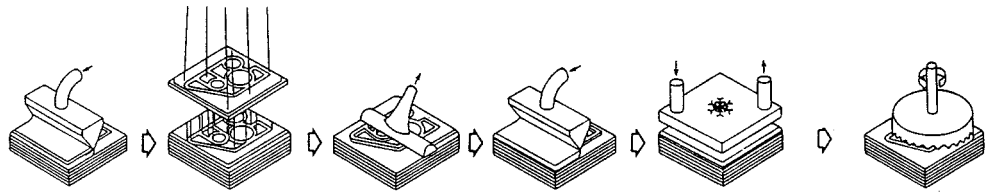


FIGURE 13.5.7 UV stereolithography (Cubital).

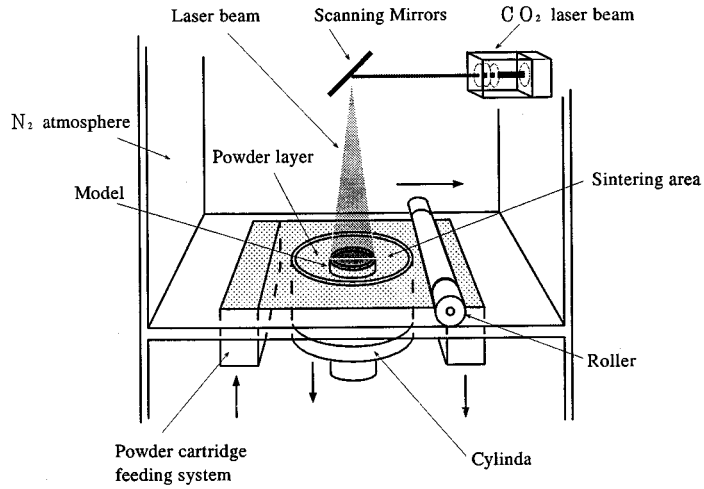


FIGURE 13.5.8 Selective laser sintering process (DTM, EOS).

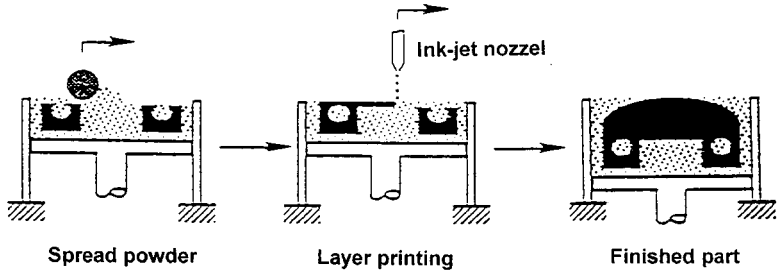


FIGURE 13.5.9 Ink-jet binding process (MIT, 3D printing).

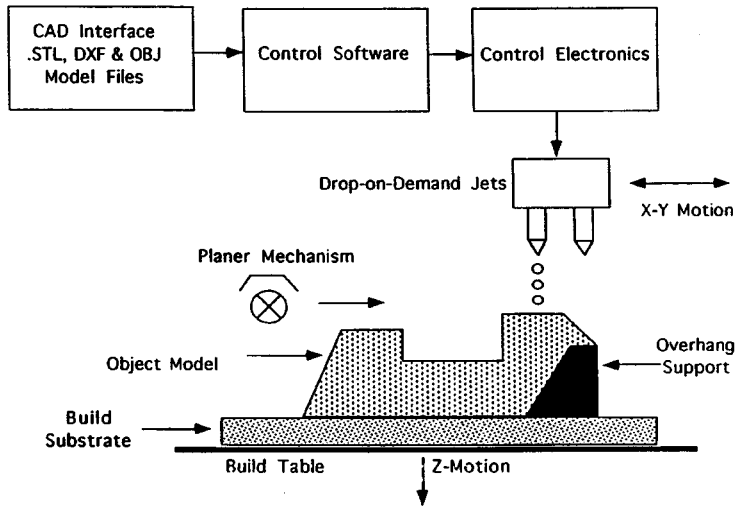


FIGURE 13.5.10 Ink-jet process (Sanders prototype).

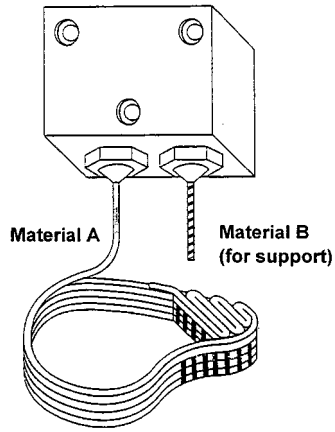


FIGURE 13.5.11 Fused deposition process (Stratasys).

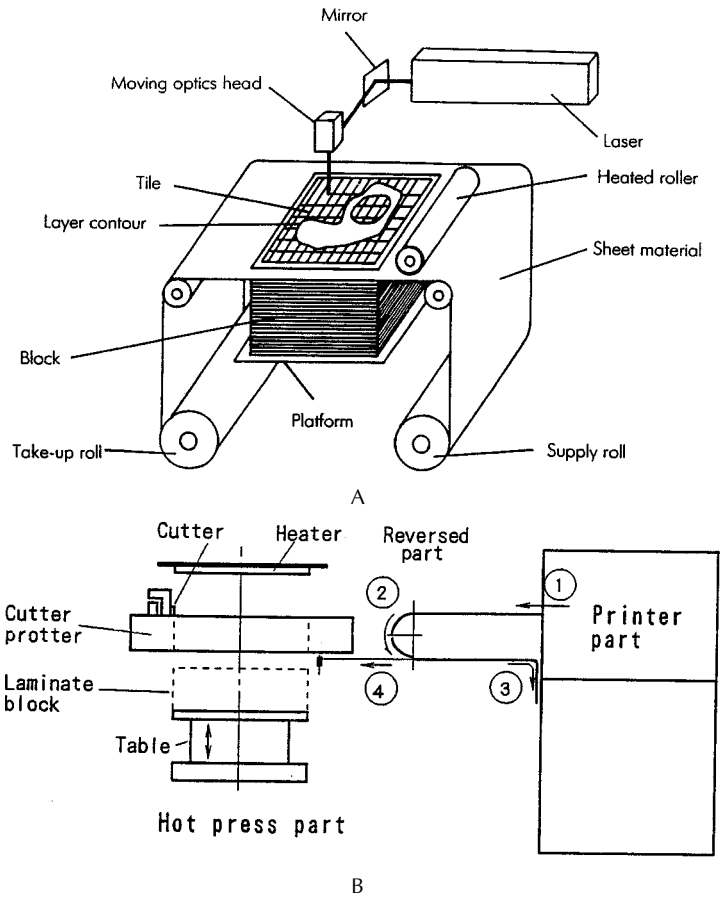


FIGURE 13.5.12 Sheet cutting process. (A) Laser (LOM). (B) Cutter (Kira).

Even in laser stereolithography, accuracy has improved to a considerable extent with the improvement of the resin and scanning method and the accumulation of know-how for positioning the reinforcing rib. It should also be possible to attain the same accuracy as injection molds by measuring formed products, correcting the data, or predicting errors.

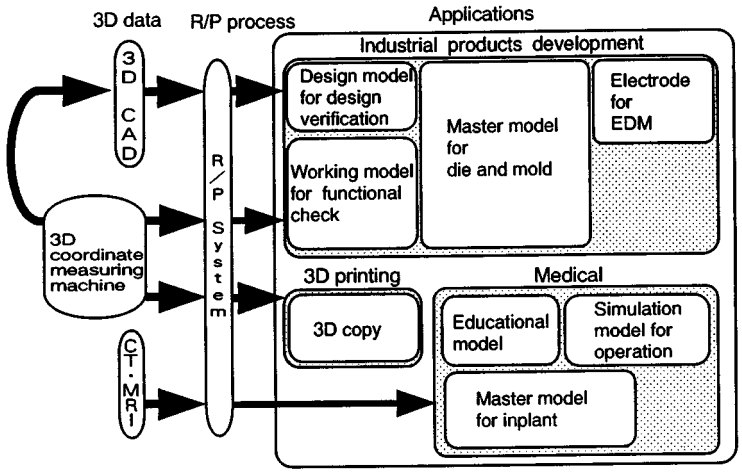


FIGURE 13.5.13 Application of rapid prototyping (source: CMET).

In general, photocurable resins are generally weak and brittle as compared with conventional polymer parts produced by injection molds. Urethane resin which is usually used for vacuum casting with a silicon rubber mold reversely copied from a rapid-prototyping model also lacks the required strength. In order to carry out the functional check of the prototypes created, other processes which can use normal thermoplastic should be used.

Figure 13.5.14 shows an intake manifold for car engines made by laser stereolithography. This serves as a test model for checking fluid performance of air. For such purposes, current photocurable resins available prove relatively satisfactory.

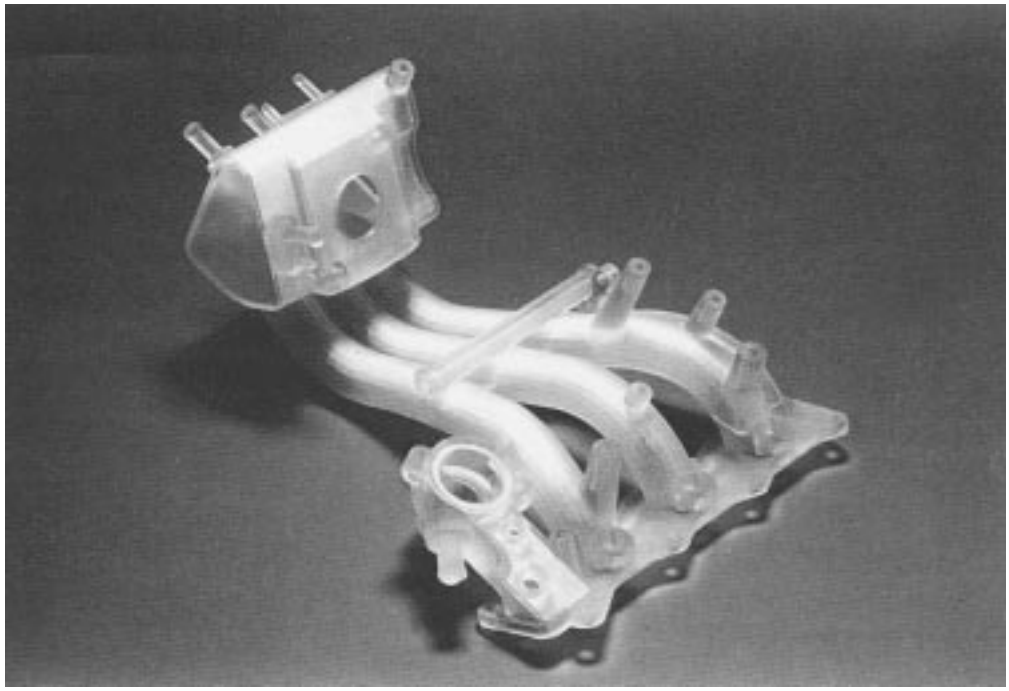


FIGURE 13.5.14 Sample for fluid dynamic analysis.

Although the powder sintering process can apply some metals, there is no suitable rapid-prototyping technique which can directly form products from metal materials at the moment. Research activities are underway to study the feasibility of producing molds from metal materials directly using a three-dimentional printing technique. In a general application, the lost wax models are first made by rapid prototyping and then used for creating metal prototypes by investment casting.

Rapid prototyping is still a relatively new technology, and therefore there are considerable opportunities for technical improvements.

General Rapid Prototyping in Production

For rapid prototyping to be carried out, three-dimentional CAD data must be available. Creating the CAD data takes far more time than creating the three-dimentional models based on the CAD data. By realizing efficient concurrent engineering, rapid prototyping will no doubt become a very important tool. In general, many other types of production systems can be included in the list of systems currently termed *general rapid prototyping* (casting and machining, etc.).

Use of Three-Dimentional CAD Data

Casting methods which are able to produce green sand molds satisfy the conditions of rapid prototyping. Expendable pattern casting is also suitable for rapid manufacturing. In this case, a three-dimentional polystyrene foam model is made by machining or binding. Most of the industrial products around us are produced with dies and molds. Because they are expensive to manufacture, dies and molds are unsuitable for making prototypes and for small-lot production. This may be a reason why rapid prototyping was developed; however, some prototype production methods do involve the use of dies and molds. Flexible prototype production has been carried out in sheet metal forming with the use of the turret punch press, laser beam cutting machine, and NC press brake. Producing dies and molds rapidly and manufacturing using such dies and molds also fit into the category of general rapid prototyping in the broad sense. Examples include what is known as the low-cost blanking dies using steel rule, deep drawing die made of zinc alloy and bismuth alloy.

Among the many general rapid-prototyping systems that exist, those newly developed rapid-prototyping methods discussed are gradually becoming methods for creating complicated products accurately with the use of three-dimentional CAD data. In terms of the total cost, applications of these new methods are still limited, but the spread of three-dimentional CAD data and technological progress of rapid prototyping should make them one of the common manufacturing techniques in the near future.

References

Ashley, S. 1992. Rapid prototyping systems, *Mech. Eng.*, April, 34–43.

Jacobs, P.F. 1992. *Rapid Prototyping and Manufacturing*, Society of Manufacturing Engineers, Dearborn, MI.

Rapid Prototyping in Europe and Japan, Japan and World Technology Evaluation Centers CJTEC/WTECS Report, September 1996, Loyola College, Baltimore, MD.

Sachs, E. et al. 1990. Three dimensional printing: rapid tooling and prototypes directly from a CAD model, *CIRP Ann.*, 39(1), 201–204.

Solid Freeform Fabrication Symposium, University of Texas, Austin, TX 1991–1993.

13.6 Underlying Paradigms in Manufacturing Systems and Enterprise Management for the 21st Century

Quality Systems

H. E. Cook

Introduction

Quality engineering has been described as the process of minimizing the sum of the total costs and the functional losses of manufactured products. Total costs include variable costs, investment, maintenance/repair costs, environmental losses, and costs of disposal or recycling. Functional losses arise from deviations from ideal performance. A subset of total quality management, quality engineering, focuses on parameter and tolerance design after the target specifications for the product have been developed as part of system design.

In contrast to quality engineering, total quality management embraces the entire product realization process. Its objective should be to maximize the net value of the product to society which includes buyer, seller, and the rest of society. Product value is determined solely by the customer and can be set equal to the maximum amount the customer would be willing to pay for the product. For a product to be purchased, its price must be less than its perceived value to the customer at the time of purchase. Consumer surplus is the difference between value and price.

The true value of a product is formed by the customer after assessing the product's performance over the complete time period that he or she used it. Functional quality loss is also known as the cost of inferior quality which is equal to the loss of value incurred by a product as a result of its attributes being off their ideal specification points (Figure 13.6.1). When manufacturing costs are added to value, the resulting sum (equal to total quality less environmental losses) is maximized when the attribute is off its ideal specification because of the impossibly high costs required to make a product perfect.

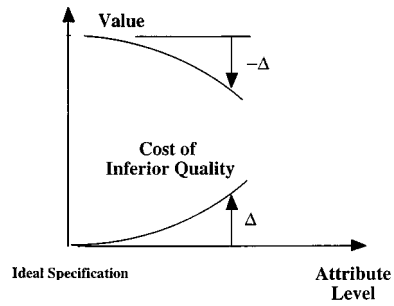


FIGURE 13.6.1 The relation of product value to the cost of inferior quality.

Requirements Flow

The systems viewpoint, as expressed by the flow of requirements shown in Figure 13.6.2, is helpful in considering the full ramifications of total quality management. Every system can be divided into subsystems and every system is but a subsystem of a larger system. Each task receives input requirements from its customer (either internal or external) and sends output requirements to its suppliers.

Task Objectives

A major objective of the system task is to assess customer needs, to translate those needs into a complete set of system-level specifications for the product, and to send a key (but partial) set of subsystem requirements to those responsible for the subsystem tasks. The system specifications and subsystem requirements developed by the system task should be such that (1) customers will want to purchase the product in a competitive marketplace and, with use, find that the product meets or exceeds their

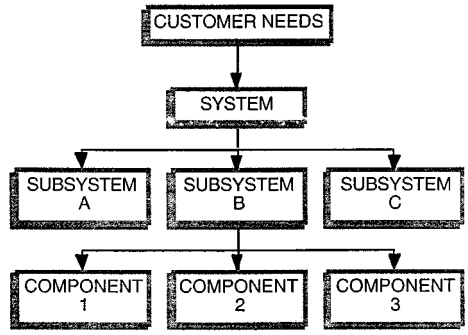


FIGURE 13.6.2 Flow of requirements from customer through enterprise.

expectations, (2) the product will meet the profitability objectives of the enterprise, and (3) all environmental rules and regulations are met. The system task also has the responsibility of resolving conflicts which arise between subsystem tasks.

The subsystem task receives the key requirements from its internal customer and translates these into a complete set of subsystem requirements and sends a key (but partial) set of component requirements to those responsible for the component tasks. In turn, those responsible for each component task translate the requirements received into a complete set of component requirements and a partial (but key) set of raw material requirements. Requirements set at each level include controls on variable costs, investment, performance, reliability, durability, service, disposal, environmental quality, package, assembly, and timing for both production and prototype parts. Synchronization is very important to total quality management as parts should be received exactly when needed with minimal inventory.

Parts Flow

The response to the requirements flow is a parts flow in the opposite direction that begins with the conversion of raw materials into components. This is followed by the assembly of components into subsystems which are shipped to the system task for final assembly. The process is completed by shipping the finished product to the customer. Thus, each task shown in Figure 13.6.1 has both a planning or design function as well as other functions including manufacturing, assembly, purchasing, marketing, service, accounting, and finance. The actions taken to meet customer needs should be traceable as the requirements flow through the enterprise. With the systems viewpoint, all parameters are measured or computed at the full system level including value, costs, and investment.

Task Management

Within each task shown in Figure 13.6.2 are subtasks. The combined flow of requirements and parts between several subtasks is shown in Figure 13.6.3 using a modified IDEF representation. Requirements are shown as controls which flow from left to right, and parts, in response, are shown as flowing from right to left.

Each task is accomplished by exercising its authority, responsibility, and capability. Authority to set requirements on parts should rest fully and undiluted with the task receiving the parts. The task which ships the parts should possess the full authority, responsibility, and capability to manufacture the parts for its customer. Before sourcing of parts, demonstration of capability by the manufacturer is a vital element of sound quality engineering. Capability is ultimately determined by the set of tools which the task has at its disposal and includes the skills and experience of the people as well as the hardware and software used by them.

Because a broad range of skills is needed, the required expertise is generated by forming a team to carry out the task. Quality tools used by the teams include structured methodologies such as Taguchi methods (design of experiments), quality function deployment, failure mode effects and criticality

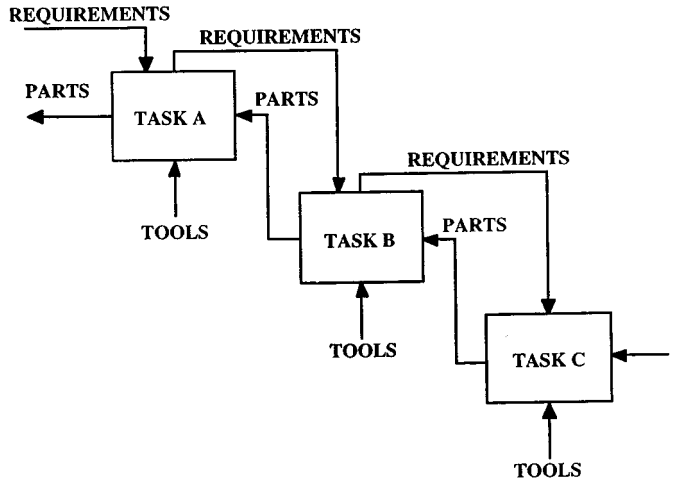


FIGURE 13.6.3 The flow of requirements and parts in a modified IDEF0 (function model) representation.

analysis, statistical process control, cost analysis, and value analysis. It is highly recommended that final authority and responsibility for each task rest with one person.

Fundamental and Bottom-Line Metrics

A variety of parameters can be used to measure the progress of quality improvements. These include things such as the degree of customer satisfaction expressed for the product, the frequency of repair, and the variance found in product dimensions and performance levels that directly impact value to the customer. Repair and operating costs borne by the customer subtract from value if they are greater than what was anticipated by the customer. Likewise, resale price subtracts form value if it is below what the customer expected. Performance degradation of the product over its lifetime of use subtracts from its value. Noise and atmospheric pollution caused by the manufacture and use of the product subtract from the net value of the product to society. The costs to manufacture and develop the product subtract from the net value received by the manufacturer. Moreover, products which are not improved in value and reduced in costs at the pace of competing products will likely be eliminated from the market in time.

These metrics can be grouped into one of three categories — value, cost, and the pace of innovation. They represent the fundamental metrics for the product because they determine what the bottom-line metrics of profitability and market share will be. Management of the fundamental metrics is the management of total quality and, likewise, the management of the total enterprise. The level of sustained profitability is the best measure of how well total quality is being managed in competitive markets.

Collaborative Manufacturing

James J. Solberg

Introduction

The world of manufacturing is undergoing rapid change in almost every aspect. It has become something of a cliché to speak of paradigm shifts, but there is no doubt that many assumptions, beliefs, and practices of the past are being seriously questioned. Meanwhile, the daily struggle of manufacturing enterprises to cope with the ordinary problems of producing products and satisfying customers goes on. In ecological terms, manufacturing companies are faced with a competitive struggle for survival (as always), augmented with the additional challenge of a changing climate. It is not surprising that many people are confused about what is happening and what to do about it.

What Is Collaborative Manufacturing (CM)?

CM is a very broad arena which incorporates many other topical themes of the day, including team design, computer-supported collaborative work, agile manufacturing, enterprise integration, virtual enterprises, high-performance distributed computing, concurrent engineering, computer-integrated manufacturing, virtual reality, global sourcing, and business process reengineering. In general, CM is defined by the following attributes:

- Integrated product and process development, including customers and suppliers;
- Flexible manufacturing distributed over networks of cooperating facilities;
- Teamwork among geographically and organizationally distributed units;
- High-technology support for the collaboration, including high-speed information networks and integration methodology;
- Multidisciplines and multiple objectives.

What is the payoff for CM? Companies that can engage effectively in CM will have the potential for

- Better market opportunities;
- A wider range of design and processing options over which to optimize;
- Fewer and looser constraints restricting their capabilities;
- Lower investment costs;
- Better utilization of resources;
- Faster response to changes.

Obviously, we are still a long way from being able to do what was described in our scenario. However, many of the pieces are already available, and many of the enabling technologies are rapidly coming into commercial use. The high-speed networks (well beyond current Internet speeds) are being developed and are certain to become both cost-effective and ubiquitous within a few years. The needed software developments, such as agent programming languages and interoperability standards, are progressing nicely. Nevertheless, an enormous agenda for needed research can be derived from unmet needs, particularly in areas that directly relate traditional manufacturing to information technology. Bridging these two research communities will not be easy, but the effort will offer great rewards to those who succeed.

Who Is Doing CM?

A great deal of the current work in CM and related themes is best accessed through the Internet. Understandably, the most active researchers in these fields are “Internet aware” and use it for both gathering and distributing their knowledge. Consequently, the best way to survey recent work is to browse the net, following links to associated sites. Unfortunately, the medium is so dynamic that material can appear or disappear at any time. A few of the better home pages are given here as starting points. Search engines can pick up more current connections.

- ACORN — A project involving Carnegie Mellon University, MIT, the University of Michigan and Enterprise Integration Technologies, Inc. (EIT):
  (http://www.edrc.cmu.edu:8888/acorn/acorn_front.html).
- Agile Manufacturing projects and organizations — A wide range of research an development projects funded through DARPA and NSF:
  (http://absu.amef.lehigh.edu/NIST-COPIES/pimain.html).
- SHARE — A DARPA-sponsored project to create a methodology and environment for collaborative product development, conducted by EIT and Stanford:
  (http://gummo. stanford.edu/html/SHARE/share.html).

- SHADE — Another DARPA-sponsored project, SHAred Dependency Engineering, information sharing aspect of concurrent engineering. It is led by the Lockheed AI Center, with help from Stanford and EIT: (http://hitchhiker.space.lockheed.com/aic/shade/papers/shadeoverview.html)
- Succeed — An NSF education coalition, which is investigating collaboration technologies to support research and education: (http://fiddle.ee.vt.edu/succeed/collaboration.html).
- Purdue Center for Collaborative Manufacturing — An NSF-sponsored engineering research center focused on the entire range of CM issues: (http://erc.www.ecn.purdue. edu/erc/).
- Groupware yellow pages: (http://www.consensus.com:8300/GWYP_TOC.html).
- Human computer interaction: (ftp://cheops.cis.ohio-state.edu/pub/hcibib/README.html).
- Computer Supported Collaborative Work yellow pages: (http://www.tft.tele.no/cscw/).
- Cross platform page: (http://www.mps.org/~ebennett/).

References

National Research Council. 1994. *Realizing the Information Future: The Internet and Beyond*, National Academy Press, Washington, D.C.

National Research Council. 1994. *Research Recommendations to Facilitate Distributed Work*, National Academy Press, Washington, D.C.

National Research Council. 1995. *Unit Manufacturing Processes: Issues and Opportunities in Research*, National Academy Press, Washington, D.C.

National Research Council. 1995. *Information Technology for Manufacturing: A Research Agenda*, National Academy Press, Washington, D.C.

Electronic Data Interchange

Chris Wang

Introduction

Electronic data interchange (EDI) is a method to exchange business information between computer systems. In a traditional purchasing environment, buyers, when placing computer-generated orders, will mail them to suppliers, and it could take days before the suppliers receive them and then rekey the orders into their computer system. Using EDI, the buyer's computer system can generate an EDI standard order transaction and transmit it directly to the supplier's inventory system for material pickup. It happens instantly. The benefit of EDI is quite obvious in this case as it reduces material lead time dramatically. Consequently, the objectives of EDI implementation should not be limited to just reducing paperwork and clerical work; instead, it should be used as a methodology to streamline company processes and become competitive in the marketplace.

EDI Elements

EDI consists of the following elements:

- Trading partners — The parties, such as a manufacturer and a supplier, who agree to exchange information.
- Standards — The industry-supplied national, or international formats to which information is converted, allowing disparate computer systems and applications to interchange it. This will be discussed in more detail later.
- Applications — The programs that process business information. For example, an orders application can communicate with an orders entry application of the trading partner.
- Translation — The process of converting business information, usually from a format used by an application, to a standard format, and vice versa.

- Electronic transmission — The means by which the information is delivered, such as a public network. Some companies may choose to build their own transmission facilities. For others, VAN (value-added network) seems to be a good choice as companies do not have to invest heavily in communication equipment and personnel to support it. The VAN provider can handle disparate communication hardware and software and provide wide-area network access at a reasonable cost.

EDI in Manufacturing

In the present-day business environment, many companies are turning to just-in-time (JIT) and other techniques to compete as effectively as possible. EDI can make an important contribution to the success of JIT by ensuring that information exchanged between business partners is also just in time.

In traditional manufacturing, material is stored in quantities much larger than required because of faulty components and possible waste in the production process. To solve problems of carrying safety stock and still producing high-quality product, JIT seems to be an effective technology.

JIT systems are designed to pull raw materials and subassemblies through the manufacturing process only when they are needed and exactly when they are needed. Also, with rapidly changing production needs, orders are getting smaller and are issued more frequently. The traditional paperwork environment simply cannot effectively cope with this change. This is why EDI comes in to play a key role to provide fast, accurate information to achieve these JIT goals. In other words, EDI can provide JIT information in manufacturing processes.

- EDI cuts order delivery and lead time — The more control points you have in a process, the greater the number of potential problems. EDI eliminates "control points" for the order process. It eliminates the need to mail orders and rekey order information at the receiving end. It reduces the material lead time for production use.
- Connect applications and processes — With EDI capability, information, such as scheduling, orders, advance delivery notice, statistical process control data, and material safety data sheets can pass quickly and accurately from the supplier's computer application to the customer's computer application, so that arriving material can be put to production use with confidence. This meets one of the important goals of JIT, that is, to turn the supplier's entire production line into a vast stockroom so a company does not have to maintain a huge warehouse and the working capital tied to excess inventory.
- Improve relationship with customers and suppliers — In the supply chain environment, the quicker the chain moves, the better the customers' needs can be met. With quicker orders, acknowledgments, order changes, and invoices, EDI can satisfy customers' needs more quickly. Also, the time spent on order tracking and error recovery can now be used in a more productive way and can improve customer/supplier relationships. Companies deeply involved in EDI may see the number of suppliers reduced. This is because through the EDI process, a company can weed out many suppliers who are not efficient and reliable.

EDI Standards

When two organizations exchange business forms electronically, information is encoded and decoded by the computer software of both parties. Therefore, the information must be unambiguous, in order to avoid different interpretations. This relates to the meaning of the terms used, the representation of data used, the codes to be used for data, and the sequence in which data are to be transmitted. All these parameters must be arranged between the two parties on a detailed level.

There are many standard types — it can be based on bilateral agreement, imposed by a dominating party in a certain marketplace, or jointly developed by an industrial group. Some standards have been ratified by international organizations.

The pioneer of EDI standard development was the transportation industry. TDCC (Transportation Data Coordinating Committee) developed sets of standards for transportation mode — air, ocean, motor,

and rail. Later, the U.S. grocery industry developed a set of standards, UCS (Uniform Communication Standard), based on TDCC structure.

The TDCC and UCS are more geared toward the business forms exchanged by shipper/carrier, for example, bill of lading. Not until the American National Standards Institute got involved did a general use standard for all industries start to develop, leading to the birth of ANSI X12 standards.

The ANSI X12 is popular in the U.S. Although ANSI X12 is intended for all industries, different user groups still come up with their own conventions to address their specific needs but remain under the X12 umbrella. To name a few, there are AIAG (Automotive Industry Action Group) for the auto industry, CIDX for the chemical industry, and EIDC for the electronic industry.

In Europe, at approximately the same time period, under the leadership of the United Kingdom, the TDI (Trade Data Interchange) was developed. The TDI syntax and structure are quite different from the ANSI X12. To resolve the incompatibility, the U.N. organization UNJEDI was formed to develop an EDI international standard containing features from both TDI and ANSI X12. The result was EDIFACT (EDI for Administrative, Commerce, and Transport). This is the standard to which the world is trying to convert.

EDI Implementation

Before implementation of EDI, planning is critical to success. First, get all the right people involved in planning and implementation of EDI. Ensure that every employee gets EDI education on how to use EDI as a business tool to manage his or her job. Prepare a strategic plan to get approval and support from top management. Top management should be aware of the significant benefits of EDI and has to appreciate the potential of EDI as a business methodology to improve the bottom line.

As part of a strategic plan, it is crucial to perform an operational evaluation. This evaluation details how the internal departments of the company function. For each paper document under evaluation, information flow is tracked, processing procedures are scrutinized, time is measured, and costs are calculated. This will provide top management with valuable information as important as industry trends and competition information. The operational evaluation provides the company with detailed documentation about how it does business in a paper-based environment. This information then serves as a benchmark against which to measure projected costs and benefits of the EDI model.

Once the strategic plan is in place, available resources must be allocated to the departments that will generate the most benefits for the company. Once EDI is implemented in the company, the next step is to sell it to trading partners to maximize the EDI investment.

Summary

EDI is on a fast growing path. EDI software and communication services are available and not expensive. It will not be too long before EDI becomes mandatory as a business practice. If a company is determined to implement EDI, it should look beyond just connecting two computer systems. To achieve the best return on the EDI investment, one should try to use EDI to improve the existing processes within the organization and the relationship with customers and suppliers.

References

Gerf, V.G. 1991. Prospects for electronic data interchange, *Telecommunications*, Jan., 57–60.
Mandell, M. 1991. EDI speeds Caterpillar's global march, *Computerworld*, 25(32), 58.

Lewis, F.L.; et. al. “Robotics”
Mechanical Engineering Handbook
Ed. Frank Kreith
Boca Raton: CRC Press LLC, 1999

Robotics

Frank L. Lewis
University of Texas at Arlington

John M. Fitzgerald
University of Texas at Arlington

Ian D. Walker
Rice University

Mark R. Cutkosky
Stanford University

Kok-Meng Lee
Georgia Tech

Ron Bailey
University of Texas at Arlington

Frank L. Lewis
University of Texas at Arlington

Chen Zhou
Georgia Tech

John W. Priest
University of Texas at Arlington

G. T. Stevens, Jr.
University of Texas at Arlington

John M. Fitzgerald
University of Texas at Arlington

Kai Liu
University of Texas at Arlington

14.1 Introduction

The word "robot" was introduced by the Czech playright Karel Čapek in his 1920 play *Rossum's Universal Robots.* The word "robota" in Czech means simply "work." In spite of such practical beginnings, science fiction writers and early Hollywood movies have given us a romantic notion of robots. Thus, in the 1960s robots held out great promises for miraculously revolutionizing industry overnight. In fact, many of the more far-fetched expectations from robots have failed to materialize. For instance, in underwater assembly and oil mining, teleoperated robots are very difficult to manipulate and have largely been replaced or augmented by "smart" quick-fit couplings that simplify the assembly task. However, through good design practices and painstaking attention to detail, engineers have succeeded in applying robotic systems to a wide variety of industrial and manufacturing situations where the environment is structured or predictable. Today, through developments in computers and artificial intelligence techniques and often motivated by the space program, we are on the verge of another breakthrough in robotics that will afford some levels of autonomy in unstructured environments.

On a practical level, robots are distinguished from other electromechanical motion equipment by their dexterous manipulation capability in that robots can work, position, and move tools and other objects with far greater dexterity than other machines found in the factory. Process robot systems are functional components with grippers, end effectors, sensors, and process equipment organized to perform a controlled sequence of tasks to execute a process — they require sophisticated control systems.

The first successful commercial implementation of process robotics was in the U.S. automobile industry. The word "automation" was coined in the 1940s at Ford Motor Company, as a contraction of "automatic motivation." By 1985 thousands of spot welding, machine loading, and material handling applications were working reliably. It is no longer possible to mass produce automobiles while meeting currently accepted quality and cost levels without using robots. By the beginning of 1995 there were over 25,000 robots in use in the U.S. automobile industry. More are applied to spot welding than any other process. For all applications and industries, the world's stock of robots is expected to exceed 1,000,000 units by 1999.

The single most important factor in robot technology development to date has been the use of microprocessor-based control. By 1975 microprocessor controllers for robots made programming and executing coordinated motion of complex multiple degrees-of-freedom (DOF) robots practical and reliable. The robot industry experienced rapid growth and humans were replaced in several manufacturing processes requiring tool and/or workpiece manipulation. As a result the immediate and cumulative dangers of exposure of workers to manipulation-related hazards once accepted as necessary costs have been removed.

A distinguishing feature of robotics is its multidisciplinary nature — to successfully design robotic systems one must have a grasp of electrical, mechanical, industrial, and computer engineering, as well as economics and business practices. The purpose of this chapter is to provide a background in all these areas so that design for robotic applications may be confronted from a position of insight and confidence. The material covered here falls into two broad areas: function and analysis of the single robot, and design and analysis of robot-based systems and workcells.

Section 14.2 presents the available configurations of commercial robot manipulators, with Section 14.3 providing a follow-on in mathematical terms of basic robot geometric issues. The next four sections provide particulars in end-effectors and tooling, sensors and actuators, robot programming languages, and dynamics and real-time control. Section 14.8 deals with planning and intelligent control. The next three sections cover the design of robotic systems for manufacturing and material handling. Specifically, Section 14.9 covers workcell layout and part feeding, Section 14.10 covers product design and economic analysis, and Section 14.11 deals with manufacturing and industrial processes. The final section deals with some special classes of robots including mobile robots, lightweight flexible arms, and the versatile parallel-link arms including the Stewart platform.

14.2 Commercial Robot Manipulators

John M. Fitzgerald

In the most active segments of the robot market, some end-users now buy robots in such large quantities (occasionally a single customer will order hundreds of robots at a time) that market prices are determined primarily by configuration and size category, not by brand. The robot has in this way become like an economic commodity. In just 30 years, the core industrial robotics industry has reached an important level of maturity, which is evidenced by consolidation and recent growth of robot companies. Robots are highly reliable, dependable, and technologically advanced factory equipment. There is a sound body of practical knowledge derived from a large and successful installed base. A strong foundation of theoretical robotics engineering knowledge promises to support continued technical growth.

The majority of the world's robots are supplied by established stable companies using well-established off-the-shelf component technologies. All commercial industrial robots have two physically separate basic elements: the manipulator arm and the controller. The basic architecture of all commercial robots is fundamentally the same. Among the major suppliers the vast majority of industrial robots uses digital servo-controlled electrical motor drives. All are serial link kinematic machines with no more than six axes (degrees of freedom). All are supplied with a proprietary controller. Virtually all robot applications require significant effort of trained skilled engineers and technicians to design and implement them. What makes each robot unique is how the components are put together to achieve performance that yields a competitive product. Clever design refinements compete for applications by pushing existing performance envelopes, or sometimes creating new ones. The most important considerations in the application of an industrial robot center on two issues: Manipulation and Integration.

Commercial Robot Manipulators

Manipulator Performance Characteristics

The combined effects of kinematic structure, axis drive mechanism design, and real-time motion control determine the major manipulation performance characteristics: reach and dexterity, payload, quickness, and precision. Caution must be used when making decisions and comparisons based on manufacturers' published performance specifications because the methods for measuring and reporting them are not standardized across the industry. Published performance specifications provide a reasonable comparison of robots of similar kinematic configuration and size, but more detailed analysis and testing will insure that a particular robot model can reach all of the poses and make all of the moves with the required payload and precision for a specific application.

Reach is characterized by measuring the extents of the space described by the robot motion and *dexterity* by the angular displacement of the individual joints. Horizontal reach, measured radially out from the center of rotation of the base axis to the furthest point of reach in the horizontal plane, is usually specified in robot technical descriptions. For Cartesian robots the range of motion of the first three axes describes the reachable workspace. Some robots will have unusable spaces such as dead zones, singular poses, and wrist-wrap poses inside of the boundaries of their reach. Usually motion test, simulations, or other analysis are used to verify reach and dexterity for each application.

Payload weight is specified by the manufacturer for all industrial robots. Some manufacturers also specify inertial loading for rotational wrist axes. It is common for the payload to be given for extreme velocity and reach conditions. Load limits should be verified for each application, since many robots can lift and move larger-than-specified loads if reach and speed are reduced. Weight and inertia of all tooling, workpieces, cables, and hoses must be included as part of the payload.

Quickness is critical in determining throughput but difficult to determine from published robot specifications. Most manufacturers will specify a maximum speed of either individual joints or for a specific kinematic tool point. Maximum speed ratings can give some indication of the robot's quickness but may be more confusing and misleading than useful. Average speed in a working cycle is the quickness

characteristic of interest. Some manufacturers give cycle times for well-described motion cycles. These motion profiles give a much better representation of quickness. Most robot manufacturers address the issue by conducting application-specific feasibility tests for customer applications.

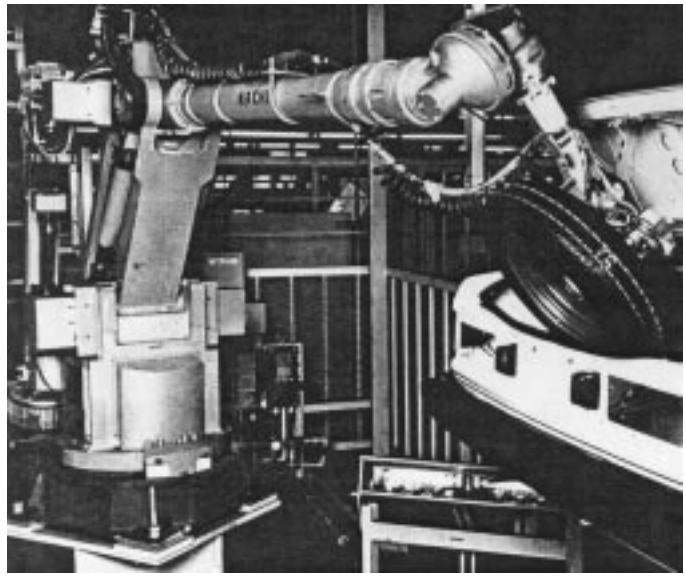
Precision is usually characterized by measuring repeatability. Virtually all robot manufacturers specify static position repeatability. Usually, tool point repeatability is given, but occasionally repeatability will be quoted for each individual axis. *Accuracy* is rarely specified, but it is likely to be at least four times larger than repeatability. Dynamic precision, or the repeatability and accuracy in tracking position, velocity, and acceleration on a continuous path, is not usually specified.

Common Kinematic Configurations

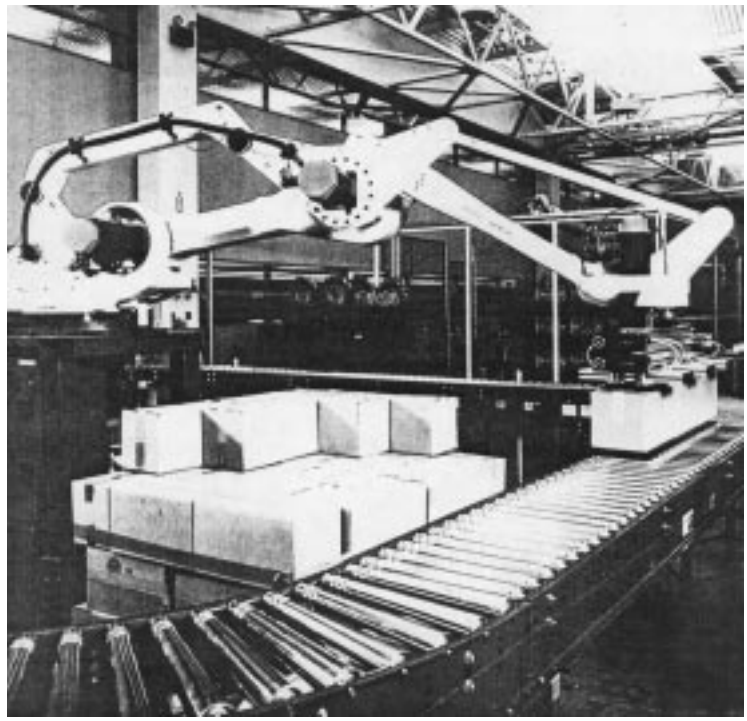
All common commercial industrial robots are serial link manipulators with no more than six kinematically coupled axes of motion. By convention, the axes of motion are numbered in sequence as they are encountered from the base on out to the wrist. The first three axes account for the spatial positioning motion of the robot; their configuration determines the shape of the space through which the robot can be positioned. Any subsequent axes in the kinematic chain provide rotational motions to orient the end of the robot arm and are referred to as wrist axes. There are, in principle, two primary types of motion that a *robot axis* can produce in its driven link: either *revolute* or *prismatic*. It is often useful to classify robots according to the orientation and type of their first three axes. There are four very common commercial robot configurations: Articulated, Type 1 *SCARA,* Type 2 SCARA, and Cartesian. Two other configurations, Cylindrical and Spherical, are now much less common.

Articulated Arms. The variety of commercial articulated arms, most of which have six axes, is very large. All of these robots' axes are revolute. The second and third axes are parallel and work together to produce motion in a vertical plane. The first axis in the base is vertical and revolves the arm sweeping out a large work volume. The need for improved reach, quickness, and payload have continually motivated refinements and improvements of articulated arm designs for decades. Many different types of drive mechanisms have been devised to allow wrist and forearm drive motors and gearboxes to be mounted close in to the first and second axis rotation to minimize the extended mass of the arm. Arm structural designs have been refined to maximize stiffness and strength while reducing weight and inertia. Special designs have been developed to match the performance requirements of nearly all industrial applications and processes. The workspace efficiency of well-designed articulated arms, which is the degree of quick dexterous reach with respect to arm size, is unsurpassed by other arm configurations when five or more degrees of freedom are needed. Some have wide ranges of angular displacement for both the second and third axis, expanding the amount of overhead workspace and allowing the arm to reach behind itself without making a 180° base rotation. Some can be inverted and mounted overhead on moving gantries for transportation over large work areas. A major limiting factor in articulated arm performance is that the second axis has to work to lift both the subsequent arm structure and payload. Springs, pneumatic struts, and counterweights are often used to extend useful reach. Historically, articulated arms have not been capable of achieving accuracy as well as other arm configurations. All axes have joint angle position errors which are multiplied by link radius and accumulated for the entire arm. However, new articulated arm designs continue to demonstrate improved repeatability, and with practical calibration methods they can yield accuracy within two to three times the repeatability. An example of extreme precision in articulated arms is the Staubli Unimation RX arm (see Figure 14.2.1).

Type I SCARA. The Type I SCARA (selectively compliant assembly robot arm) arm uses two parallel revolute joints to produce motion in the horizontal plane. The arm structure is weight-bearing but the first and second axes do no lifting. The third axis of the Type 1 SCARA provides work volume by adding a vertical or Z axis. A fourth revolute axis will add rotation about the Z axis to control orientation in the horizontal plane. This type of robot is rarely found with more than four axes. The Type 1 SCARA is used extensively in the assembly of electronic components and devices, and it is used broadly for the assembly of small- to medium-sized mechanical assemblies. Competition for robot sales in high speed electronics assembly has driven designers to optimize for quickness and precision of motion. A

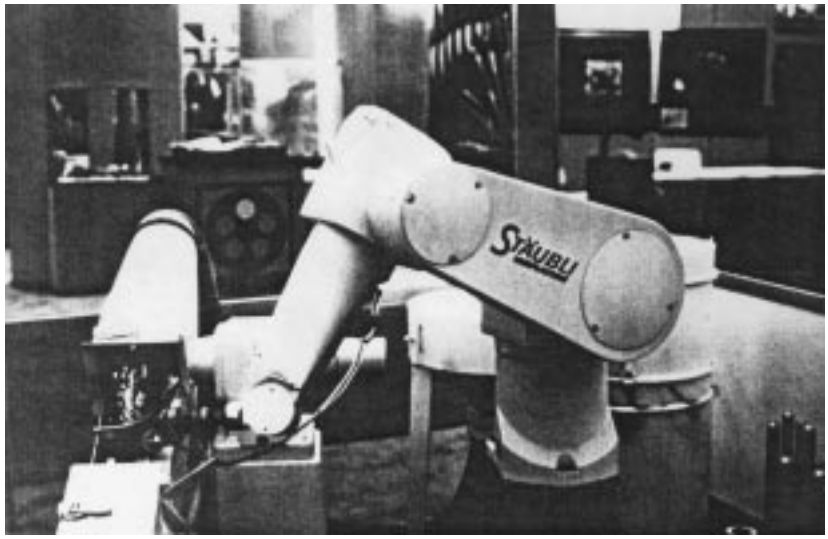


(a)

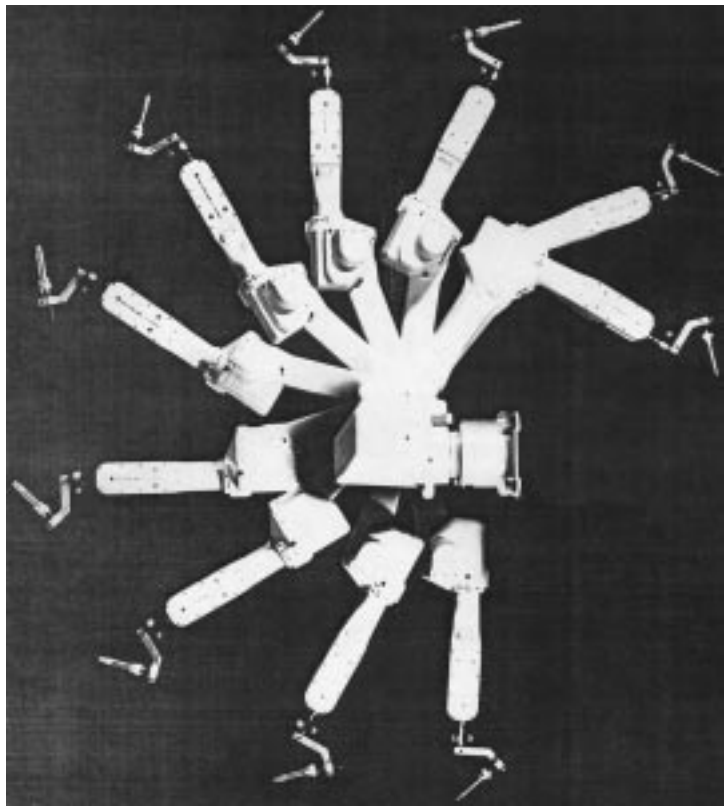


(b)

FIGURE 14.2.1 Articulated arms. (a) Six axes are required to manipulate spare wheel into place (courtesy Nachi, Ltd.); (b) four-axis robot unloading a shipping pallet (courtesy Fanuc Robotics, N.A.); (c) six-axis arm grinding from a casting (courtesy of Staubli Unimation, Inc.); (d) multiple exposure sideview of five-axis arc welding robot (courtesy of Fanuc Robotics, N.A.).



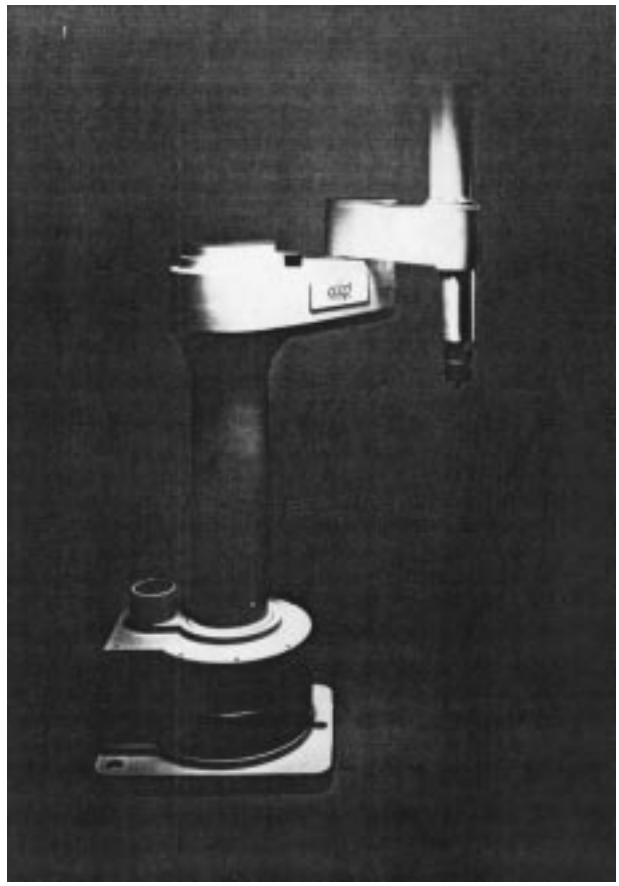
(c)



(d)

FIGURE 14.2.1 continued

well-known optimal SCARA design is the AdeptOne robot shown in Figure 14.2.2a. It can move a 20-lb payload from point "A" up 1 in. over 12 in. and down 1 in. to point "B" and return through the same path back to point "A" in less than 0.8 sec (see Figure 14.2.2).

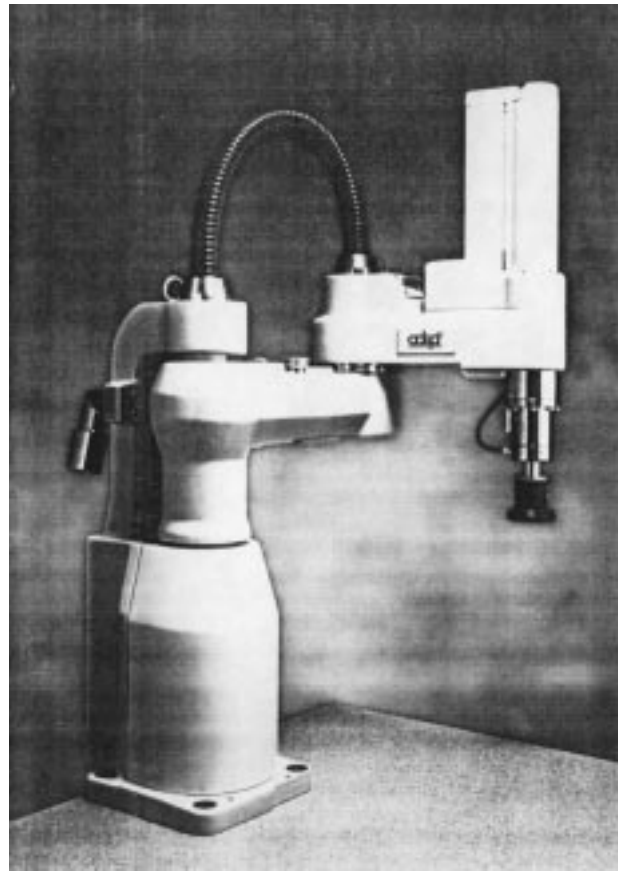


(a)

FIGURE 14.2.2 Type 1 SCARA arms (courtesy of Adept Technologies, Inc.). (a) High precision, high speed midsized SCARA; (b) table top SCARA used for small assemblies.

Type II SCARA. The Type 2 SCARA, also a four-axis configuration, differs from Type 1 in that the first axis is a long, vertical, prismatic Z stroke which lifts the two parallel revolute axes and their links. For quickly moving heavier loads (over approximately 75 lb) over longer distances (over about 3 ft), the Type 2 SCARA configuration is more efficient than the Type 1. The trade-off of weight vs. inertia vs. quickness favors placement of the massive vertical lift mechanism at the base. This configuration is well suited to large mechanical assembly and is most frequently applied to palletizing, packaging, and other heavy material handling applications (see Figure 14.2.3).

Cartesian Coordinate Robots. Cartesian coordinate robots use orthogonal prismatic axes, usually referred to as X, Y, and Z, to translate their end-effector or payload through their rectangular workspace. One, two, or three revolute wrist axes may be added for orientation. Commercial robot companies supply several types of Cartesian coordinate robots with workspace sizes ranging from a few cubic inches to tens of thousands of cubic feet, and payloads ranging to several hundred pounds. Gantry robots are the most common Cartesian style. They have an elevated bridge structure which translates in one horizontal direction on a pair of runway bearings (usually referred to as the X direction), and a carriage which



(b)

FIGURE 14.2.2 continued

moves along the bridge in the horizontal "Y" direction also usually on linear bearings. The third orthogonal axis, which moves in the Z direction, is suspended from the carriage. More than one robot can be operated on a gantry structure by using multiple bridges and carriages. Gantry robots are usually supplied as semicustom designs in size ranges rather than set sizes. Gantry robots have the unique capacity for huge accurate work spaces through the use of rigid structures, precision drives, and work-space calibration. They are well suited to material handling applications where large areas and/or large loads must be serviced. As process robots they are particularly useful in applications such as arc welding, waterjet cutting, and inspection of large, complex, precision parts.

Modular Cartesian robots are also commonly available from several commercial sources. Each module is a self-contained completely functional single axis actuator. Standard liner axis modules which contain all the drive and feedback mechanisms in one complete structural/functional element are coupled to perform coordinated three-axis motion. These modular Cartesian robots have work volumes usually on the order of 10 to 30 in. in X and Y with shorter Z strokes, and payloads under 40 lb. They are typically used in many electronic and small mechanical assembly applications where lower performance than Type 1 SCARA robots is suitable (see Figure 14.2.4).

Spherical and Cylindrical Coordinate Robots. The first two axes of the spherical coordinate robot are revolute and orthogonal to one another, and the third axis provides prismatic radial extension. The result is a natural spherical coordinate system and a work volume that is spherical. The first axis of cylindrical coordinate robots is a revolute base rotation. The second and third are prismatic, resulting in a natural cylindrical motion.

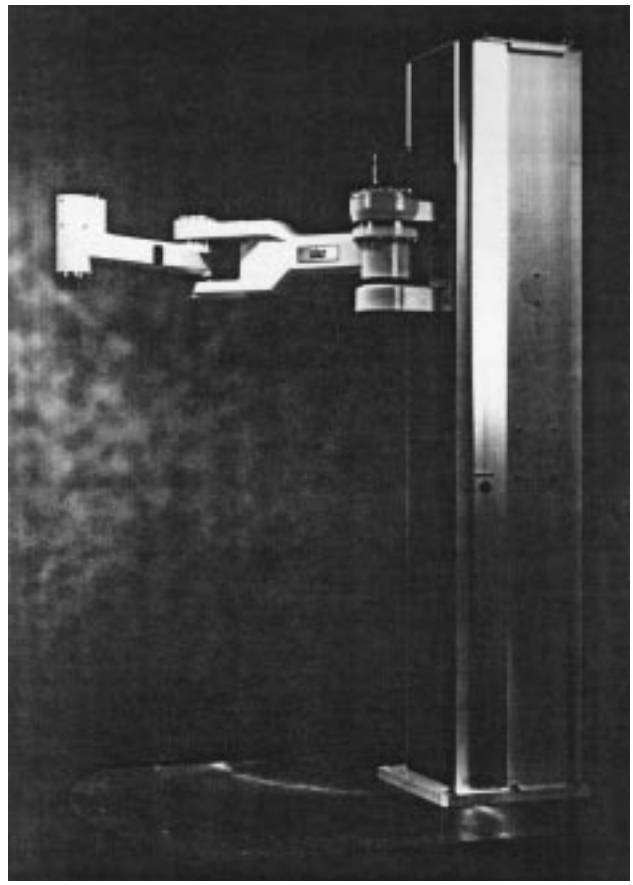
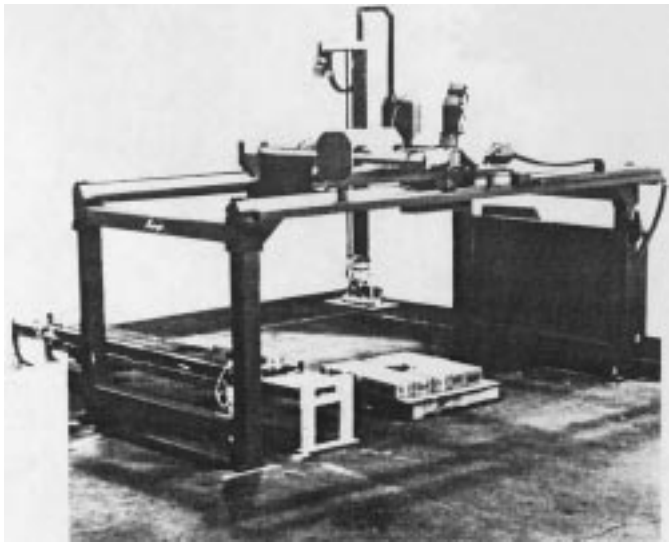


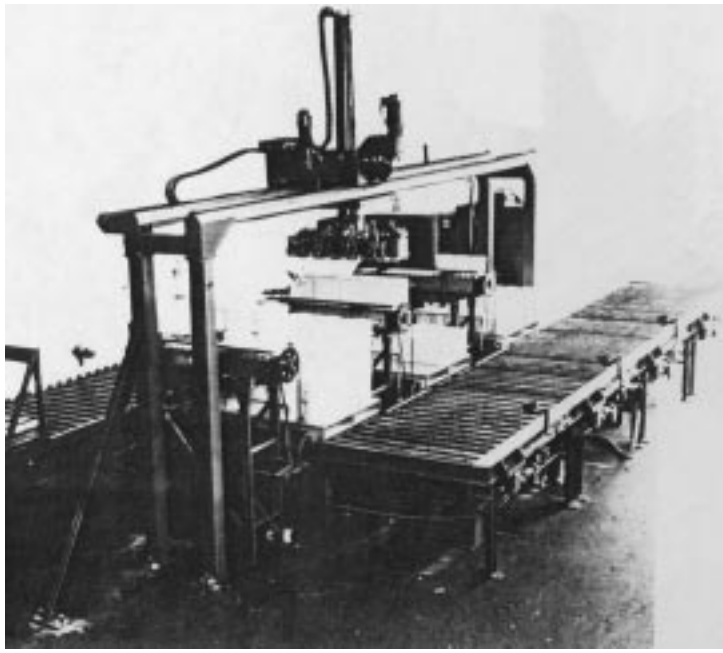
FIGURE 14.2.3 Type 2 SCARA (courtesy of Adept Technologies, Inc.).

Commerical models of spherical and cylindrical robots were originally very common and popular in machine tending and material handling applications. Hundreds are still in use but now there are only a few commercially available models. The Unimate model 2000, a hydraulic-powered spherical coordinate robot, was at one time the most popular robot model in the world. Several models of cylindrical coordinate robots were also available, including a standard model with the largest payload of any robot, the Prab model FC, with a payload of over 600 kg. The decline in use of these two configuations is attributed to problems arising from use of the prismatic link for radial extension/retraction motion. A solid boom requires clearance to fully retract. Hydraulic cylinders used for the same function can retract to less than half of their fully extended length. Type 2 SCARA arms and other revolute jointed arms have displaced most of the cylindrical and spherical coordinate robots (see Figure 14.2.5).

Basic Performance Specifications. Figure 14.2.6 sumarizes the kinematic configurations just described. Table 14.2.1 is a table of basic performance specifications of selected robot models that illustrates the broad spectrum of manipulator performance available from commercial sources. The information contained in the table has been supplied by the respective robot manufacturers. This is not an endorsement by the author or publisher of the robot brands selected, nor is it a verification or validation of the performance values. For more detailed and specific information on the availability of robots, the reader is advised to contact the Robotic Industries Association, 900 Victors Way, P.O. Box 3724, Ann Arbor, MI 48106, or a robot industry trade association in your country for a listing of commercial robot suppliers and system integrators.



(a)

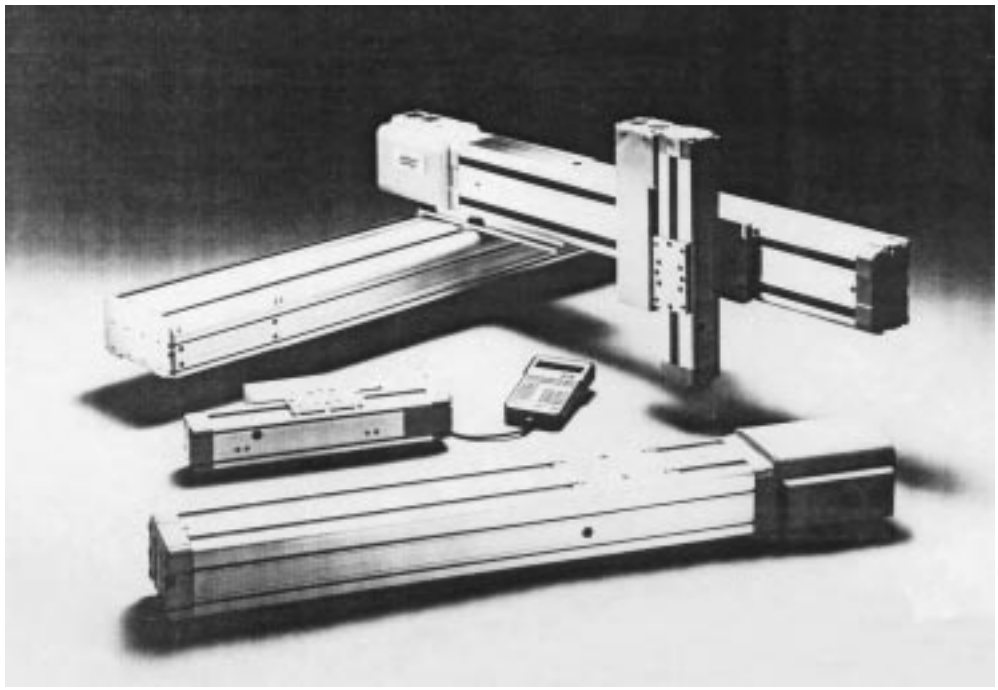


(b)

FIGURE 14.2.4 Cartesian robots. (a) Four-axis gantry robot used for palletizing boxes (courtesy of C&D Robotics, Inc.); (b) three-axis gantry for palletizing (courtesy of C&D Robotics, Inc.); (c) three-axis robot constructed from modular single-axis motion modules (courtesy of Adept Technologies, Inc.).

Drive Types of Commerical Robots

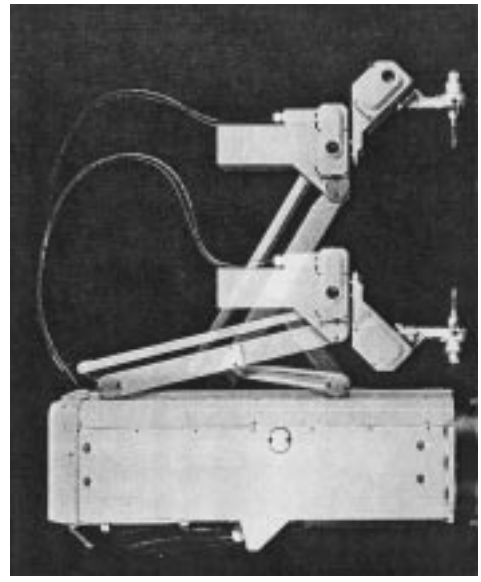
The vast majority of commerical industrial robots uses electric servo motor drives with speed-reducting transmissions. Both AC and DC motors are popular. Some servo hydraulic articulated arm robots are available now for painting applications. It is rare to find robots with servo pneumatic drive axes. All types of mechanical transmissions are used, but the tendency is toward low and zero backlash-type drives. Some robots use direct drive methods to eliminate the amplification of inertia and mechanical backlash associated with other drives. The first axis of the AdeptOne and AdeptThree Type I SCARA



(c)

FIGURE 14.2.4 continued

(a)



(b)

FIGURE 14.2.5 Spherical and cylindrical robots. (a) Hydraulic-powered spherical robot (courtesy Kohol Systems, Inc.); (b) cylindrical arm using scissor mechanism for radial prismatic motion (courtesy of Yamaha Robotics).

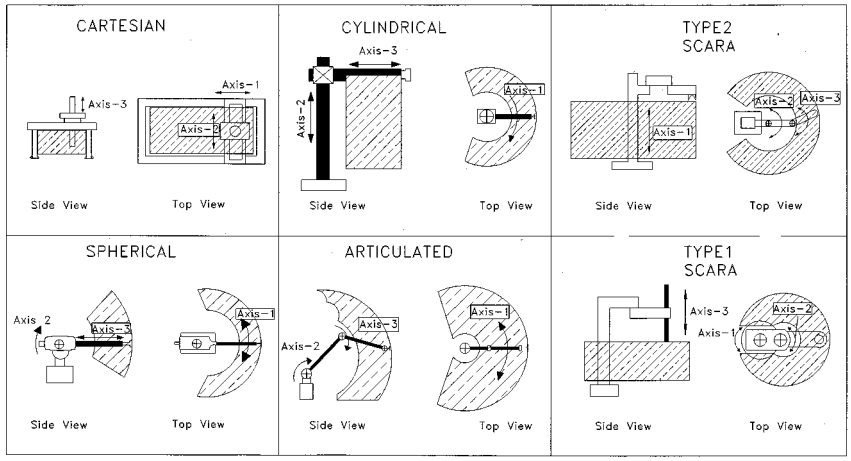


FIGURE 14.2.6 Common kinematic configurations for robots.

TABLE 14.2.1Basic Performance Specifications of Selected Commercial Robots

Configuration	Model	Axes	Payload (kg)	Reach (mm)	Repeatability (mm)	Speed
Articulated	Fanuc M-410i	4	155	3139	+/–0.5	axis 1, 85 deg/sec axis 2, 90 deg/sec axis 3, 100 deg/sec axis 4, 190 deg/sec
	Nachi 8683	6	200	2510	+/–0.5	N/A
	Nachi 7603	6	5	1405	+/–0.1	axis 1, 115 deg/sec axis 2, 115 deg/sec axis 3, 115 deg/sec
	Staubli RX90	6	6	985	+/–0.02	axis 1, 240 deg/sec axis 2, 200 deg/sec axis 3, 286 deg/sec
Type 1 SCARA	AdeptOne	4	9.1	800	+/–0.025	(est.) 1700 mm/sec
	Fanuc A-510	4	20	950	+/–0.065	N/A
Type 2 SCARA	Adept 1850	4	70	1850	X,Y +/–0.3 Z +/–0.3	axis 1, 1500 mm/sec axis 2, 120 deg/sec axis 3, 140 deg/sec axis 4, 225 deg/sec
	Staubli RS 184	4	60	1800	+/–0.15	N/A
Cartesian	PaR Systems XR225	5	190	X 18000 Y 5500 Z 2000	+/–0.125	N/A
	AdeptModules	3	15	X 500 Y 450	+/–0.02	axis 1, 1200 mm/sec axis 2, 1200 mm/sec axis 3, 600 mm/sec
Cylindrical	Kohol K45	4	34	1930	+/–0.2	axis 1, 90 deg/sec axis 2, 500 mm/sec axis 3, 1000 mm/sec
Spherical	Unimation 2000 (Hydraulic, not in production)	5	135		+/–1.25	axis 1, 35 deg/sec axis 2, 35 deg/sec axis 3, 1000 mm/sec

robots is a direct drive motor with the motor stator integrated into the robot base and its armature rotor integral with the first link. Other more common speed-reducing low backlash drive transmissions include toothed belts, roller chains, roller drives, and harmonic drives.

Joint angle position and velocity feedback devices are generally considered an important part of the drive axis. Real-time control performance for tracking position and velocity commands and precision is often affected by the fidelity of feedback. Resolution, signal-to-noise, and innate sampling frequency are important motion control factors ultimately limited by the type of feedback device used.

Given a good robot design, the quality of fabrication and assembly of the drive components must be high to yield good performance. Because of their precision requirements, the drive components are sensitive to manufacturing errors which can readily translate to less than specified manipulator performance.

Commercial Robot Controllers

Commercial robot controllers are specialized multiprocessor computing systems that provide four basic processes allowing integration of the robot into an automation system. These functions which must be factored and weighed for each specific application are Motion Generation, Motion/Process Integration, Human Integration, and Information Integration.

Motion Generation

There are two important controller-related aspects of industrial robot motion generation. One is the extent of manipulation that can be programmed; the other is the ability to execute controlled programmed motion. The unique aspect of each robot system is its real-time kinematic motion control. The details of real-time control are typically not revealed to the user due to safety and proprietary information secrecy reasons. Each robot controller, through its operating system programs, converts digital data into coordinated motion through precise coordination and high speed distribution and communication of the individual axis motion commands which are executed by individual joint controllers. The higher level programming accessed by the end user is a reflection of the sophistication of the real-time controller.

Of greatest importance to the robot user is the motion programming. Each robot manufacturer has its own proprietary programming language. The variety of motion and position command types in a programming language is usually a good indication of the robot's motion generation capability. Program commands which produce complex motion should be available to support the manipulation needs of the application. If palletizing is the application, then simple methods of creating position commands for arrays of positions are essential. If continuous path motion is needed, an associated set of continuous motion commands should be available. The range of motion generation capabilities of commercial industrial robots is wide. Suitability for a particular application can be determined by writing test code.

Motion/Process Integration

Motion/process integration involves methods available to coordinate manipulator motion with process sensor or process controller devices. The most primitive process integration is through discrete digital I/O. For example, an external (to the robot controller) machine controller might send a one-bit signal indicating whether it is ready to be loaded by the robot. The robot control must have the ability to read the signal and to perform logical operations (if then, wait until, do until, etc.) using the signal. At the extreme of process integration, the robot controller can access and operate on large amounts of data in real time during the execution of motion-related processes. For example, in arc welding, sensor data are used to correct tool point positions as the robot is executing a weld path. This requires continuous communication between the welding process sensor and the robot motion generation functions so that there are both a data interface with the controller and motion generation code structure to act on it. Vision-guided high precision pick and place and assembly are major applications in the electronics and semiconductor industries. Experience has shown that the best integrated vision/robot performance has come from running both the robot and the vision system internal to the same computing platform. The

reasons are that data communication is much more efficient due to data bus access, and computing operations are coordinated by one operating system.

Human Integration

Operator integration is critical to the expeditious setup, programming, and maintenance of the robot system. Three controller elements most important for effective human integration are the human *I/O devices*, the information available to the operator in graphic form, and the modes of operation available for human interaction. Position and path teaching effort is dramatically influenced by the type of manual I/O devices available. A teach pendant is needed if the teacher must have access to several vantage points for posing the robot. Some robots have teleoperator-style input devices which allow coordinated manual motion command inputs. These are extremely useful for teaching multiple complex poses. Graphical interfaces, available on some industrial robots, are very effective for conveying information to the operator quickly and efficiently. A graphical interface is most important for applications which require frequent reprogramming and setup changes. Several very useful off-line programming software systems are available from third-party suppliers. These systems use computer models of commercially available robots to simulate path motion and provide rapid programming functions.

Information Integration

Information integration is becoming more important as the trend toward increasing flexibility and agility impacts robotics. Automatic and computer-aided robot task planning and process control functions will require both access to data and the ability to resolve relevant information from CAD systems, process plans and schedules, upstream inspections, and other sources of complex data and information. Many robot controllers now support information integration functions by employing integrated PC interfaces through the communications ports, or in some through direct connections to the robot controller data bus.

14.3 Robot Configurations

Ian D. Walker

Fundamentals and Design Issues

A robot manipulator is fundamentally a collection of *links* connected to each other by *joints*, typically with an *end effector* (designed to contact the environment in some useful fashion) connected to the mechanism. A typical arrangement is to have the links connected serially by the joints in an open-chain fashion. Each joint provides one or more degree of freedom to the mechanism.

Manipulator designs are typically characterized by the number of independent degrees of freedom in the mechanism, the types of joints providing the degrees of freedom, and the geometry of the links connecting the joints. The degrees of freedom can be revolute (relative rotational motion θ between joints) or prismatic (relative linear motion d between joints). A joint may have more than one degree of freedom. Most industrial robots have a total of six independent degrees of freedom. In addition, most current robots have essentially rigid links (we will focus on rigid-link robots throughout this section).

Robots are also characterized by the type of actuators employed. Typically manipulators have hydraulic or electric actuation. In some cases where high precision is not important, pneumatic actuators are used.

A number of successful manipulator designs have emerged, each with a different arrangement of joints and links. Some "elbow" designs, such as the PUMA robots and the SPAR Remote Manipulator System, have a fairly anthropomorphic structure, with revolute joints arranged into "shoulder," "elbow," and "wrist" sections. A mix of revolute and prismatic joints has been adopted in the Stanford Manipulator and the SCARA types of arms. Other arms, such as those produced by IBM, feature prismatic joints for the "shoulder," with a spherical wrist attached. In this case, the prismatic joints are essentially used as positioning devices, with the wrist used for fine motions.

The above designs have six or fewer degrees of freedom. More recent manipulators, such as those of the Robotics Research Corporation series of arms, feature seven or more degrees of freedom. These arms are termed kinematically redundant, which is a useful feature as we will see later.

Key factors that influence the design of a manipulator are the tractability of its geometric (kinematic) analysis and the size and location of its workspace. The workspace of a manipulator can be defined as the set of points that are reachable by the manipulator (with fixed base). Both shape and total volume are important. Manipulator designs such as the SCARA are useful for manufacturing since they have a simple semicylindrical connected volume for their workspace (Spong and Vidyasagar, 1989), which facilitates workcell design. Elbow manipulators tend to have a wider volume of workspace, however the workspace is often more difficult to characterize. The kinematic design of a manipulator can tailor the workspace to some extent to the operational requirements of the robot.

In addition, if a manipulator can be designed so that it has a simplified kinematic analysis, many planning and control functions will in turn be greatly simplified. For example, robots with spherical wrists tend to have much simpler inverse kinematics than those without this feature. Simplification of the kinematic analysis required for a robot can significantly enhance the real-time motion planning and control performance of the robot system. For the rest of this section, we will concentrate on the kinematics of manipulators.

For the purposes of analysis, a set of *joint variables* (which may contain both revolute and prismatic variables), are augmented into a vector q, which uniquely defines the geometric state, or configuration of the robot. However, task description for manipulators is most naturally expressed in terms of a different set of *task coordinates*. These can be the position and orientation of the robot end effector, or of a special task frame, and are denoted here by Y. Thus Y most naturally represents the performance of a task, and q most naturally represents the mechanism used to perform the task. Each of the coordinate systems q and Y contains information critical to the understanding of the overall status of the manipulator. Much of the kinematic analysis of robots therefore centers on transformations between the various sets of coordinates of interest.

Manipulator Kinematics

The study of manipulator kinematics at the position (geometric) level separates naturally into two subproblems: (1) finding the position/orientation of the end effector, or task, frame, given the angles and/or displacements of the joints *(Forward Kinematics);* and (2) finding possible angles/displacements of the joints given the position/orientation of the end effector, or task, frame *(Inverse Kinematics)*. At the velocity level, the *Manipulator Jacobian* relates joint velocities to end effector velocities and is important in motion planning and for identifying *Singularities*. In the case of *Redundant Manipulators*, the Jacobian is particularly crucial in planning and controlling robot motions. We will explore each of these issues in turn in the following subsections.

Example 14.3.1

Figure 14.3.1 shows a planar three-degrees-of-freedom manipulator. The first two joints are revolute, and the third is prismatic. The end effector position (x, y) is expressed with respect to the (fixed) world coordinate frame (x_0, y_0), and the orientation of the end effector is defined as the angle of the second link ϕ measured from the x_0 axis as shown. The link length l_1 is constant. The joint variables are given by the angles θ_1 and θ_2 and the displacement d_3, and are defined as shown. The example will be used throughout this section to demonstrate the ideas behind the various kinematic problems of interest.

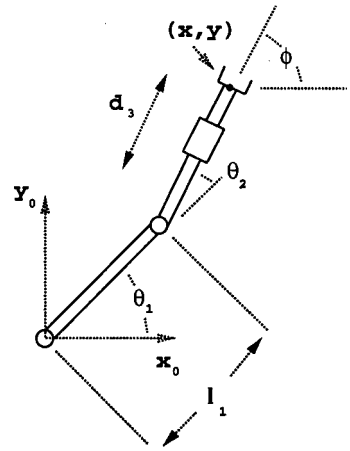


FIGURE 14.3.1 Planar RRP manipulator.

Forward (Direct) Kinematics

Since robots typically have sensors at their joints, making available measurements of the joint configurations, and we are interested in performing tasks at the robot end effector, a natural issue is that of determining the end effector position/orientation Y given a joint configuration q. This problem is the forward kinematics problem and may be expressed symbolically as

$$Y = f(q) \tag{14.3.1}$$

The forward kinematic problem yields a unique solution for Y given q. In some simple cases (such as the example below) the forward kinematics can be derived by inspection. In general, however, the relationship f can be quite complex. A systematic method for determining the function f for any manipulator geometry was proposed by Denavit and Hartenberg (Denavit and Hartenberg, 1955).

The Denavit/Hartenberg (or D-H) technique has become the standard method in robotics for describing the forward kinematics of a manipulator. Essentially, by careful placement of a series of coordinate

frames fixed in each link, the D-H technique reduces the forward kinematics problem to that of combining a series of straightforward consecutive link-to-link transformations from the base to the end effector frame. Using this method, the forward kinematics for any manipulator is summarized in a table of parameters (the D-H parameters). A maximum of three nonzero parameters per link are sufficient to uniquely specify the map *f*. Lack of space prevents us from detailing the method further. The interested reader is referred to Denavit and Hartenberg (1955) and Spong and Vidyasagar (1989).

To summarize, forward kinematics is an extremely important problem in robotics which is also well understood, and for which there is a standard solution technique

Example 14.3.2

In our example, we consider the task space to be the position and orientation of the end effector, i.e., $Y = [x, y, \phi]^T$ as shown. We choose joint coordinates (one for each degree of freedom) by $q = [\theta_1, \theta_2, d_3]^T$. From Figure 14.3.1, with the values as given it may be seen by inspection that

$$x = l_1 \cos(\theta_1) + d_3 \cos(\theta_1 + \theta_2) \tag{14.3.2}$$

$$y = l_1 \sin(\theta_1) + d_3 \sin(\theta_1 + \theta_2) \tag{14.3.3}$$

$$\phi = \theta_1 + \theta_2 \tag{14.3.4}$$

Equations (14.3.2) to (14.3.4) form the forward kinematics for the example robot. Notice that the solution for $Y = [x, y, \phi]^T$ is unique given $q = [\theta_1, \theta_2, d_3]^T$.

Inverse Kinematics

The *inverse kinematics* problem consists of finding possible joint configurations q corresponding to a given end effector position/orientation Y. This transformation is essential for planning joint positions of the manipulator which will result in desired end effector positions (note that task requirements will specify Y, and a corresponding q must be planned to perform the task). Conceptually the problem is stated as

$$q = f^{-1}(Y) \tag{14.3.5}$$

In contrast to the forward kinematics problem, the inverse kinematics cannot be solved for arbitrary manipulators by a systematic technique such as the Denavit-Hartenberg method. The relationship (1) does not, in general, invert to a unique solution for q, and, indeed, for many manipulators, expressions for q cannot even be found in closed form!

For some important types of manipulator design (particularly those mechanisms featuring spherical wrists), closed-form solutions for the inverse kinematics can be found. However, even in these cases, there are at best multiple solutions for q (corresponding to "elbow-up," "elbow-down" possibilities for the arm to achieve the end effector configuration in multiple ways). For some designs, there may be an infinite number of solutions for q given Y, such as in the case of kinematically redundant manipulators discussed shortly.

Extensive investigations of manipulator kinematics have been performed for wide classes of robot designs (Bottema and Roth, 1979; Duffy, 1980). A significant body of work has been built up in the area of inverse kinematics. Solution techniques are often determined by the geometry of a given manipulator design. A number of elegant techniques have been developed for special classes of manipulator designs, and the area continues to be the focus of active research. In cases where closed-form solutions cannot be found, a number of iterative numerical techniques have been developed.

Example 14.3.3

For our planar manipulator, the inverse kinematics requires the solution for $q = [\theta_1, \theta_2, d_3]^T$ given $Y = [x, y, \phi]^T$. Figure 14.3.2 illustrates the situation, with $[x, y, \phi]^T$ given as shown. Notice that for the Y specified in Figure 14.3.2, there are two solutions, corresponding two distinct configurations q.

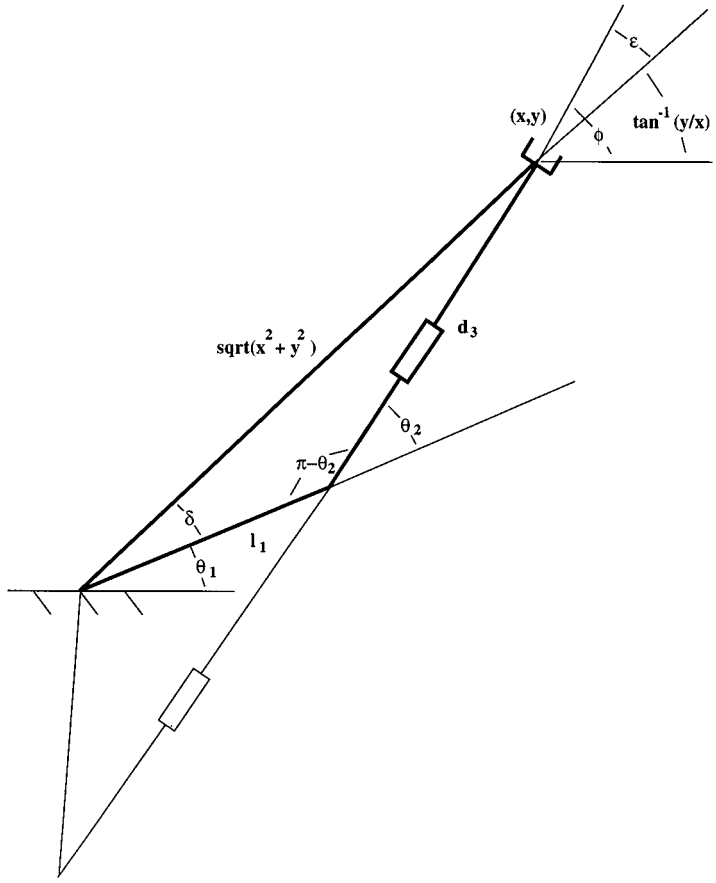


FIGURE 14.3.2 Planar RRP arm inverse kinematics.

The two solutions are sketched in Figure 14.3.2, with the solution for the configuration in bold the focus of the analysis below. The solutions may be found in a number of ways, one of which is outlined here. Consider the triangle formed by the two links of the manipulator and the vector (x, y) in Figure 14.3.2. We see that the angle ε can be found as

$$\varepsilon = \phi - \tan^{-1}(y/x)$$

Now, using the sine rule, we have that

$$l_1/\sin(\varepsilon) = \left(\sqrt{x^2 + y^2}\right)\Big/\sin(\pi - \theta_2) = \left(\sqrt{x^2 + y^2}\right)\Big/\sin(\theta_2)$$

and thus

$$\sin(\theta_2) = \left(\sqrt{x^2 + y^2}\right)\sin(\varepsilon)/l_1$$

The above equation could be used to solve for θ_2. Alternatively, we can find θ_2 as follows. Defining D to be $(\sqrt{x^2 + y^2})\sin(\varepsilon)/l_1$ we have that $\cos(\theta_2) = \pm\sqrt{1 - D^2}$. Then θ_2 can be found as

$$\theta_2 = \tan^{-1}\left[D/\pm\sqrt{1 - D^2}\right] \tag{14.3.6}$$

Notice that this method picks out both possible values of θ_2, corresponding to the two possible inverse kinematic solutions. We now take the solution for θ_2 corresponding to the positive root of $\pm(\sqrt{1 - D^2})$ (i.e., the bold robot configuration in the figure).

Using this solution for θ_2, we can now solve for θ_1 and d_3 as follows. Summing the angles inside the triangle in Figure 14.3.2, we obtain $\pi - [(\pi - \theta_2) + \varepsilon + \delta] = 0$ or

$$\delta = \theta_2 - \varepsilon$$

From Figure 14.3.2 we see that

$$\theta_1 = \tan^{-1}(y/x) - \delta \tag{14.3.7}$$

Finally, use of the cosine rule leads us to a solution for d_3:

$$d_3^2 = l_1^2 + \left(x^2 + y^2\right) - 2l_1\left(\sqrt{x^2 + y^2}\right)\cos(\delta)$$

or

$$d_3 = \sqrt{l_1^2 + \left(x^2 + y^2\right) - 2l_1\left(\sqrt{x^2 + y^2}\right)\cos(\delta)} \tag{14.3.8}$$

Equations (14.3.6) to (14.3.8) comprise an inverse kinematics solution for the manipulator.

Velocity Kinematics: The Manipulator Jacobian

The previous techniques, while extremely important, have been limited to positional analysis. For motion planning purposes, we are also interested in the relationship between joint velocities and task (end effector) velocities. The (linearized) relationship between the joint velocities $\dot{q}$ and the end effector velocities $\dot{Y}$ can be expressed (from Equation (14.3.1)) as

$$\dot{Y} = \left[J(q)\right]\dot{q} \tag{14.3.9}$$

where J is the *manipulator Jacobian* and is given by $\partial f/\partial q$. The manipulator Jacobian is an extremely important quantity in robot analysis, planning, and control. The Jacobian is particularly useful in determining singular configurations, as we shall see shortly.

Given the forward kinematic function f, the Jacobian can be obtained by direct differentiation (as in the example below). Alternatively, the Jacobian can be obtained column by column in a straightforward fashion from quantities in the Denavit-Hartenberg formulation referred to earlier. Since the Denavit-Hartenberg technique is almost always used in the forward kinematics, this is often an efficient and preferred method. For more details of this approach, see Spong and Vidyasagar (1989).

The Jacobian can be used to perform inverse kinematics at the velocity level as follows. If we define $[J^{-1}]$ to be the inverse of the Jacobian (assuming J is square and nonsingular), then

$$\dot{q} = \left[J^{-1}(q)\right]\dot{Y} \tag{14.3.10}$$

and the above expression can be solved iteratively for $\dot{q}$ (and hence q by numerical integration) given a desired end effector trajectory $\dot{Y}$ and the current state q of the manipulator. This method for determining joint trajectories given desired end effector trajectories is known as Resolved Rate Control and has become increasingly popular. The technique is particularly useful when the positional inverse kinematics is difficult or intractable for a given manipulator.

Notice, however, that the above expression requires that *J is* both nonsingular and square. Violation of the nonsingularity assumption means that the robot is in a singular configuration, and if J has more columns than rows, then the robot is kinematically redundant. These two issues will be discussed in the following subsections.

Example 14.3.4

By direct differentiation of the forward kinematics derived earlier for our example,

$$\begin{bmatrix} \dot{x} \\ \dot{y} \\ \dot{\phi} \end{bmatrix} = \begin{bmatrix} -l_1 \sin(\theta_1) - d_3 \sin(\theta_1 + \theta_2) & -d_3 \sin(\theta_1 + \theta_2) & \cos(\theta_1 + \theta_2) \\ l_1 \cos(\theta_1) + d_3 \cos(\theta_1 + \theta_2) & d_3 \cos(\theta_1 + \theta_2) & \sin(\theta_1 + \theta_2) \\ 1 & 1 & 0 \end{bmatrix} \begin{bmatrix} \dot{\theta}_1 \\ \dot{\theta}_2 \\ \dot{d}_3 \end{bmatrix} \tag{14.3.11}$$

Notice that each column of the Jacobian represents the (instantaneous) effect of the corresponding joint on the end effector motions. Thus, considering the third column of the Jacobian, we confirm that the third joint (with variable d_3) cannot cause any change in the orientation (ϕ) of the end effector.

Singularities

A significant issue in kinematic analysis surrounds so-called singular configurations. These are defined to be configurations q_s at which $J(q_s)$ has less than full rank (Spong and Vidyasagar, 1989). Physically, these configurations correspond to situations where the *robot joints* have been aligned in such a way that there is at least one direction of motion (the singular direction[s]) for the end effector that physically cannot be achieved by the mechanism. This occurs at workspace boundaries, and when the axes of two (or more) joints line up and are redundantly contributing to an end effector motion, at the cost of another end effector degree of freedom being lost. It is straightforward to show that the singular direction is orthogonal to the column space of $J(q_s)$.

It can also be shown that every manipulator must have singular configurations, i.e., the existence of singularities cannot be eliminated, even by careful design. Singularities are a serious cause of difficulties in robotic analysis and control. Motions have to be carefully planned in the region of singularities. This is not only because at the singularities themselves there will be an unobtainable motion at the end effector, but also because many real-time motion planning and control algorithms make use of the (inverse of the) manipulator Jacobian. In the region surrounding a *singularity*, the Jacobian will become ill-conditioned, leading to the generation of joint velocities in Equation (14.3.10) which are extremely high, even for relatively small end effector velocities. This can lead to numerical instability, and unexpected wild motions of the arm for small, desired end effector motions (this type of behavior characterizes motion near a singularity).

For the above reasons, the analysis of singularities is an important issue in robotics and continues to be the subject of active research.

Example 14.3.5

For our example manipulator, we can find the singular configurations by taking the determinant of its Jacobian found in the previous section and evaluating the joint configurations that cause this determinant to become zero. A straightforward calculation yields

$$\det(J) = l_1 \cos(\theta_1) \tag{14.3.12}$$

and we note that this determinant is zero exactly when θ_1 is a multiple of $\pi/2$. One such configuration ($\theta_1 = \pi/2$, $\theta_2 = -\pi/2$) is shown in Figure 14.3.3. For this configuration, with $l_1 = 1 = d_3$, the Jacobian is given by

$$\begin{bmatrix} -1 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 1 & 0 \end{bmatrix}$$

and by inspection, the columns of J are orthogonal to $[0, -1, 1]^T$, which is therefore a singular direction of the manipulator in this configuration. This implies that from the (singular) configuration shown in Figure 14.3.3, the direction $\dot{Y} = [0, -1, 1]^T$ cannot be physically achieved. This can be confirmed by considering the physical device (motion in the negative y direction cannot be achieved while simultaneously increasing the orientation angle ϕ).

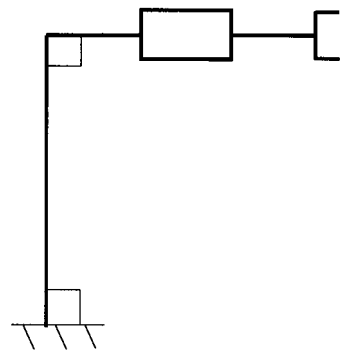


FIGURE 14.3.3 Singular configuration of planar RRP arm.

Redundant Manipulator Kinematics

If the dimension of q is n, the dimension of Y is m, and n is larger than m, then a manipulator is said to be *kinematically redundant* for the task described by Y. This situation occurs for a manipulator with seven or more degrees of freedom when Y is a six-dimensional position/orientation task, or, for example, when a six-degrees-of-freedom manipulator is performing a position task and orientation is not specified.

In this case, the robot mechanism has more degrees of freedom than required by the task. This gives rise to extra complexity in the kinematic analysis due to the extra joints. However, the existence of these extra joints gives rise to the extremely useful *self-motion* property inherent in redundant arms. A self-motion occurs when, with the end effector location held constant, the joints of the manipulator can move (creating an "orbit" of the joints). This allows a much wider variety of configurations (typically an infinite number) for a given end effector location. This added maneuverability is the key feature and advantage of kinematically redundant arms. Note that the human hand/arm has this property. The key question for redundant arms is how to best utilize the self-motion property while still performing specified

end effector motions $\dot{Y}$. A number of motion-planning algorithms have been developed in the last few years for redundant arms (Siciliano, 1990). Most of them center on the Jacobian pseudoinverse as follows.

For kinematically redundant arms, the Jacobian has more columns than rows. If J is of full rank, and we choose $[J^+]$ to be a pseudoinverse of the Jacobian such that $JJ^+ = I$ [for example $J^+ = J^T(JJ^T)^{-1}$], where I is the $m \times m$ identity matrix, then from Equation (14.3.9) a solution for $\dot{q}$ which satisfies end effector velocity of $\dot{Y}$ is given by

$$\dot{q} = \left[J^+(q)\right]\dot{Y} + \left[I - J^+(q)\,J(q)\right]\varepsilon \tag{14.3.13}$$

where ε *is* an $(n \times 1)$ column vector whose values may be arbitrarily selected. Note that conventional nonredundant manipulators have $m = n$, in which case the pseudoinverse becomes J^{-1} and the problem reduces to the resolved rate approach (Equation 14.3.10).

The above solution for $\dot{q}$ has two components. The first component, $[J^+(q)]\,\dot{Y}$, are joint velocities that produce the desired end effector motion $\dot{Y}$ (this can be easily seen by substitution into Equation (14.3.9)). The second term, $[I - J^+(q)J(q)]\varepsilon$, comprises joint velocities which produce no end effector velocities (again, this can be seen by substitution of this term into Equation (14.3.9)). Therefore, the second term produces a self-motion of the arm, which can be tuned by appropriately altering ε. Thus different choices of ε correspond to different choices of the self-motion and various algorithms have been developed to exploit this choice to perform useful subtasks (Siciliano, 1990).

Redundant manipulator analysis has been an active research area in the past few years. A number of arms, such as those recently produced by Robotics Research Corporation, have been designed with seven degrees of freedom to exploit kinematic redundancy. The self-motion in redundant arms can be used to configure the arm to evade obstacles, avoid singularities, minimize effort, and a great many more subtasks in addition to performing the desired main task described by $\dot{Y}$. For a good review of the area, the reader is referred to Siciliano (1990).

Example 14.3.6

If, for our example, we are only concerned with the position of the end effector in the plane, then the arm becomes kinematically redundant. Figure 14.3.4 shows several different (from an infinite number of) configurations for the arm given one end effector position. In this case, J becomes the 2×3 matrix formed by the top two rows of the Jacobian in Equation (14.3.11). The pseudoinverse J^+ will therefore be a 3×2 matrix. Formation of the pseudoinverse is left to the reader as an exercise.

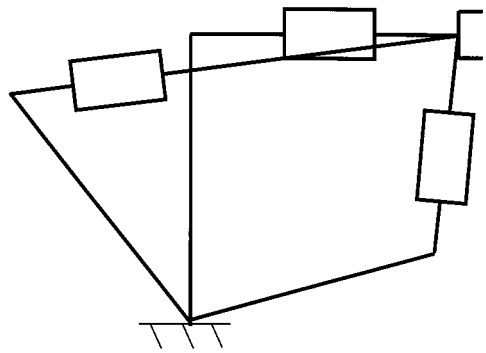


FIGURE 14.3.4 Multiple configurations for RRP arm for specified end effector position only.

Summary

Kinematic analysis is an interesting and important area, a solid understanding of which is required for robot motion planning and control. A number of techniques have been developed and are available to the robotics engineer. For positional analysis, the Denavit-Hartenberg technique provides a systematic approach for forward kinematics. Inverse kinematic solutions typically have been developed on a manipulator (or class of manipulator)-specific basis. However, a number of insightful effective techniques exist for positional inverse kinematic analysis. The manipulator Jacobian is a key tool for analyzing singularities and motion planning at the velocity level. Its use is particularly critical for the emerging generation of kinematically redundant arms

14.4 End Effectors and Tooling

Mark R. Cutkosky and Peter McCormick

End effectors or end-of-arm tools are the devices through which a robot interacts with the world around it, grasping and manipulating parts, inspecting surfaces, and working on them. As such, end effectors are among the most important elements of a robotic application — not "accessories" but an integral component of the overall tooling, fixturing, and sensing strategy. As robots grow more sophisticated and begin to work in more demanding applications, end effector design is becoming increasingly important.

The purpose of this chapter is to introduce some of the main types of end effectors and tooling and to cover issues associated with their design and selection. References are provided for the reader who wishes to go into greater depth on each topic. For those interested in designing their own end effectors, a number of texts including Wright and Cutkosky (1985) provide additional examples.

A Taxonomy of Common End Effectors

Robotic end effectors today include everything from simple two-fingered grippers and vacuum attachments to elaborate multifingered hands. Perhaps the best way to become familiar with end effector design issues is to first review the main end effector types.

Figure 14.4.1 is a taxonomy of common end effectors. It is inspired by an analogous taxonomy of grasps that humans adopt when working with different kinds of objects and in tasks requiring different amounts of precision and strength (Wright and Cutkosky, 1985). The left side includes "passive" grippers that can hold parts, but cannot manipulate them or actively control the grasp force. The right-hand side includes active servo grippers and *dextrous* robot hands found in research laboratories and teleoperated applications.

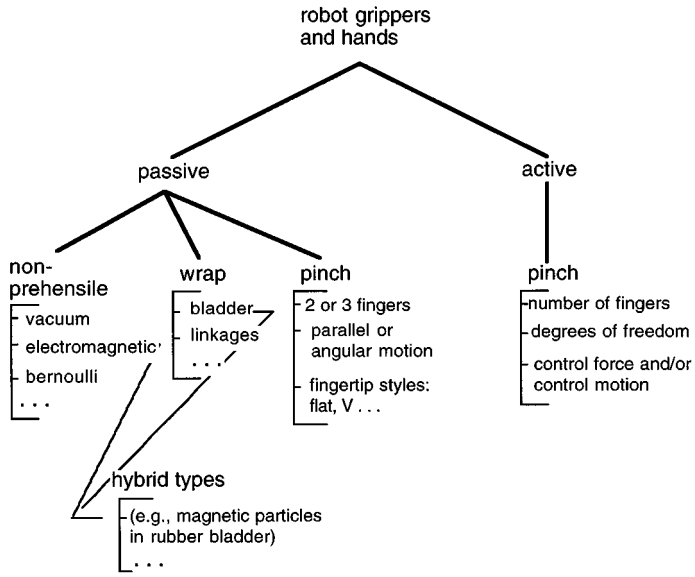


FIGURE 14.4.1 A taxonomy of the basic end effector types.

Passive End Effectors

Most end effectors in use today are passive; they emulate the grasps that people use for holding a heavy object or tool, without manipulating it in the fingers. However, a passive end effector may (and generally should) be equipped with sensors, and the information from these sensors may be used in controlling the robot arm.

The left-most branch of the "passive" side of the taxonomy includes vacuum, electromagnetic, and Bernoulli-effect end effectors. Vacuum grippers, either singly or in combination, are perhaps the most commonly used gripping device in industry today. They are easily adapted to a wide variety of parts — from surface mount microprocessor chips and other small items that require precise placement to large, bulky items such as automobile windshields and aircraft panels. These end effectors are classified as "nonprehensile" because they neither enclose parts nor apply grasp forces across them. Consequently, they are ideal for handling large and delicate items such as glass panels. Unlike grippers with fingers, vacuum grippers to not tend to "center" or relocate parts as they pick them up. As discussed in Table 14.4.1, this feature can be useful when initial part placement is accurate.

TABLE 14.4.1 Task Considerations in End Effector Design

Initial Accuracy. Is the initial accuracy of the part high (as when retrieving a part from a fixture or lathe chuck) or low (as when picking unfixtured components off a conveyor)? In the former case, design the gripper so that it will conform to the part position and orientation (as do the grippers in Figures 14.4.5 and 14.4.6. In the latter case, make the gripper center the part (as will most parallel-jaw grippers).
Final Accuracy. Is the final accuracy of the part high or low? In the former case (as when putting a precisely machined peg into a chamfered hole) the gripper and/or robot arm will need *compliance*. In the latter case, use an end effector that centers the part.
Anticipated Forces. What are the magnitudes of the expected task forces and from what directions will they come? Are these forces resisted directly by the gripper jaws, or indirectly through friction? High forces may lead to the adoption of a "wrap"-type end effector that effectively encircles the part or contacts it at many points.
Other Tasks. Is it useful to add sensing or other tooling at the end effector to reduce cycle time? Is it desirable for the robot to carry multiple parts to minimize cycle time? In such cases consider *compound end effectors*.
Speed and Cycle Time. Are speeds and accelerations large enough that inertial forces and moments should be considered in computing the required grip force?

If difficulties are encountered with a vacuum gripper, it is helpful to remember that problem can be addressed in several ways, including increasing the suction cup area through larger cups or multiple cups, redesigning the parts to be grasped so that they present a smoother surface (perhaps by affixing smooth tape to a surface), and augmenting suction with grasping as discussed below. Figure 14.4.2 shows a large gripper with multiple suction cups for handling thermoplastic auto body panels. This end effector also has pneumatic actuators for providing local left/right and up/down motions.

An interesting noncontact variation on the vacuum end effector is illustrated in Figure 14.4.3. This end effector is designed to lift and transport delicate silicon wafers. It lifts the wafers by blowing gently on them from above so that aerodynamic lift is created via the Bernoulli effect. Thin guides around the periphery of the wafers keep them centered beneath the air source.

The second branch of end effector taxonomy includes "wrap" grippers that hold a part in the same way that a person might hold a heavy hammer or a grapefruit. In such applications, humans use *wrap grasps* in which the fingers envelop a part, and maintain a nearly uniform pressure so that friction is used to maximum advantage. Figures 14.4.4 and 14.4.5 show two kinds of end effectors that achieve a similar effect.

Another approach to handling irregular or soft objects is to augment a vacuum or magnetic gripper with a bladder containing particles or a fluid. When handling ferrous parts, one can employ an electromagnet and iron particles underneath a membrane. Still another approach is to use fingertips filled with an electrorheological fluid that stiffens under the application of an electrostatic field.

The middle branch of the end effector taxonomy includes common two-fingered grippers. These grippers employ a strong "pinch" force between two fingers, in the same way that a person might grasp a key when opening a lock. Most such grippers are sold without fingertips since they are the most product-specific part of the design. The fingertips are designed to match the size of components, the

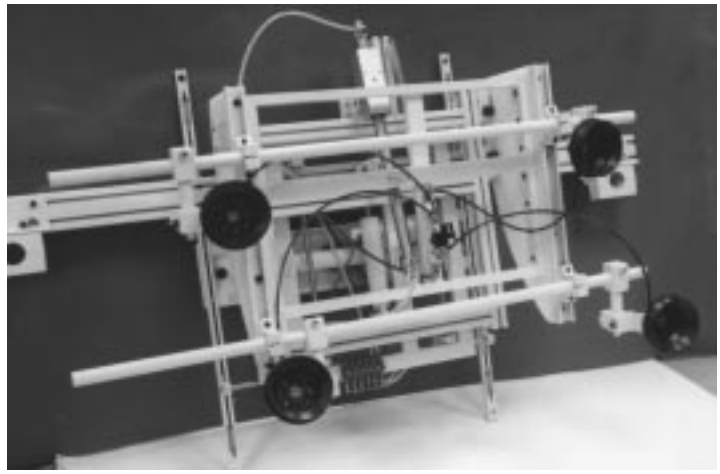


FIGURE 14.4.2 A large end effector for handling autobody panels with actuators for local motions. (Photo courtesy of EOA Systems Inc., Dallas, TX.)

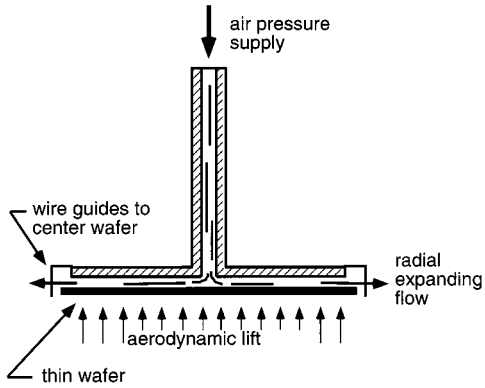


FIGURE 14.4.3 A noncontact end effector for acquiring and transporting delicate wafers.

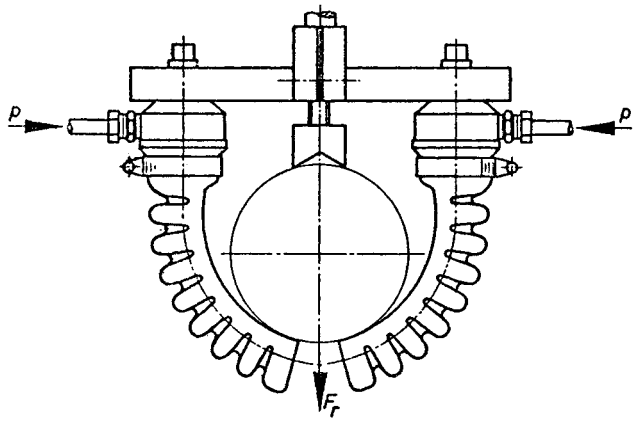


FIGURE 14.4.4 A compliant pneumatic gripper that executes a gentle wrap grasp. (From U.S. Patent No. 3981528, Simrit Corp., Arlington Hts., IL, 1984.)

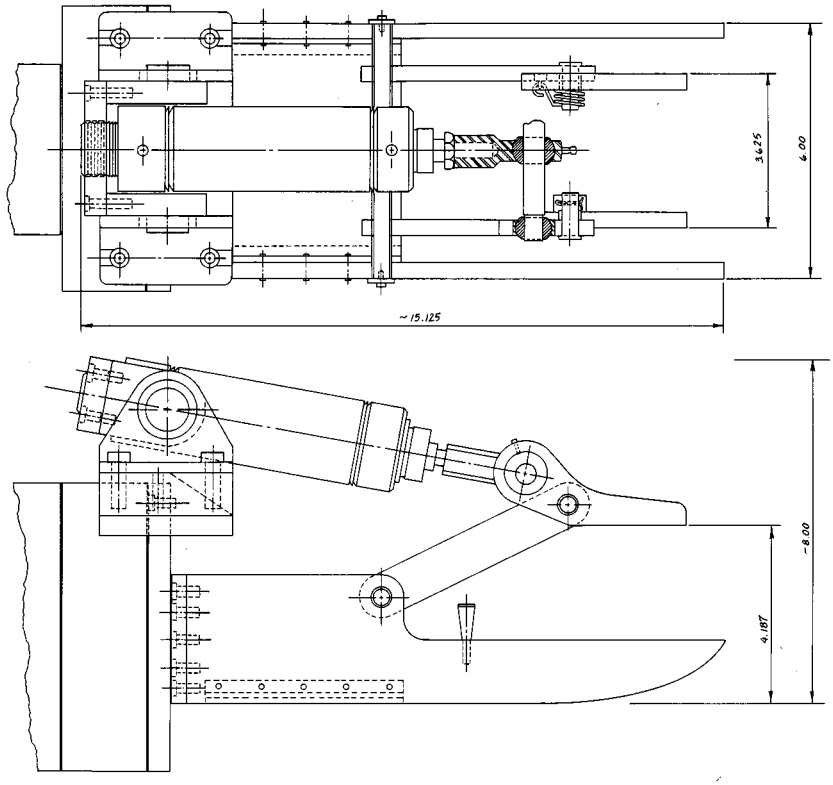


FIGURE 14.4.5 A gripper with pivoted fingers designed to conform to the position and orientation of heavy, irregular parts and to hold them securely. (From U.S. Patent No. 4,545,722, Cutkosky and Kurokawa, 1985.)

shape of components (e.g., flat or V-grooved for cylindrical parts), and the material (e.g., rubber or plastic to avoid damaging fragile objects).

Note that since two-fingered end effectors typically use a single air cylinder or motor that operates both fingers in unison, they will tend to center parts that they grasp. This means that when they grasp constrained parts (e.g., pegs that have been set in holes or parts held in fixtures) some compliance must be added, perhaps with a compliant wrist as discussed in "Wrists and Other End-of-Arm Tooling" below.

Active End Effectors and Hands

The right-hand branch of the taxonomy includes servo grippers and dextrous multifingered hands. Here the distinctions depend largely on the number of fingers and the number of joints or degrees of freedom per finger. For example, the comparatively simple two-fingered servo gripper of Figure 14.4.6 is confined to *"pinch" grasps*, like commercial two-fingered grippers.

Servo-controlled end effectors provide advantages for fine-motion tasks. In comparison to a robot arm, the fingertips are small and light, which means that they can move quickly and precisely. The total range of motion is also small, which permits fine-resolution position and velocity measurements. When equipped with force sensors such as strain gages, the fingers can provide force sensing and control, typically with better accuracy than can be obtained with robot wrist- or joint-mounted sensors. A servo gripper can also be programmed either to control the position of an unconstrained part or to accommodate to the position of a constrained part as discussed in Table 14.4.1.

The sensors of a servo-controlled end effector also provide useful information for robot programming. For example, position sensors can be used to measure the width of a grasped component, thereby providing a check that the correct component has been grasped. Similarly, force sensors are useful for weighing grasped objects and monitoring task-related forces.

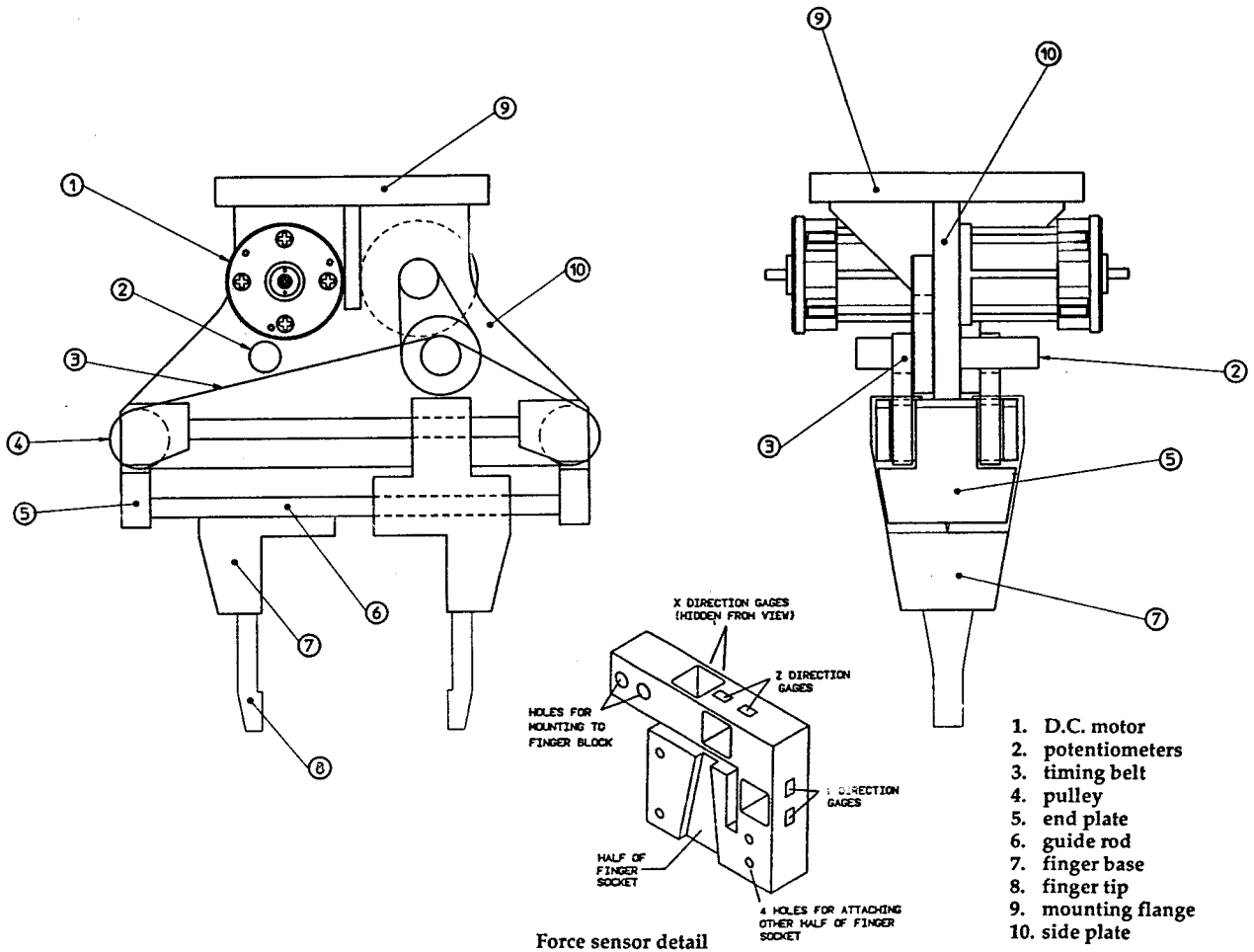


FIGURE 14.4.6 A two-finger servo gripper with force sensing and changeable fingertips. (From E. Pearce et. al, ME210 Report, Stanford University, 1987.)

For applications requiring a combination of dexterity and versatility for grasping a wide range of objects, a dextrous multifingered hand is the ultimate solution. A number of multifingered hands have been described in the literature (see, for example, Jacobsen et al. [1984]) and commercial versions are available. Most of these hands are frankly anthropomorphic, although kinematic criteria such as workspace and *grasp isotropy* (basically a measure of how accurately motions and forces can be controlled in different directions) have also been used.

Despite their practical advantages, dextrous hands have thus far been confined to a few research laboratories. One reason is that the design and control of such hands present numerous difficult tradeoffs among cost, size, power, flexibility and ease of control. For example, the desire to reduce the dimensions of the hand, while providing adequate power, leads to the use of cables that run through the wrist to drive the fingers. These cables bring attendant control problems due to elasticity and friction (Jacobsen et al., 1984).

A second reason for slow progress in applying dextrous hands to manipulation tasks is the formidable challenge of programming and controlling them. The equations associated with several fingertips sliding and rolling on a grasped object are complex — the problem amounts to coordinating several little robots at the end of a robot. In addition, the mechanics of the hand/object system are sensitive to variations in the contact conditions between the fingertips and object (e.g., variations in the object profile and local coefficient of friction). Moreover, during manipulation the fingers are continually making and breaking contact with the object, starting and stopping sliding, etc., with attendant changes in the dynamic and kinematic equations which must be accounted for in controlling the hand. A survey of the dextrous manipulation literature can be found in Pertin-Trocaz (1989).

Wrists and Other End-of-Arm Tooling

In many applications, an active servo gripper is undesirably complicated, fragile, and expensive, and yet it is desirable to obtain some of the compliant force/motion characteristics that an actively controlled gripper can provide. For example, when assembling close-fitting parts, compliance at the end effector can prevent large contact forces from arising due to minor position errors of the robot or manufacturing tolerances in the parts themselves. For such applications a compliant wrist, mounted between the gripper and the robot arm, may be the solution. In particular, *remote center of compliance (RCC)* wrists allow the force/deflection properties of the end effector to be tailored to suit a task. Active wrists have also been developed for use with end effectors for precise, high-bandwidth control of forces and fine motions (Hollis et al., 1988).

Force sensing and quick-change wrists are also commercially available. The former measure the interaction forces between the end effector and the environment and typically come with a dedicated microprocessor for filtering the signals, computing calibration matrices, and communicating with the robot controller. The latter permit end effectors to be automatically engaged or disengaged by the robot and typically include provisions for routing air or hydraulic power as well as electrical signals. They may also contain provisions for overload sensing.

End Effector Design Issues

Good end effector design is in many ways the same as good design of any mechanical device. Foremost, it requires:

- A formal understanding of the functional specifications and relevant constraints. In the authors, experience, most design “failures” occurred not through faulty engineering, but through incompletely articulated requirements and constraints. In other words, the end effector solved the wrong problem.
- A “concurrent engineering” approach in which such issues as ease of maintenance, as well as related problems in fixturing, robot programming, etc., are addressed in parallel with end effector design.

- An attention to details in which issues such as power requirements, impact resistance, and sensor signal routing are not left as an afterthought.

Some of the main considerations are briefly discussed below.

Sensing

Sensors are vital for some manufacturing applications and useful in many others for detecting error conditions. Virtually every end effector design can benefit from the addition of limit switches, proximity sensors, and force overload switches for detecting improperly grasped parts, dropped parts, excessive assembly forces, etc. These binary sensors are inexpensive and easy to connect to most industrial controllers. The next level of sophistication includes analog sensors such as strain gages and thermocouples. For these sensors, a dedicated microprocessor as well as analog instrumentation is typically required to interpret the signals and communicate with the robot controller. The most complex class of sensors includes cameras and tactile arrays. A number of commercial solutions for visual and tactile imaging are available, and may include dedicated microprocessors and software. Although vision systems are usually thought of as separate from end effector design, it is sometimes desirable to build a camera into the end effector; this approach can reduce cycle times because the robot does not have to deposit parts under a separate station for inspecting them.

Actuation

The actuation of industrial end effectors is most commonly pneumatic, due to the availability of compressed air in most applications and the high power-to-weight ratio that can be obtained. The grasp force is controlled by regulating air pressure. The chief drawbacks of pneumatic actuation are the difficulties in achieving precise position control for active hands (due primarily to the compressibility of air) and the need to run air lines down what is otherwise an all-electric robot arm. Electric motors are also common. In these, the grasp force is regulated via the motor current. A variety of drive mechanisms can be employed between the motor or cylinder and the gripper jaws, including worm gears, rack and pinion, toggle linkages, and cams to achieve either uniform grasping forces or a self-locking effect. For a comparison of different actuation technologies, with emphasis on servo-controlled applications, see Hollerbach et al. (1992).

Versatility

Figure 14.4.7 shows a how/why diagram for a hypothetical design problem in which the designer has been asked to redesign an end effector so that it can grasp a wide range of part shapes or types. Designing a versatile end effector or hand might be the most obvious solution, but it is rarely the most economical. A good starting point in such an exercise is to examine the end effector taxonomy in conjunction with the guidelines in Tables 14.4.1 and 14.4.2 to identify promising classes of solutions for the desired range of parts and tasks. The next step is to consider how best to provide the desired range of solutions. Some combination of the following approaches is likely to be effective.

Interchangeable End Effectors. These are perhaps the most common solution for grasping a wider array of part sizes and shapes. The usual approach is to provide a magazine of end effectors and a quick-change wrist so the robot can easily mount and dismount them as required. A similar strategy, and a simpler one if sensory information is to be routed from the end effector down the robot arm, is to provide changeable fingertips for a single end effector.

Compound End Effectors. These are a “Swiss army knife” approach that consists of putting a combination of end effectors on a single arm, or a combination of fingertips on a single end effector. As long as the end effectors or fingertips do not interfere with each other and the ensemble does not weigh too much for the robot arm, this solution combines the advantage of not having to pause to change end effectors with the advantages of custom-designed tooling. Figure 14.4.8 shows a compound end effector with tools for feeding, measuring, cutting, and laying down wires in a cable harness.

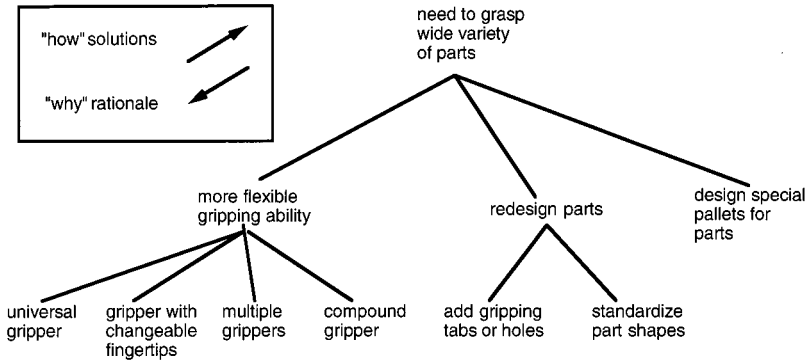


FIGURE 14.4.7 A "how/why" diagram of solutions and rationale for a design problem involving a need to grasp a wide range of parts.

TABLE 14.4.2 Part Characteristics and Associated End Effector Solutions

Size, weight	
Large, heavy	Grippers using wrap grips, taking advantage of friction or *vacuum* or electromagnetic holding
Small, light	Two-fingered gripper; vacuum cup if smooth surface, electromagnet if ferrous alloy
Shape	
Prismatic	Two-fingered parallel-jaw gripper; angular motion if all parts have approximately same dimensions
Cylindrical	Parallel or angular motion two-finger gripper with V-jaw fingertips if light; wrap gripper if heavy; consider gripping on end with three-finger gripper if task or fixtures permit
Flat	Parallel or angular motion gripper or vacuum attachment
Irregular	Wrap grasp using linkages or bladder; consider augmenting grasp with vacuum or electromagnetic holding for heavy parts
Surface	
Smooth	Good for vacuum attachments, simple electromagnets, two-fingered grippers with flat fingertips
Rough	Compliant material (e.g., low durometer rubber) on fingertips or compliant membrane filled with powder or magnetic particles; grippers that use a wrap grasp are less sensitive to variations in surface quality
Slippery	Consider electromagnet or vacuum to help hold onto slippery material; grippers that use a wrap grasp are less sensitive to variations in friction
Material	
Ferrous	Electromagnet (provided that other concerns do not rule out the presence of strong magnetic fields)
Soft	Consider vacuum or soft gripping materials
Very delicate	Soft wrap grippers and vacuum grippers such as those in Figure 14.4.4 can grip very gently; compliant fingertips with foam rubber, or a membrane covering a powder, can also be used to distribute the contact pressure; if the part is very light and fragile consider lifting it using the Bernoulli effect

Redesigned Parts and Fixtures. Stepping back from the end effector, it is useful to recall that the design of the end effector is coupled with the design of fixtures, parts, and the robot. Perhaps we can design special pallets or adapters for the parts that make them simpler to grasp. Another solution is to standardize the design of the parts, using Group Technology principles to reduce the variability in sizes and geometries. When it is difficult to reduce the range of parts to a few standard families (or when the parts are simply hard to grip), consider adding special nonfunctional features such as tabs or handles so that a simple end effector can work with them.

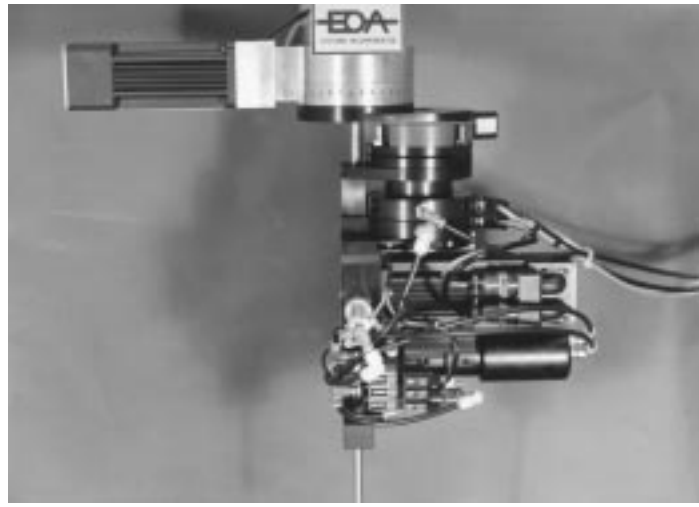


FIGURE 14.4.8 A "compound" end effector with tools for feeding, measuring, cutting, and laying down wires in a cable harness. (Photo courtesy of EOA Systems Inc., Dallas, TX.)

Summary

In summary, we observe that end effector design and selection are inextricably coupled with the design of parts, robots, fixtures, and tooling. While this interdependence complicates end effector design, it also provides opportunities because difficult problems involving geometry, sensing, or task-related forces can be tackled on all of these fronts.

14.5 Sensors and Actuators

Kok-Meng Lee

Sensors and actuators play an important role in robotic manipulation and its applications. They must operate precisely and function reliably as they directly influence the performance of the robot operation. A transducer, a sensor or actuator, like most devices, is described by a number of characteristics and distinctive features. In this section, we describe in detail the different sensing and actuation methods for robotic applications, the operating principle describing the energy conversion, and various significant designs that incorporate these methods. This section is divided into four subsections, namely, tactile and proximity sensors, force sensors, vision, and actuators.

By definition, tactile sensing is the continuously variable sensing of forces and force gradients over an area. This task is usually performed by an $m \times n$ array of industrial sensors called forcels. By considering the outputs from all of the individual forcels, it is possible to construct a tactile image of the targeted object. This ability is a form of sensory feedback which is important in development of robots. These robots will incorporate tactile sensing pads in their end effectors. By using the tactile image of the grasped object, it will be possible to determine such factors as the presence, size, shape, texture, and thermal conductivity of the grasped object. The location and orientation of the object as well as reaction forces and moments could also be detected. Finally, the tactile image could be used to detect the onset of part slipping. Much of the tactile sensor data processing is parallel with that of the vision sensing. Recognition of contacting objects by extracting and classifying features in the tactile image has been a primary goal. Thus, the description of tactile sensor in the following subsection will be focused on transduction methods and their relative advantages and disadvantages.

Proximity sensing, on the other hand, is the detection of approach to a workplace or obstacle prior to touching. Proximity sensing is required for really competent general-purpose robots. Even in a highly structured environment where object location is presumably known, accidental collision may occur, and foreign object could intrude. Avoidance of damaging collision is imperative. However, even if the environment is structured as planned, it is often necessary to slow a working manipulator from a high slew rate to a slow approach just prior to touch. Since workpiece position accuracy always has some tolerance, proximity sensing is still useful.

Many robotic processes require sensors to transduce contact force information for use in loop closure and data gathering functions. Contact sensors, wrist force/torque sensors, and force probes are used in many applications such as grasping, assembly, and part inspection. Unlike tactile sensing which measures pressure over a relatively large area, force sensing measures action applied to a spot. Tactile sensing concerns extracting features of the object being touched, whereas quantitative measurement is of particular interest in force sensing. However, many transduction methods for tactile sensing are appropriate for force sensing.

In the last three decades, computer vision has been extensively studied in many application areas which include character recognition, medical diagnosis, target detection, and remote sensing. The capabilities of commercial vision systems for robotic applications, however, are still limited. One reason for this slow progress is that robotic tasks often require sophisticated vision interpretation, yet demand low cost and high speed, accuracy, reliability, and flexibility. Factors limiting the commercially available computer vision techniques and methods to facilitate vision applications in robotics are highlights of the subsection on vision.

Tactile and Proximity Sensors

A review of past investigations (see Nichols and Lee [1989] for details) has shown that a tactile sensor should have the following characteristics: most important, the sensor surface should be both compliant and durable, and the response of individual forcels should be stable, repeatable, free from hysteresis. The response must be monotonic, though not necessarily linear. The forcels should be capable of detecting

TABLE 14.5.1 Advantages and Disadvantages of Different Tactile Transduction Methods

Type	Advantages	Disadvantages
Resistive and conductive	Wide dynamic range Durability Good overload tolerance Compatibility with integrated circuitry	Hysteresis in some designs Limited spatial resolution Monotonic response, but often not linear
Capacitive	Wide dynamic range Linear response Robust	Susceptible to noise Temperature-sensitive Limiting spatial resolution
Magnetoelastic	Wide dynamic range Low hysteresis Linear response Robust	Susceptibility to stray fields and noise as circuitry requires
Optical	Very high resolution Compatible with vision technology No electrical interference problems	Some hysteresis, depends on elastomer in some designs
Piezoelectric and pyroelectric	Wide dynamic range Durability Good mechanical properties Capable of temperature as well as force sensing	Difficult to separate piezoelectric from pyroelectric effects Inherently dynamic
Thermal	Combined force and temperature	Slow in response

loads ranging from 0 to 1000 g, having a 1-g sensitivity, a dynamic range of 1000:1, and a bandwidth of approximately 100 Hz. Furthermore, forcers should be spaced no more than 2 mm apart and on at least a 10 × 10 grid. A wide range of transduction techniques have been used in the designs of the present generation of tactile sensors. These techniques are compared in Table 14.5.1 and the principles of transduction methods are described as follows.

Resistive and Conductive Transduction

This technique involves measuring the resistance either through or across the thickness of a conductive elastomer. As illustrated in Figure 14.5.1, the measured resistance changes with the amount of force applied to the materials, resulting from the deformation of the elastomer altering the particle density within it. Most commonly used elastomers are made from carbon or silicon-doped rubber, and the construction is such that the sensor is made up of a grid of discrete sites at which the resistance is measured.

A number of the conductive and resistive designs have been quite successful. A design using carbon-loaded rubber originated by Purbrick at MIT formed the basis for several later designs. It was constructed

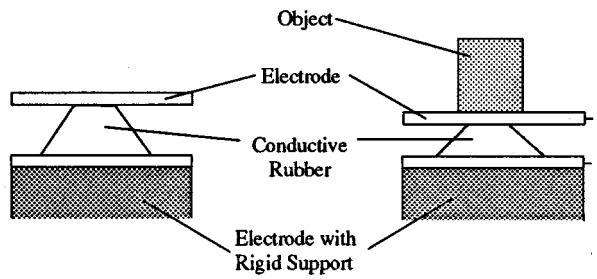


FIGURE 14.5.1 Resistive tactile element.

from a simple grid of silicon rubber conductors. Resistance at the electrodes was measured, which corresponds to loads. A novel variation of this design developed by Raibeit is to place the conductive sheet rubber over a printed circuit board (PCB) which incorporates VLSI circuitry, each forcel not only transduces its data but processes it as well. Each site performs transduction and processing operations at the same time as all the others. The computer is thus a parallel processor.

Capacitive Transduction

Capacitive tactile sensors are concerned with measuring capacitance, which is made to vary under applied load. A common sensor design is to use an elastomeric separator between the plates to provide compliance such that the capacitance will vary according to applied load. The capacitance of a parallel plate capacitor is proportional to its congruous area and the permitivity of dielectric, and inversely proportional to the separation of the plates. Alteration of any of the three parameters causes a change of capacitance. Since the capacitance decreases with decreasing congruous area, the sensor becomes rather cumbersome for design of small forcels.

To allow for a more compact design, an alternative tactile sensor array can be designed based on a moving dielectric element as illustrated in Figure 14.5.2. Each sensing element has two coaxial capacitor cylinders, acting as plates, fixed to a PCB. A dielectric element is spring-mounted in the space between the cylinders. The dielectric is displaced by contact with an external stimulus; hence it moves up and down between the capacitor plates as contact loads vary. A force-displacement relationship is thereby established.

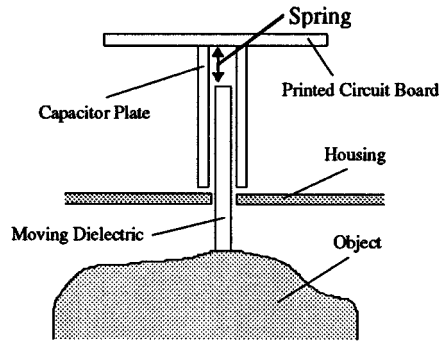


FIGURE 14.5.2 Mechanical/capacitive tactile element.

A novel slip sensor using the change in capacitance caused by relative contact movement between sensor and object is described by Luo (Nichols and Lee, 1989). The contacting sensor surface comprises a set of parallel rollers. Each roller is a half cylinder of conductive material, and a half cylinder of nonconductive material. The rollers are mounted in a nonconductive material.

The casing and rollers act as a variable capacitor. A slipping object will rotate the rollers, causing the capacitance to change, which is then measured, thereby facilitating a slip sensor. The sensor measures the change of phase angle, with the amount of phase shift providing a measure of the scale of slip. A highly linear relationship between detected phase shift angle and sensor output was established.

Magnetoelastic Transduction

Magnetoelastic sensors are a kind of inductive sensor that differs from those described above; they are not based on a change of geometry or on the position of conductive or capacitive materials. Instead, they are based on the Villari effect, consisting of reversible changes in the magnetization curve of a ferromagnetic material when it is subjected to a mechanical stress. It consists of changes of shape and volume during the magnetization process. Magnetoelastic materials undergo changes in their magnetic field when subjected to stress and therefore suggest themselves as possible transducers in tactile sensors.

Figure 14.5.3 illustrates this transduction principle in tactile sensor design. This method of tactile transduction has seen little development in robotics, although there are several papers on the subject (Fraden, 1993).

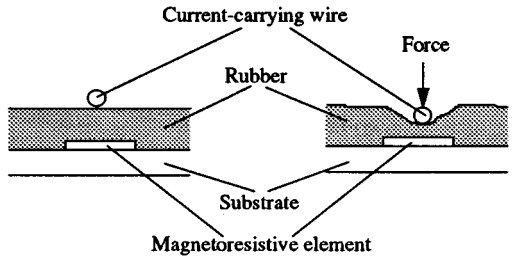


FIGURE 14.5.3 Magnetoresistive tactile element.

Fiber Optics Proximity and Tactile Sensors

The development of optical fiber technology and solid-state cameras has led to some interesting new tactile sensor designs. The capability for high-spatial-resolution images, freedom from electrical interference, and ease of separation of sensor from processing electronics are some of the attractions of incorporating optical transduction methods into tactile sensors. The following illustrates two different fiber optic sensor designs, a proximity sensor and a tactile sensor.

Figure 14.5.4 illustrates the basic principle of fiber optic proximity sensor. Light from a light-emitting diode (LED) is passed down a fiber optic cable to illuminate any proximal objects. A second cable picks up any reflected light from illuminated objects within a detection zone and directs it onto a photodiode. This simple technique can be built into a finger. The finger can sense contacts perpendicular to the finger axis, radially, and also axial contact at the fingertip. Several fiber optic cable pairs can be evenly spaced around the fingers and incorporated into a gripping system. Figure 14.5.4 illustrates this transduction method for proximity sensing.

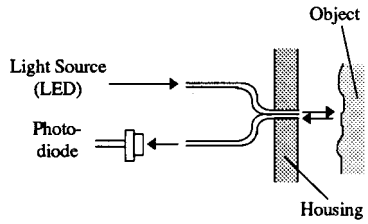


FIGURE 14.5.4 Optical proximity sensing.

Optical fibers are a type of dielectric waveguide. These waveguides channel light energy by "trapping" it between cylindrical layers of dielectric materials. In the most simple case, the fiber core is surrounded by a cladding which has a small refractive index. Light is lost from the core of a fiber when a mechanical bend or perturbation results in coupling between guided and radiation modes. The concept of monitoring light losses due to microbending can be found in several tactile sensor designs (Nichols and Lee, 1989; Tzou and Fukuda, 1992).

Piezoelectric/Pyroelectric Effect

The piezoelectric effect is the generation of a voltage across the sensing element when pressure is applied to it. Correspondingly, the pyroelectric effect is the generation of a voltage when the sensing element is heated or cooled. No external voltage is required, and a continuous analog output is available from such a sensor. Such sensors are most suited for sensing pressure changes or thermal variations. Figure 14.5.5 shows a design based on the piezoelectric effect for robotic applications (Fraden, 1993). The

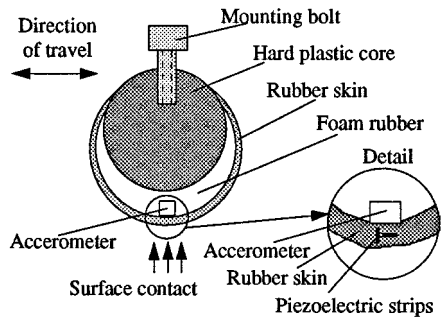


FIGURE 14.5.5 Schematic of piezoelectric sensor for a soft fingertip.

sensor includes piezoelectric strips directly interfaced with a rubber skin; thus the electric signal produced by the strips reflects movements of the elastic rubber which results from the friction forces.

Thermal Tactile Sensors

The thermal sensor (Nichols and Lee, 1989) is based on the detection of the change of thermal properties through contact of an object. The main function of the thermal sensor is to provide information about the material makeup of objects. The essential parts of each element of the thermal sensor are a heat source (such as a power transistor), a layer of material of known thermal conductivity (for example, copper) to couple the heat source to the touched object, and a temperature transducer (thermistor) to measure the contact-point temperature. The response time of the thermal sensor is relatively slow, typically in the order of several seconds. However, images representing the material constitution of the touching objects provide useful tactile data.

Force Sensors

Force sensors measure the force and represent its value in terms of an electrical signal. Examples of these sensors are strain gauges and load cells.

Strain Gauge-Based Force Sensor

A strain gauge is a resistive elastic sensor whose resistance is a function of applied strain or unit deformation. The relationship between the normalized incremental resistance and the strain is generally known as the piezoresistive effect. For metallic wire, the piezoresistance ranges from 2 to 6. For semiconductor gauges, it is between 40 and 200. Many metals can be used to fabricate strain gauges. Typical resistances vary from 100 to several thousand ohms. Strain gauges may be arranged in many ways to measure strains and are used typically with Wheatstone bridge circuits. As strain gauges are often sensitive to temperature variations, interfacing circuits or gauges must contain temperature-compensating networks.

Strain gauges are commonly used for six-degrees-of-freedom force/torque wrist sensors, force probes, flexural assemblies for force control, and micromotion detection. The Scheinman force-sensing wrist is a Maltese cross design, with one strain guage mounted on each of the 16 faces of the cross-webbings. The gauges are operated in eight voltage-divider pairs to measure distortions, and therefore forces, in six degrees of freedom in the hand coordinate system.

Other Force Sensors

Other methods include the vacuum diode force sensor, quartz force sensor, and piezoelectric force sensor. A piezoelectric sensor converts mechanical stress into an electric signal (Fraden, 1993). It is sensitive to changing stimuli only and insensitive to a constant force. As shown in Figure 14.5.6, the sensor consists of three layers where the PVDF film is laminated between a backing material (for example, silicon rubber) and a plastic film. When the PVDF is stressed, it results in a generation of electric charge

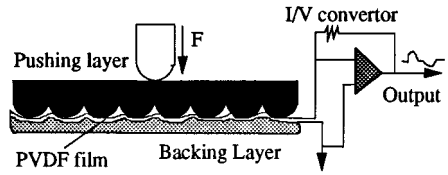


FIGURE 14.5.6 Piezoelectric force rate sensor.

flowing out of the film through a current-to-voltage (I/V) converter. The resulting output voltage is proportional to the applied force.

Figure 14.5.7 shows a typical structure fabricated by micromachining technology in a silicon wafer. As shown in the figure, the diode sensor has a cold field emission cathode, which is a sharp silicon tip, and a movable diaphragm anode. When a positive potential difference is applied between the tip and the anode, an electric field is generated which allows electrons to tunnel from inside the cathode to the vacuum. The field strength at the tip and quantity of electrons emitted (emission current) are controlled by the anode potential. When an external force is applied, the anode deflects and changes the field and the emission current.

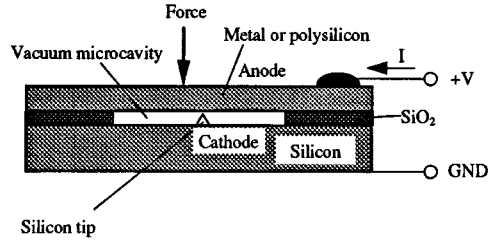


FIGURE 14.5.7 Schematic of a vacuum diode force sensor.

Figure 14.5.8 shows a quartz crystal force sensor. A quartz crystal is often used as a resonator in electrical oscillators. The basic idea behind the quartz force sensor's operation is that certain cuts of quartz crystal shift the resonant frequency when mechanically loaded.

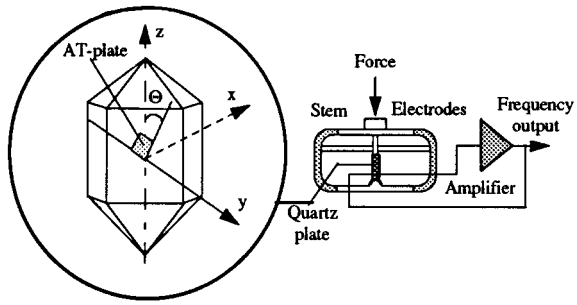


FIGURE 14.5.8 Quartz force sensor.

Vision

Many industrial tasks require sophisticated vision interpretation, yet demand low cost, high speed, accuracy, and flexibility. To be fully effective, machine vision systems must be able to handle complex industrial parts. This includes verifying or recognizing incoming parts and determining the location and orientation of the part within a short cycle time. Typical video-based vision systems conform to the RS-170 standard established in the 1950s, which defines the composite video and synchronizing signal that

the television industry uses. It specifies a standard frame rate for visual interpretation. The components required for building a video-based vision system generally include a video camera which outputs standard RS170 video signal, a frame grabber board which uses a flash analog-to-digital (A/D) converter to change the RS170 video signal into a series of n bit brightness values (gray levels) and fast memory components to store them, and a microcomputer which processes the images and computes the location and orientation of the part. See Ballard and Brown (1982) for information on vision processing techniques.

In addition to the error resulting from the timing mismatching between image acquisition hardware and the computer hardware, the RS170 video signal limits the readout of a complete frame at a rate of 30 fps (frames per second). An image of m rows by n columns has m × n *pixels* and so requires a substantial amount of memory and loading time. Among these m × n pixels, only a few carry the information on which a vision system will base a decision. This generally makes "frame grabbing" inherently wasteful.

Apart from the lack of appropriate hardware and the high equipment cost for robotic applications, a major problem often associated with the use of the RS170 video vision system is the excessive image processing time which depends on the illumination technique, the complexity of the geometry, and the surface reflectance of both the background and the objects to be handled.

Flexible Integrated Vision System

To overcome these problems, several vision systems were designed for robotic applications. Among these is a Flexible Integrated Vision System (FIVS) developed at Georgia Tech (Lee and Blenis, 1994), which offers performance and cost advantages by integrating the imaging sensor, control, illumination, direct digitization, computation, and data communication in a single unit. By eliminating the host computer and frame grabber, the camera is no longer restricted by the RS-170 standard and thus frame rate higher than 30 fps can be achieved.

Flexible Integrated Vision System Hardware. As shown in Figure 14.5.9, the central control unit of the flexible integrated vision system is a microprocessor-based control board. The design is to have all of the real-time processing performed using the microprocessor control board without relying on any other system or computer. Thus, it is desired to have the following features: (1) the microprocessor has an on-chip program memory and independent on-chip data memories. These memories must be externally expandable and accessible with zero wait states; (2) it has independent execution units which are connected by independent buses to the on-chip memory blocks. This feature provides the parallelism needed for high performance digital signal processing and high-powered computation of mathematically intensive algorithms. For these reasons, a digital signal processor (DSP) chip has been chosen.

The DSP-based control board is designed to communicate with several option boards in parallel to tailor the system for a number of applications. Each of these option boards is controlled independently by a programmable logic device (PLD) which receives a peripheral select signal, a read/write signal, and an address signal from the microprocessor control board. Typical examples of the option boards for the FIVS are the digital video head, a real-time video record/display/playback board, and an expandable memory board.

The video head consists of a m × n CCD array, the output of which is conditioned by high bandwidth amplification circuitry. The output is then sampled by a "flash" analog-to-digital converter (ADC). The DSP-based control board provides a direct software control of CCD array scanning and integration time, the intensity of the collocated illumination, and the real-time execution of a user-selectable vision algorithm imbedded in the EEPROM. In operation, the PLD decodes the control signals to initiate row shifts and column shifts in response to commands from the DSP-based control board. Particular row shifts and column shifts enable retrieving only a specific relevant area from an image. The PLD also provides control signals to ADC for performing the analog-to-digital conversion synchronized with row shifts, and enables the video buffer when the DSP reads or writes data to the VRAM.

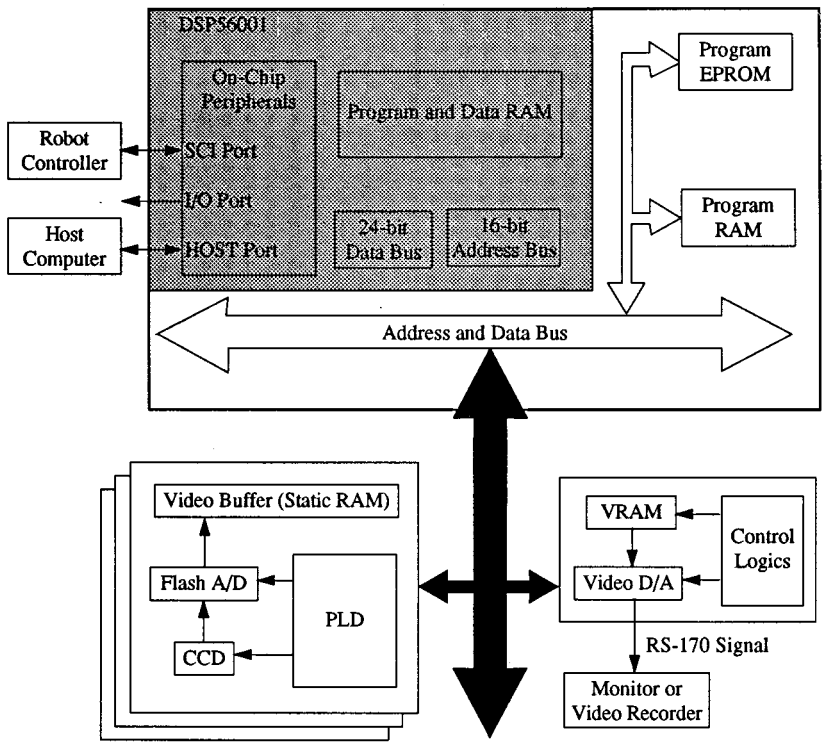


FIGURE 14.5.9 Schematic of a flexible integrated vision system.

Unlike conventional RS170-based systems which require pixel data to be stored in a video buffer before processing of pixel data can commence, the FIVS design provides an option to completely by-pass the video buffer and thus offers a means to process and/or to store the digitized pixel data by directly transferring the ADC output to the DSP. For real-time vision-based object tracking and motion control system applications, the scheme represents a significant saving in time and video buffer size required for processing an image. As an illustration, consider an image array of m × n pixels. The time needed to store the entire image (with no computation) in a memory at K MHz is (m × n)/K s and requires (m × n) bytes of memory. Typical array size of a CCD ranges from 200 × 160 to 4096 × 4096 of pixels. The corresponding video buffer and time required simply to store the entire image at a clock rate of 10 MHz would range from 32K btyes to 16 Mbytes and 3.2 to 1600 msec, respectively! Clearly, the option to completely bypass the video buffer offers a potentially usefu) solution to eliminate the frame storage prerequisite which is often required in conventional vision systems. Furthermore, this scheme completely eliminates the special hardware needed in acquiring the digitized pixel data for storage.

Flexible Integrated Vision System Imbedded Software. The vision system imbedded software includes the following functions. The first function is to give users the flexibility to control the CCD array scanning, integration time, and the intensity of the illumination. With the CCD under software control, partial frames can be “captured” instead of the customary full frame, reducing the cycle time required to capture and process an image. The ability to shift out partial frames is ideal for high-speed tracking applications where the approximate location is known from a prior image. By reducing the time to capture an image, the effective frame rate is increased. For example, shifting out 1/4 of an image can increase the frame rate up to 480 fps, not including the time required for illumination and image processing. This frame rate is 16 times the rate achievable from the RS-170 standard.

The second function is to offer an option to process the pixel data from the ADC directly without having to store the pixel data prior to processing. Although windowing process methods have been

suggested to perform object tracking under software control, these methods required that a partial window is stored before scanning can begin. The differences between the direct computation and the windowing process for object tracking are as follows: (1) in windowing process, the entire image must be stored and analyzed at least once before any subsequent windowing process can be performed in order to provide a reasonable estimate of the object location. Furthermore, if the initial field of view does not contain the object, this estimate must be repeated until an approximate area containing the object can be reasonably found. This prerequisite of storing the image is not necessary if the pixel data are directly processed; (2) after the initial estimate, a fixed window which must be sufficiently large in order to include the object in the field of view must be specified in the windowing process. In most conventional systems which output their ADC to the video buffer directly, a partial frame of the image as specified by the window must be stored. By providing a direct transfer the ADC output to the DSP and thus eliminating the windowing storing process, a significant fraction of time can be saved. This function provides an attractive feature to vision-based motion control applications.

The third function allows image processing to be performed in real time without a host computer. The algorithm that allows the user to customize the system for a specified task is preprogrammed in the EEPROM (electrically erasable programmable read only memory). Because it is impractical to preprogram every possible vision processing algorithm into the FIVS camera, it is desirable that the system can be reprogrammed easily. The main kernel provides a user interface whereby the user can customize the real-time processing for a particular task, from a library of algorithms. This function also provides an effective means to resolve software implementation issues prior to an on-line application. By previewing images of a sample part, the user may select an appropriate vision algorithm for an accurate computation of the location and orientation in real time. Once the algorithms and data are downloaded into the on-board EEPROM, the FIVS can function as an intelligent sensor and communicate directly with the robot controller without a host computer.

The fourth function, which incorporates a real-time display, allows the process controller to set up, to calibrate the vision system, or to analyze a failure mode (if any).

Illumination Considerations

Imaging sensors are characterized by their specific bandwidths or wavelengths of light which maximize the response of the sensor and will provide it an optimum operating environment. It is desired that the photodetector responds only to the light from the illumination source structured for the object but not that of ambient lighting. Otherwise, software compensation must be considered. To accomplish the objective, a typical sensor/illumination system design must consider the spectral matching of the camera imaging sensor/filter and a spectral illuminator while minimizing the effect of the ambient lighting.

Spectral Responsivity of Sensor. The two most commonly used camera imaging sensors are the charge-coupled device (CCD) and the charge injection device (CID). The CCD is responsive to wavelengths of light from below 350 nm (ultraviolet) to 1100 nm (near infrared) and has a peak response approximately at 800 nm. The CID offers a similar spectral response and has a peak spectral response about 650 nm. The relative response of a vidicon camera, however, depends significantly on the materials.

Spectral Characteristic of Typical Ambient Lighting. Depending on the spectral emissions of illumination sources used as general lighting in the factory environment, the influences of the ambient lighting can be effectively minimized or eliminated by means of spectral filtering. Gas discharge lamps generally have relatively high emission in the visible range and have little or no emission for wavelengths larger than 800 nm. Sun, tungsten lamps, and quartz-halogen-type lamps have a wide spectral emission.

Illumination Source. The spectral characteristics of three different spectral sources, namely, laser diodes, light-emitting diode (LED), and xenon strobes, are of particular interest since the spectral wavelengths of these sources well match the optimal response of the CCD and/or CID detectors. Pulsed GaAlAs laser diodes emit single frequency power in the 790- to 850-nm wavelength range. Irradiance at spectral wavelength in the range of 810 to 830 nm can also be produced from a xenon lamp. An

TABLE 14.5.2 Comparison Between Three Spectral Light Sources

Source	Wavelength (nm)	Unit cost (US $)	Life	Power
LED	570–630	1.00	5,000,000 hours (MTBF)	100 mW
Laser diode	790–840	200.00	250,000 hours (MTTF)	1 W (peak pulse power)
Xenon flashtubes	830–1000	10.00	1,000,000 flashes (0.3–4 flashes/sec)	25 W (500 V nominal)

AlGaAs LED is designed to concentrate the luminous flux into a narrow radiation pattern to achieve a narrow high peak intensity. A comparison of these sources is provided in Table 14.5.2.

Object and Background Reflectance. If the orientation of the parts can be characterized by the two-dimensional object silhouette and the environment can be structured, back lighting can be used to create silhouettes of the object. Alternatively, retroreflective materials (Lee and Blenis, 1994) can be used to create a unique background. Since most of the incident illuminance from the object is reflected or diffused away from the aperture, whereas that on the background surface is retroreflected, the object appears as a dark silhouette against a reliable bright-field background.

Retroreflective materials can be used as background in part presentation or as a landmark on parts. The choice clearly depends on the part design and manufacturing process. The most common retroreflective surface is in the form of sheeting due to its reliability and ease of application. Flexible retroreflective sheeting is made of countless microcube-corners or spheres enclosed in a weather-resistant transparent plastic film. Pigment or dye can be inserted into the film or the reflecting surface to reflect color. Four typical retroreflective sheetings are described as follows: (1) cube-corner retroreflective sheeting, (2) exposed glass beads, (3) enclosed glass beads, and 4) encapsulated glass beads. A detailed study of retroreflective sensing for robotic applications is given by Lee and Li (Lee and Blenis, 1994).

Vision Algorithms for Robotic Applications

Figure 14.5.10 illustrates a vision system for robotic part pickup applications (see also Section 14.9). Here the camera is mounted along with the gripper on the end effector mount of the robot. This allows complete freedom in positioning and orienting the camera for viewing. Placing the camera on the last link of a six-DOF robot enables the machine vision to view objects (parts) individually. The camera is oriented so that its line of sight is perpendicular to the plane on which the part is placed. However, at each position, the 3D position and orientation of the feature measured by the vision system are only relative to the vision sensor. The robot is driven by sensory information from the vision system as well as inputs from the off-line calibration.

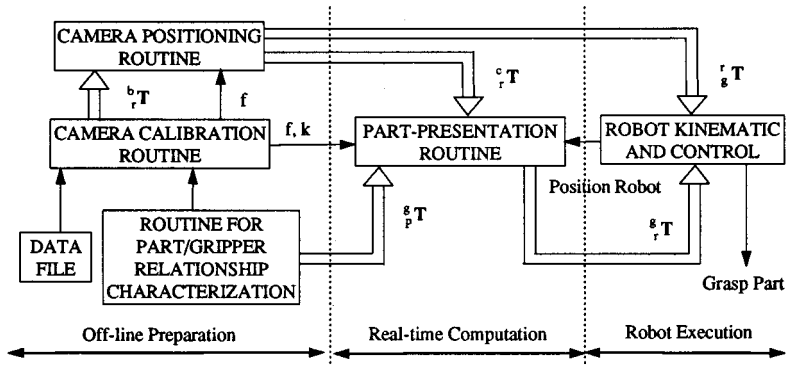


FIGURE 14.5.10 Vision system for robotic applications.

Thus, the basic components of the vision system for robotic part pickup include (1) system calibration, (2) image segmentation and feature extraction, and (3) communication protocol.

System Calibration. In order to determine the 3D position and orientation of the feature with respect to robot world coordinate, it is necessary to calibrate the relative homogeneous transformation between the two coordinate frames, one centered at the camera and the other at the gripper. Thus, the system calibration is to establish the relationship between the 3D world coordinates as seen by the robot and their corresponding 2D image coordinates as seen by the computer. The calibration algorithm consists of off-line calibrating the intrinsic parameters of the camera and the camera-gripper relationship and on-line calibrating the pallet location. The camera calibration technique originally established by Tsai and Lenz (1989) has been the basis for several later calibration routines.

Image Segmentation and Feature Extraction. Most existing industrial-vision systems and algorithms extract features from industrial objects against a high contrast background with controlled lighting. The processing of feature extraction usually begins by generating a binary image from the original gray-scale image by choosing an appropriate threshold. To eliminate noise caused by electromagnetic interference, and ignoring the other objects in the field of view, image segmentation is performed before the computation of the part location and orientation. An image segmentation algorithm is written to locate regions of pixels that are connected and to label each region (object) so that it can easily be picked out from the other regions in the image. After segmentation is complete, only the largest object in the image is examined for its features.

There are many practical methods for the identification of a given object in a scene. A part-recognition system consists of three major components, namely, feature extraction, object modeling, and matching. Most industrial parts-recognition systems are model-based systems in which recognition involves matching the input image with a set of predefined models of part. Models based on geometric properties of an object's visible surfaces or silhouette are commonly used because they describe objects in terms of their constituent shape features. Image features such as edge, corner, line, curve, hole, and boundary curvature define individual feature components of an image. Given a set of models that describes all aspects of all parts to be recognized, the process of model-based recognition consists of matching features extracted from a given input image with those of the models. There are many practical methods for identification of a given object in a scene. A tutorial on binary image processing for robot-vision applications is given by Kitchin and Pugh (1983). A more general comparative study of model-based object-recognition algorithms for robot vision is described by Chin and Dyer (1986).

Communication Protocol. To ensure data integrity during communications, DEC's Digital Data Communications Message Protocol (DDCMP) is used for communications with the vision system. DDCMP is an industrial standard communication protocol that is used to communicate with industrial robots. DDCMP ensures a reliable communications link with a minimum amount of overhead. DDCMP precedes all data transmissions with a header block describing the type of message and the length of any message data that follows.

Actuators

Actuators used for robotic manipulators can be broadly classified as follows: (1) electromechanical actuators, (2) fluid power actuators, and (3) new alternative actuators. Table 14.5.3 summarizes a further subdivision based on their operating principles. The choice of actuators for robotic applications depends on specific tasks. Relative comparisons to guide selection of common actuators are given in Table 14.5.4 and Figure 14.5.11, which shows the force vs. speed comparison for common actuators.

Direct-Drive Joint Motor

The direct-drive joint motor has been developed to eliminate the transmission mechanism between the motor and the links, thus eliminating friction and backlash introduced by gear motors. This results in an arm suitable for high-speed, fine torque control. A direct, drive design also allows for elegant

TABLE 14.5.3 Lower Power Actuator Principles

Electro-mechanical	Fluid power	Alternative concepts
Direct Current (DC) motor	Hydraulic actuators	Piezoelectric
Alternating Current (AC) motor	Pneumatic actuators	Magnetostrictive
Stepper motor		Electrochemical
Electromagnetic		Thermo-bimetal
Linear motor		Shape Memory Alloy
		Electrostatic

TABLE 14.5.4 Comparison between Common Actuators

		Non-linearity			Accuracy
Actuator type	**Static linearity**	**Friction**	**Backlash**	**Hysteresis**	**mm**
AC/DC motor with feed	A	B–C	B–C	B–C	0.005–100
Stepper motor with feed	A	B–C	B–C	B–C	0.01–50
Hydraulic cylinder		C			0.01–100
Pneumatic cylinder		C			0.1–100

Symbols: A good, negligible; B: average, common; C: bad, significant.

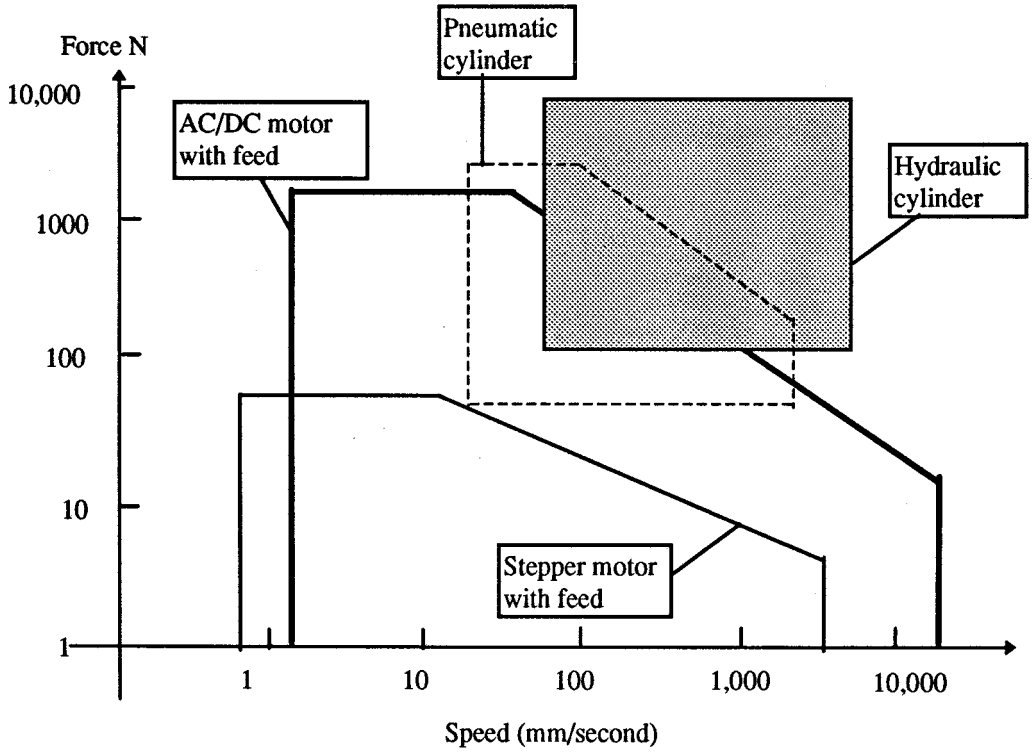


FIGURE 14.5.11 Force vs. speed for common actuators.

mechanical construction. All of the joints in a direct-drive arm are essentially identical in structure, consisting of a motor, a shaft encoder, bearings, and a housing. These components are mounted on a single shaft; a single bearing is used in the entire joint assembly. As a result, a direct drive joint generally has few but compact and more easily manufactured components than a gear-driven joint, an attractive feature for commercial production of arms.

Shape Memory Alloy (SMA) Wire

The properties of the shape memory alloy are associated with appearance and disappearance of martensite in the alloy structure. There are two kinds of martensites, namely, thermal martensite which is generated by cooling the shape memory alloy below martensite transition temperature, and stress-induced martensite which is generated by loading the stress on a shape memory alloy having an austenite structure. Shape memory effect (SME) is associated with the former, and superconductivity (SE) is associated with the latter. By making use of SME and SE, it is possible to use the shape memory alloy as an element of the actuator of a joint mechanism as shown in Figures 14.5.12 and 14.5.13. In Figure 14.5.12, mass 1 is driven toward the right side by heating the SMA wire A and cooling SMA wire B. Similarly, by reversing the direction of the heating and cooling, mass 1 can be made to move toward the left. An alternative SMA wire-actuated revolute joint using an ordinary spring is shown in Figure 14.5.13. The shape memory alloy joint mechanism has the advantage of being light in weight and simple. However, it generally has a relatively low efficiency and is slow in response.

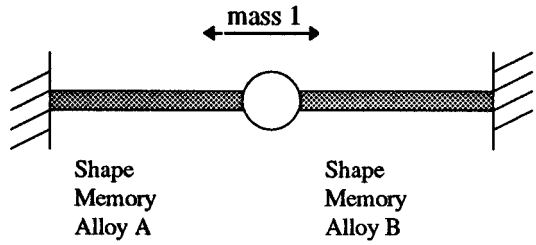


FIGURE 14.5.12 Model illustrating SME wire for joint mechanism.

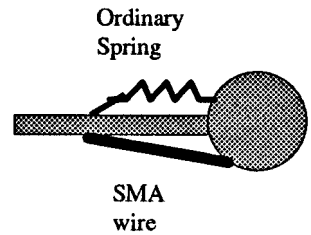


FIGURE 14.5.13 Schematic illustrating SME wire for revolute joint.

An increasing need for high performance robotic applications has motivated several researchers to direct their investigation efforts to new actuator concepts. In some applications such as high-speed plasma, water-jet and laser cutting, active joystick, and coordinate-measuring machines, the demands on workspace and the wrist force/torque are low, but the end effector must be oriented quickly, continuously, and isotropically in all directions. The performance of the popular three-consecutive-rotational-joints wrist, which possesses singularities within its workspace, is less than optimum. Several alternative designs have been developed, which present some attractive possibilities by combining pitch, roll, and yaw motion in single balljoint-like actuators. Among these design concepts are the spherical induction motor, the DC spherical servo motor, and the variable-reluctance (VR) spherical motor. Major developments for robotics are given as follows.

Spherical Induction Motor

In a spherical induction motor, three sets of windings are necessary to realize rotation about an arbitrary axis. The three windings are positioned to give rotations about the x, y, and z axes. By independently controlling the strength and phase of any two windings, one can realize a rotation vector at any point in the rotation plane of the two windings. Analyses of fields and torques in the spherical induction motor have been performed; however, realization of a prototype spherical induction motor remains to be

demonstrated. The mechanical design of a spherical motor is complex. Laminations are required to prevent unwanted eddy currents. Complicated three-phase windings must be mounted in recessed grooves in addition to the rolling supports for the rotor in a static configuration.

Spherical DC Servo Motor

The rotor of a spherical DC servo motor is a disk comprising a yoke with four permanent magnets attached to its periphery. The pivot bearing is constructed of three small radial ball bearings and has three DOF. Therefore, the rotor can incline and rotate around the three axes of Cartesian coordinates relative to the stator. The inclined and rotated angles are detected by rotary encoders attached to each axis. Three sets of windings are set at 30° apart around the z axis such that four electromagnetic force vectors can be obtained at the locations where the currents intersect the magnetic flux. They are controlled like three separated brushless motors. Although the DC spherical motor is characterized by its constructional simplicity, the range of inclination and the torque constant are rather limited.

Variable-Reluctance (VR) Spherical Motor

The structure of a VR spherical motor is shown in Figure 14.5.14, which consists of three subassemblies: a rotor, a stator, and a measuring system. The rotor is a smooth sphere in which m magnetic poles are embedded. The stator is a hollow sphere with n stator coils radially mounted on its inner surface. It also serves as a structure that holds together all the other functional elements which include the stator coils, the bearing, and the measurement system.

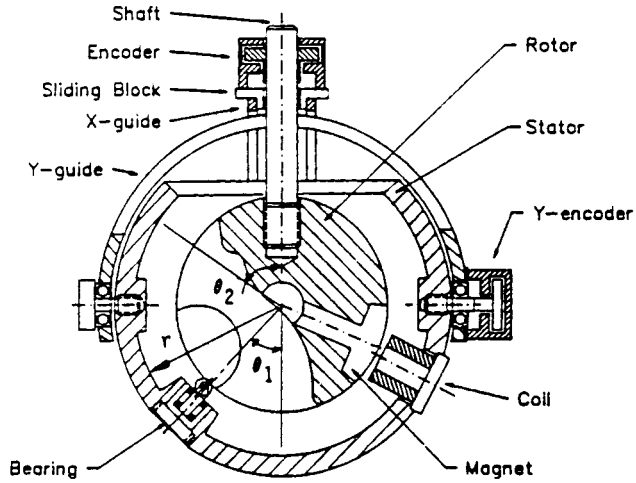


FIGURE 14.5.14 Schematic illustrating VR spherical motor.

In the operation of the VR spherical motor, the stator coils are energized individually using the control circuitry. A magnetic field is established which stores magnetic energy in the airgaps. The stored energy is a function of the relative position of the rotor and the stator. The motion of the spherical motor is thus generated as the rotor tends to move to a position such that the energy in the airgap is minimized. The VR spherical motor is characterized by the following features: (1) it offers a potential advantage of distributing the input power among several coils, each of which contributes a small fraction of the total mmf's required to generate a specified torque, and, thus, it allows a relatively low current per coil but a large surface area for heat dissipation; (2) there are multiple solutions to the selection of coil excitations, which allow an optimal electrical input vector to be chosen to minimize a preselected cost function.

As compared to its counterpart, a VR spherical motor has a relatively large range of inclination, possesses isotropic properties in motion, and is relatively simple and compact in design. The tradeoff, however, is that a sophisticated control scheme is required.

Multi-DOF Microactuators

Silicon exhibits very useful mechanical properties and its applications to microsensors and actuators have been extensively researched around the world. The fabrication of silicon-based components typically employs micromachining. Thin film and photolithographic fabrication procedures make it possible to realize a great variety of extremely small, high precision mechanical structures using the same processes that have been developed for electronic circuits. This technology has enabled the realization of many innovative microactuators operating on the basis of electrostatic to mechanical transduction.

14.6 Robot Programming Languages

Ron Bailey

The earliest industrial robots were simple sequential machines controlled by a combination of servo motors, adjustable mechanical stops, limit switches, and programmable logic controllers. These machines were generally programmed by a record and play-back method with the operator using a teach pendant to move the robot through the desired path. MHI, the first robot programming language, was developed at MIT during the early 1960s. It was an interpreted language which ran on a TX-O computer. WAVE, developed at Stanford during the early 1970s, was a general-purpose language which ran as an assembler on a DEC PDP-10 minicomputer. MINI, developed at MIT during the mid-1970s, was an expandable language based on LISP. It allowed programming in Cartesian coordinates with independent control of multiple joints. AL (Mujtaba, 1982), developed at Stanford, included some programming features of ALGOL and Pascal. VAL and VAL II (Shimano et al., 1984), developed by Unimation, Inc., was an interpreted language designed to support the PUMA series of industrial robots.

AML (A Manufacturing Language) (Taylor et al., 1982) was a completely new programming language developed by IBM to support the R/S 1 assembly robot. It was a subroutine-oriented, interpreted language which ran on the Series/1 minicomputer. Later versions were compiled to run on IBM-compatible personal computers to support the 7535 series of SCARA robots. MCL, developed by McDonnell Douglas, was an extension of the *APT* language (Automatic Programming of Tools) originally written for numerically controlled machine tools. RAIL, developed by AUTOMATIX, Inc., was an extension of Pascal designed to control robot welding and vision systems. Several additional languages (Gruver et al., 1984; Lozano-Perez, 1983) were introduced during the late 1980s to support a wide range of new robot applications which were developed during this period.

V+, developed by Adept Technologies, Inc., is a representative modern robot programming language. It has several hundred program instructions and reserved keywords. V+ will be used in the following sections to demonstrate important features of robot programming.

Robot Control

Program instructions required to control robot motion specify location, trajectory, speed, acceleration, and obstacle avoidance. Examples of V+ robot control commands are as follows:

MOVE	Move the robot to a new location
APPRO	Move to a location set back from a named position
DEPART_	Back away from the current location
DELAY	Stop the motion for a specified period of time
SPEED	Set the speed for subsequent motions
ACCEL	Set the acceleration and deceleration for subsequent motions
SINGLE	Limit motion of a joint
MULTIPLE	Allow full motion of the wrist joints
OPEN_	Open the hand
CLOSE_	Close the hand
RELAX_	Turn off air pressure to the hand

System Control

In addition to controlling robot motion, the system must support program editing and debugging, program and data manipulation, program and data storage, program control, system definitions and control, system status, and control/monitoring of external sensors. Examples of V+ control instructions are as follows:

EDIT	Initiate line-oriented editing
SEE	Initiate screen-oriented editing

STORE_	Store information from memory onto a disk file
LOAD	Read the contents of a disk file into memory
COPY	Copy an existing disk file into a new program
SPEED	Set the overall speed of robot motions
EXECUTE	Initiate execution of a program
ABORT	Stop program execution
DO	Execute a single program instruction
WATCH	Set and clear breakpoints for diagnostic execution
WHERE	Display current robot location
TEACH	Define a series of robot location variables
CALIBRATE	Initiate the robot positioning system
STATUS	Display the status of the system
TIME	Set and display current date and time
ENABLE	Turn on one or more system switches
DISABLE	Turn off one or more system switches

Structures and Logic

Program instructions are needed to organize and control execution of the robot program and interaction with the user. Examples include the following:

FOR	Execute a group of instructions a number of times
WHILE	Continue execution of a group of instructions until a condition is satisfied
DO	Execute a group of instructions until a condition is satisfied
IF	Control whether a group of instructions is executed
TYPE	Output a message to the terminal
ATTACH	Make the control pendant active
WRITE	Output a message to the manual control pendant
PENDANT	Receive input from the manual control pendant
PARAMETER	Set the value of a system parameter

Special Functions

Various special functions are required to facilitate robot programming. These include mathematical expressions and instructions for data conversion, manipulation, and transformation. Examples are as follows:

ABS	Absolute value
COS	Cosine function
SQRT	Square root
BCD	Convert from real to Binary Coded Decimal
DCB	Convert from binary to real
FRAME	Compute the reference frame based on given locations
TRANS	Compose a transformation from individual components
INVERSE	Return the inverse of the specified transformation

Program Execution

Organization of a program into a sequence of executable instructions requires scheduling of tasks, control of subroutines, and error trapping/recovery. Examples include the following:

PCEXECUTE	Initiate the execution of a process control program
PCABORT	Stop execution of a process control program

PCPROCEED	Resume execution of a process control program
PCRETRY	After an error, resume execution at the last step tried
PCEND	Stop execution of the program at the end of the current execution cycle

Example Program

This program demonstrates a simple pick and place operation.

```
 1  .PROGRAM move.parts()
 2  ; Pick up parts at location “pick” and put them down at “place”
 3  parts = 100                          ; Number of parts to be processed
 4  height1 = 25                         ; Approach/depart height at “pick”
 5  height2=50;
    Approach/depart height at “place”
 6  PARAMETER.HAND.TIME = 16             ; Setup for slow hand
 7  OPEN                                 ; Make sure hand is open
 8  MOVE start                           ; Move to safe starting location
 9  For i = 1 TO parts                   ; Process the parts
10  APPRO pick, height1                  ; Go toward the pick-up
11  MOVES pick                           ; Move to the part
12  CLOSEI                               ; Close the hand
13  DEPARTS height1                      ; Back away
14  APPRO place, height2                 ; Go toward the put-down
15  MOVES place                          ; Move to the destination
16  OPENI                                ; Release the part
17  DEPARTS height2                      ; Back away
18  END                                  ; Loop for the next part
19  TYPE “ALL done.”, /I3, parts, “parts processed”
20  STOP                                 ; End of the program
21  .END
```

Off-Line Programming and Simulation

Computer-integrated manufacturing operations require off-line programming and simulation in order to layout production facilities, model and evaluate design concepts, optimize motion of devices, avoid interference and collisions, minimize process cycle times, maximize productivity, and ensure maximum return on investment. Commercially available software (e.g., ROBCAD and SILMA [SILMA Inc., 1992]) provides support for 3D workable layouts including robots, end effectors, fixtures, conveyors, part positioners, and automatic guided vehicles. Dynamic simulation allows off-line creation, animation, and verification of robot motion programs. However, these techniques are limited to verification of overall system layout and preliminary robot program development. With support for data exchange standards (e.g., *IGES* [International Graphics Exchange Specification], *VDAFS* [Virtual Data Acquisition and File Specification], *SET* [Specification for Exchange of Text]), these software tools can pass location and trajectory data to a robot control program which, in turn, must provide the additional functions (operator guidance, logic, error recovery, sensor monitoring/control, system management, etc.) required for full operation.

14.7 Robot Dynamics and Control

Frank L. Lewis

This section deals with the real-time motion control of robot manipulators. In Section 14.8 are covered higher-level planning and control functions, including the generation of the prescribed trajectory that is assumed given in this section. Robot manipulators have complex nonlinear dynamics that might make accurate and robust control difficult. Fortunately, robots are in the class of Lagrangian dynamical systems, so that they have several extremely nice physical properties that make their control straightforward. In this section will be discussed several control techniques including computed torque (e.g., *feedback linearization*), classical joint control, digital control, *adaptive control*, *robust control*, *learning control*, *force control*, and teleoperation. More information may be found in Lewis et al. (1993). The advances that made possible this modern approach to robot control were made by Craig (1985), Slotine and Li (Slotine, 1988), Spong and Ortega (Spong and Vidyasagar, 1989), and others.

Robot Dynamics and Properties

A robot manipulator can have either revolute joints or prismatic joints. The latter are actuators that function like an automobile antenna, extending and contracting on a linear axis. The values of the angles, for revolute joints, and link lengths, for prismatic joints, are called the link variables, and are denoted $q_1(t)$, $q_2(t)$, …, $q_n(t)$ for joints one, two, and so on. The number of links is denoted n; for complete freedom of motion in space, six degrees of freedom are needed, three for positioning and three for orientation. Thus, most commercial robots have six links. We discuss here robots which are rigid, that is which have no flexibility in the links or in the gearing of the joints; flexible robots are discussed in the section on control of flexible-link and flexible-joint robots.

The dynamics of robot manipulators with rigid links can be written as

$$M(q)\ddot{q} + V_m(q,\dot{q})\dot{q} + F(\dot{q}) + G(q) + \tau_d = \tau \tag{14.7.1}$$

where $M(q)$ is the inertia matrix, $V_m(q,\dot{q})$ is the coriolis/centripetal matrix, $F(\dot{q})$ are the friction terms, $G(q)$ is the gravity vector, $\tau_d(t)$ represents disturbances, and $\tau(t)$ is the control input torque. The joint variable $q(t)$ is an n-vector containing the joint angles for revolute joints and lengths for prismatic joints. It is often convenient to write the robot dynamics as

$$M(q)\ddot{q} + N(q,\dot{q})\dot{q} + \tau_d = \tau \tag{14.7.2}$$

$$N(q,\dot{q}) \equiv V_m(q,\dot{q})\dot{q} + F(\dot{q}) + G(q) \tag{14.7.3}$$

where $N(q,\dot{q})$ represents a vector of the nonlinear terms.

The objective of robot control is generally to select the control torques $\tau(t)$ so that the robot follows a desired prescribed motion trajectory or exerts a desired force. Examples include spray painting, grinding, or manufacturing assembly operations. The position control objective can be achieved by first defining a desired trajectory $q_d(t)$, which is a vector containing the desired values vs. time $q_{d_i}(t)$ of each of joint of the manipulator. This desired trajectory vector $q_d(t)$ is determined in a higher-level *path planner*, based on a even higher-level task decomposition, and then fed to the real-time motion control system. This section discusses the real-time motion control problem assuming that $q_d(t)$, or a similar desired force vector, is given.

The robot dynamics in Equation (14.7.1) satisfies some important physical properties as a consequence of the fact that they are a Lagrangian system. These properties significantly simplify the robot control problem. The main properties of which one should be aware are the following.

Properties of Robot Arm Dynamics

P1 The inertia matrix $M(q)$ is symmetric, positive definite, and bounded so that $\mu_1 I \le M(q) \le \mu_2 I$ for all $q(t)$. For revolute joints, the only occurrences of the joint variables q_i are as $\sin(q_i)$, $\cos(q_i)$. For arms with no prismatic joints, the bounds μ_1, μ_2 are constants.

P2 The coriolis/centripetal vector $V_m(q,\dot{q})\dot{q}$ is quadratic in $\dot{q}$ and bounded so that $\|V_m\dot{q}\| \le v_B\|\dot{q}\|^2$.

P3 The coriolis/centripetal matrix can always be selected so that the matrix $\dot{M}(q) - 2V_m(q,\dot{q})$ is *skew symmetric*. This is a statement of the fact that the fictitious forces in the robot system do no work.

P4 The friction terms have the approximate form $F(\dot{q}) = F_v\dot{q} + F_d(\dot{q})$, with F_v a diagonal matrix of constant coefficients representing the viscous friction, and $F_d(\cdot)$ a vector with entries like $K_{d_i}\,\mathrm{sgn}(\dot{q}_i)$, with $\mathrm{sgn}(\cdot)$ the signum function, and K_{d_i} the coefficients of dynamic friction. These friction terms are bounded so that $\|F(\dot{q})\| \le v_B\|\dot{q}\| + k_B$ for constants v_B, k_B.

P5 The gravity vector is bounded so that $\|G(q)\| \le g_B$. For revolute joints, the only occurrences of the joint variables q_i are as $\sin(q_i)$, $\cos(q_i)$. For revolute joint arms the bound g_B is a constant

P6 The disturbances are bounded so that $\|\tau_d(t)\| \le d$.

P7 The nonlinear robot terms are *linear in the parameters* of mass and friction so that one can write

$$M(q)\ddot{q} + V_m(q,\dot{q})\dot{q} + F(\dot{q}) + G(q) = W(q,\dot{q},\ddot{q})\phi \tag{14.7.4}$$

where $W(q,\dot{q},\ddot{q})$ *is* a matrix of known robot functions and ϕ is a vector of mass and friction coefficient parameters, often unknown. The *regression matrix* $W(\cdot)$ can be computed for any specified robot arm.

The last property, P7, is especially useful in adaptive control approaches. The bounding properties are especially useful in robust control approaches. The *skew-symmetry* property P3 is vital for Lyapunov control proofs, which provide guaranteed tracking motion and often give the structure of the control loops. It essentially allows some very nice *linear systems techniques* to be used with the time-varying robot dynamics.

State Variable Representations and Computer Simulation

The nonlinear state-variable representation $\dot{x} = f(x, u)$, with $x(t)$ the internal state and $u(t)$ the control input, is very convenient for many applications, including the derivation of suitable control laws and computer simulation. Once the system has been put into state-space form, it can easily be integrated to obtain simulation time plots using, for instance, a Runge-Kutta integrator; many standard software packages have such integration routines, including MATLAB, MATRIX$_X$, and SIMNON.

It is supposed for convenience in this subsection that the disturbance $\tau_d(t)$ is equal to zero. There are three convenient state-space formulations for the robot dynamics in Equation (14.7.1). In the *position/velocity state-space form*, one defines the state as the $2n$-vector $x \equiv [q^T\ \dot{q}^T]^T$ and writes

$$\dot{x} = \begin{bmatrix} \dot{q} \\ -M^{-1}(q)N(q,\dot{q}) \end{bmatrix} + \begin{bmatrix} 0 \\ M^{-1}(q) \end{bmatrix} \tag{14.7.5}$$

which is in state-space form with $u(t) \equiv \tau(t)$.

For computer simulation purposes, the matrix inversion $M^{-1}(q)$ is required at every integration time step. For arms with simple dynamics, it is often possible to invert the inertia matrix analytically off-line, reducing the on-line computational burden. Otherwise, it is more suitable to solve Equation (14.7.2) for $\ddot{q}$, required by the integration routine, using least-squares techniques to avoid the inversion of $M(q)$.

An alternative *linear* state-space equation in the form $\dot{x} = Ax + Bu$ can be defined as

$$\dot{x} = \begin{bmatrix} 0 & I \\ 0 & 0 \end{bmatrix} x + \begin{bmatrix} 0 \\ I \end{bmatrix} u \tag{14.7.6}$$

with $u(t) \equiv -M^{-1}(q)\, N(q,\dot{q}) + M^{-1}(q)\tau$. This is known as the *Brunovsky Canonical Form.*

The third state-space formulation is the *Hamiltonian form,* which derives from Hamilton's equations of motion. Here the state is defined as the $2n$-vector $x = (q^T\ p^T)^T$, with $p(t) \equiv M(q)\dot{q}$ the *generalized momentum*. Then the state-space equation is

$$\dot{x} = \begin{bmatrix} M^{-1}(q)\, p \\ -\frac{1}{2}\left(I_n \otimes p^T\right)\frac{\partial M^{-1}(q)}{\partial q}\, p \end{bmatrix} + \begin{bmatrix} 0 \\ I_n \end{bmatrix} u \tag{14.7.7}$$

with the control input defined by $u = \tau - G(q)$ and $\otimes$ the Kronecker product (Lewis et al., 1993).

Cartesian Dynamics and Actuator Dynamics

Cartesian Dynamics

The dynamics in Equation (14.7.1) are known as the *joint-space dynamics*; they are expressed in the joint-space coordinates q. Cartesian coordinates referred to some frame, often the base of the robot manipulator, may be used to describe the position of the end effector of the robot arm. Denote the Cartesian coordinates of the end of the arm as $Y(t) = h(q)$, whose first three coordinates represent position and last coordinates represent orientation. The nonlinear function $h(q)$ gives the end effector Cartesian coordinates in termexcepts of the current joint positions q and is called the arm *kinematics transformation.* The *arm Jacobian* relates joint and Cartesian velocites and is defined as $J(q) \equiv \partial h(q)/\partial q$ so that

$$\begin{bmatrix} v \\ \omega \end{bmatrix} \equiv \dot{Y} = J(q)\dot{q} \tag{14.7.8}$$

where $v(t)$ is the linear velocity and $\omega(t)$ the angular velocity of the end effector. Both these velocities are 3-vectors. Differentiating this equation gives the *acceleration transformation* $\ddot{Y} = J\ddot{q} + \dot{J}\dot{q}$.

By differentiating Equation (14.7.1) one discovers that the dynamics may be written in Cartesian form as

$$\overline{M}\ddot{Y} + \overline{N} + f_d = F \tag{14.7.9}$$

where $\overline{M} \equiv J^{-T}MJ^{-1}$, $\overline{N} \equiv J^{-T}(N - MJ^{-1}\,\dot{J}\,J^{-1}\,\dot{y}\,)$, and the disturbance is $f_d \equiv J^{-T}\tau_d$. In the Cartesian dynamics the control input is F, which has three components of force and three of torque.

The important conclusion of this discussion is that the Cartesian dynamics are of the same form as Equation (14.7.2). Furthermore, it can be shown that the properties of the robot dynamics hold also in Cartesian form. Therefore, all the control techniques to be described in this section can be used for either the joint-space or the Cartesian dynamics.

Actuator Dynamics

The robot manipulator is driven by actuators which may be electric, hydraulic, pneumatic, and so on. Considering the case of electric motors it is direct to show that if the armature inductance is negligible, the dynamics of the arm plus actuators can be written as

$$\left(J_M + R^2M\right)\ddot{q} + \left(B_M + R^2V_m\right)\dot{q} + \left(RF_M + R^2F\right) + R^2G = RK_M v \tag{14.7.10}$$

where the robot arm dynamics are described by $M(q)$, $V_m(q,\dot{q})$, $F(\dot{q})$, $G(q)$, and J_M is the motor inertia, B_M is given by the rotor damping constant and back emf, and R has diagonal elements containing the

gear ratios of the motor/joint couplings. The control input is the motor voltage $v(t)$, with K_M the diagonal matrix of motor torque constants.

The important conclusion is that the dynamics of the arm-plus-actuators has the same form as the dynamics (Equation (14.7.1) and can be shown to enjoy the same properties of boundedness and linearityin-the-parameters. Therefore, the control methods to be described herein apply to this composite system as well. Similar comments hold for other sorts of actuators such as hydraulic. If the armature inductances of the electric motors are not negligible, then the arm-plus-actuators have a coupled form of dynamics such as those discussed in the section on control of flexible-link and flexible-joint robots. Then special control techniques must be used.

Computed-Torque (CT) Control and Feedback Linearization

For many years during the 1960s and 1970s the major techniques for robot dynamics control were based on the *computed-torque method*, which has many variants, including classical independent joint control. Recently, advanced mathematical techniques based on *feedback linearization* have been derived. For the rigid-link arms, these are equivalent.

It is assumed that the desired motion trajectory for the manipulator $q_d(t)$, as determined, for instance, by a path planner, is prescribed. Define the *tracking error* as

$$e(t) = q_d(t) - q(t) \tag{14.7.11}$$

and differentiate twice to see that the Brunovsky canonical form in Equation (14.7.6) can be written in terms of the state $x = [e^T \ \dot{e}^T]^T$ as

$$\frac{d}{dt}\begin{bmatrix} e \\ \dot{e} \end{bmatrix} = \begin{bmatrix} 0 & I \\ 0 & 0 \end{bmatrix}\begin{bmatrix} e \\ \dot{e} \end{bmatrix} + \begin{bmatrix} 0 \\ I \end{bmatrix} u \tag{14.7.12}$$

with

$$u \equiv \ddot{q}_d + M^{-1}(q)\left(N(q,\dot{q}) - \tau\right) \tag{14.7.13}$$

A two-step design procedure now suggests itself. First, use linear system design techniques to select a feedback control $u(t)$ that stabilizes the tracking error system in Equation (14.7.12), then compute the required arm torques using the inverse of Equation (14.7.13), namely,

$$\tau = M(q)\left(\ddot{q}_d - u\right) + N(q,\dot{q}) \tag{14.7.14}$$

This is a *nonlinear feedback control law* that guarantees tracking of the desired trajectory. It relies on computing the torque τ that makes the nonlinear dynamics of Equation (14.7.1), equivalent to the linear dynamics of Equation (14.7.12), which is termed *feedback linearization.*

Selecting proportional-plus-derivative (PD) feedback for $u(t)$ results in the *PD computed-torque controller*

$$\tau = M(q)\left(\ddot{q}_d + K_v\dot{e} + K_p e\right) + N(q,\dot{q}) \tag{14.7.15}$$

and yields the tracking error dynamics $\ddot{e} = -K_v\dot{e} - K_p e$, which is stable as long as the derivative gain matrix K_v and the proportional gain matrix K_p are selected positive definite. It is common to select the gain matrices diagonal, so that stability is ensured as long as all gains are selected positive.

The PD computed-torque controller is shown in Figure 14.7.1, which has a *multiloop structure*, with a nonlinear inner feedback linearization loop and an outer unity-gain tracking loop. Note that there are actually n outer loops, one for each joint. In this figure, $\underline{q} \equiv [q^T \ \dot{q}^T]^T$, $\underline{e} \equiv [e^T \ \dot{e}^T]^T$, $\underline{q}_d \equiv [q_d^T \ \dot{q}_d^T]^T$.

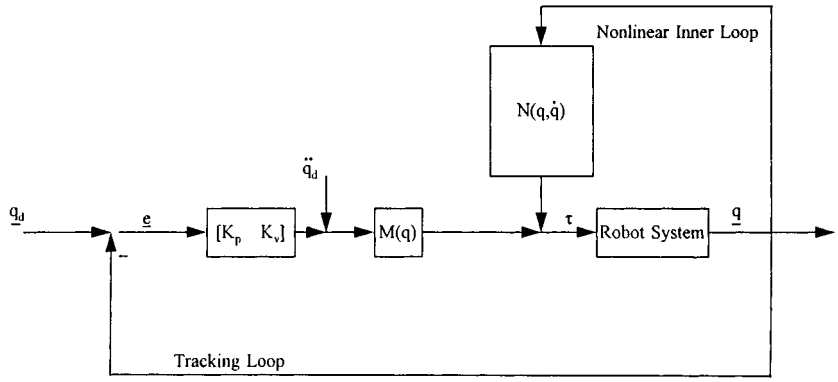


FIGURE 14.7.1 PD computed-torque controller.

To improve steady-state tracking errors, n integrators can be added, one to each joint controller, to place an integrator in the outer tracking loop in the figure. In fact, selecting $u(t)$ *as* a proportional-plus-integral-plus-derivative controller yields the *PID computed-torque controller*

$$\begin{aligned} \dot{\varepsilon} &= e \\ \tau &= M(q)\left(\ddot{q}_d + K_v\dot{e} + K_p e + K_i\varepsilon\right) + N(q,\dot{q}) \end{aligned} \tag{14.7.16}$$

which has its own dynamics and gives stable tracking as long as the integral gain K_i is not chosen too large.

Example 14.7.1 (Performance of PD and PID Computed-Torque Controllers)

The sort of performance to be expected from PD and PID CT controllers is illustrated here. It is desired for a 2-link robot arm to follow, in each of its joints, sinusoidal trajectories $q_d(t)$ with period of 2 sec.

Ideal PD CT Control. Since CT is theoretically an exact cancellation of nonlinearities, under ideal circumstances the PD CT controller yields performance like that shown in Figure 14.7.2, where the initial tracking errors go to zero quickly, so that each joint perfectly tracks its prescribed trajectory. In this figure are shown the plots for joint 1 tracking error $e_1(t)$ and joint 2 tracking error $e_2(t)$.

PD CT Control with Constant Unknown Disturbance. Now a constant unknown disturbance is added to the robot arm. As shown in Figure 14.7.3, the PD controller now exhibits steady-state tracking errors of $e_1 = -0.01$ rad, $e_2 = 0.035$ rad.

PID CT Control. If an integral term is now added to the outer loop to achieve PID CT control, even with a constant unknown disturbance, the simulation results look very much like the original plots in Figure 14.7.2; that is, the integral term has reduced the steady-state tracking errors to zero. □

A class of computed torque-like controllers is given by selecting

$$\tau = \hat{M}\left(\ddot{q}_d - u\right) + \hat{N} \tag{14.7.17}$$

where $\hat{M}, \hat{N}$ are approximations, estimates, or simplified expressions for $M(q)$, $N(q,\dot{q})$. An example is the *PD-gravity controller*

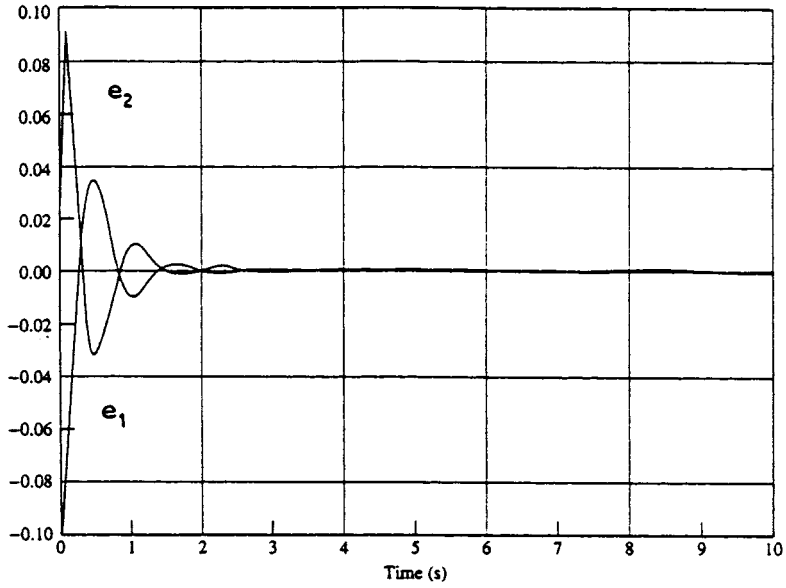


FIGURE 14.7.2 Joint tracking errors using PD computed-torque controller under ideal conditions.

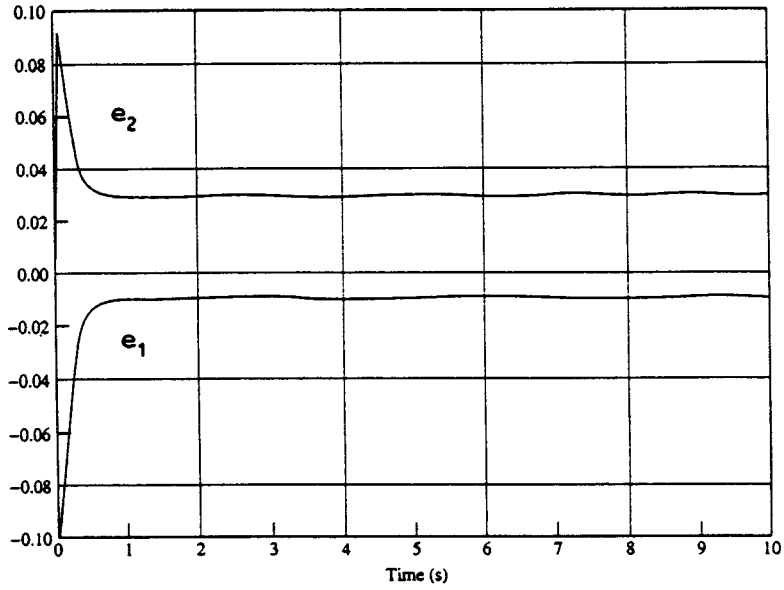


FIGURE 14.7.3 Joint tracking errors using PD computed-torque controller with constant unknown disturbance.

$$\tau = K_v\dot{e} + K_p e + G(q) \tag{14.7.18}$$

which selects $\hat{M} = I$ and only includes the gravity nonlinear terms, so that it is very easy to implement compared to full CT control. This has been used with good results in many applications.

If $\hat{M}, \hat{N}$ are selected, not as the actual inertia matrix and nonlinear terms, but only as approximations or simplified values, it is not always possible to guarantee stable tracking. In fact, the error dynamics of Equation (14.7.12) are then driven by *modeling mismatch errors*, which can degrade or even destabilize the closed-loop system.

Another computed torque-like controller is *PID classical joint control*, where all nonlinearities of the robot arm are neglected and one selects simply

$$\dot{\varepsilon} = e$$
$$\tau = K_v \dot{e} + K_p e + K_i \varepsilon \tag{14.7.19}$$

with the gain matrices diagonal, so that all the joints are decoupled. A PD classical joint controller is shown in Figure 14.7.4, which may seem familiar to many readers. The same figure may be drawn for each joint. In this figure, $d(t)$ represents the neglected nonlinear coupling effects from the other joints, and r *is* the gear ratio. The motor angle is $\theta(t)$ and $q(t)$ *is* the joint angle. The effective joint inertia and damping are J and B, respectively.

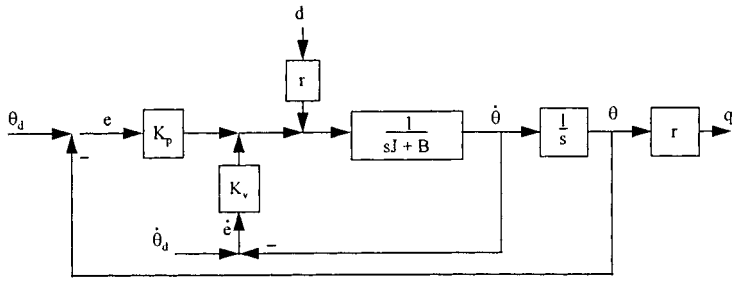


FIGURE 14.7.4 PD classical joint controller.

The simplified classical joint controller is very easy to implement, as no digital computations are needed to determine nonlinear terms. It has been found suitable in many applications if the PD gains are selected high enough, particularly when the gear ratio r is small. Unfortunately, if the gains are selected too high, the control may excite vibratory modes of the links and degrade performance. Moreover, practical applications often benefit by including additional terms such as gravity $G(q)$, desired acceleration feedforward $\ddot{q}_d(t)$, and various additional nonlinear terms.

Example 14.7.2 (Performance of PD-Gravity and Classical Joint Controllers)

The sort of performance to be expected from PD-gravity and classical joint controllers is shown in this example It is desired for a 2-link robot arm to follow, in each of its joints, sinusoidal trajectories $q_d(t)$ with period of 2 sec.

PD-Gravity Controller. The joint 1 and 2 tracking errors are shown in Figure 14.7.5. Note that the errors are small but not exactly zero, a reflection of the fact that the nonlinear coriolis/centripetal terms are missing in the controller. However, the DC error is equal to zero, since gravity compensation is used. (The gravity terms are effectively the "DC terms" of the robot dynamics.)

Classical PD Controller. The sort of behavior to be expected from classical (independent joint) control is illustrated in Figure 14.7.6. In this figure, the tracking errors are nonzero, but using large-enough PD gains can often make them small enough. Note that the DC error is no longer equal to zero; the offset is due to ignoring the gravity terms. □

Another important CT-like controller is the PD *digital controller* given by

$$\tau_k = M(q_k)\left(\ddot{q}_{d_k} + K_v \dot{e}_k + K_p e_k\right) + N(q_k, \dot{q}_k) \tag{14.7.20}$$

where the control input can only be computed at the *sample times*, $t_k = KT$, with T the sample period and k taking on integer values. Digital control is usually required in modern applications, as robot control

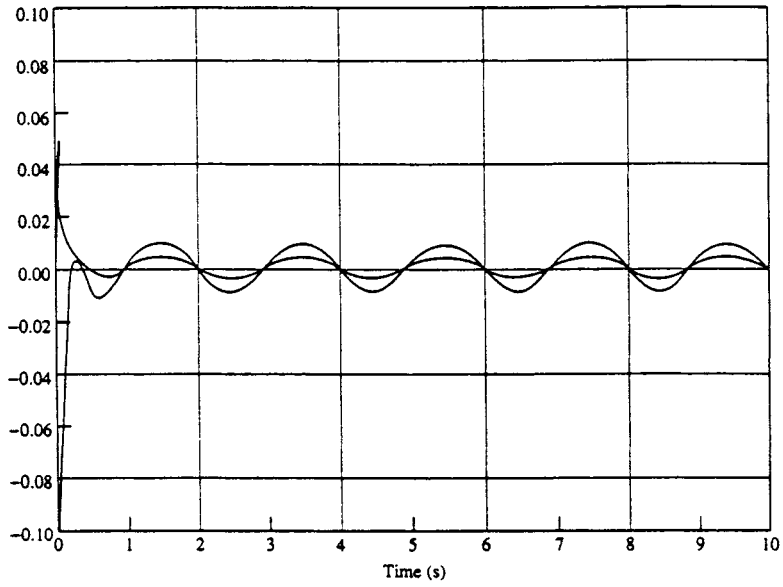


FIGURE 14.7.5 Joint tracking errors using PD-gravity controller.

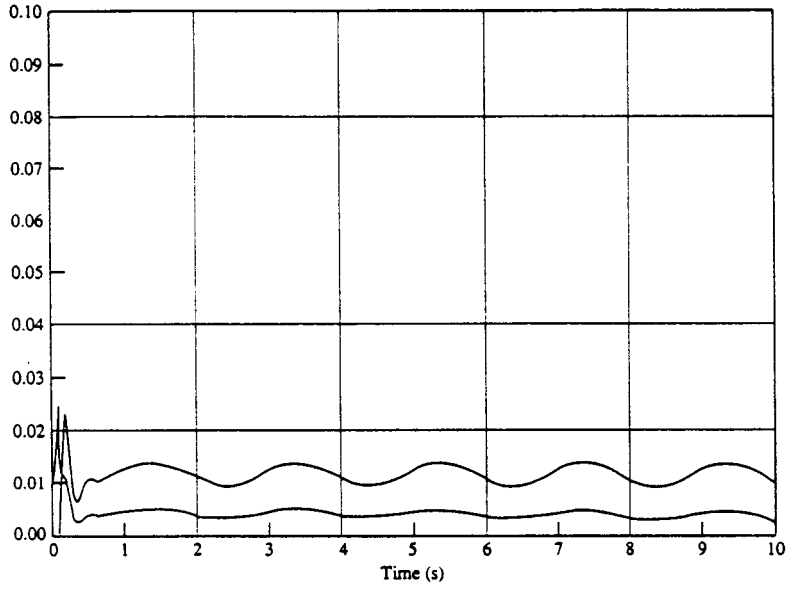


FIGURE 14.7.6 Joint tracking errors using classical independent joint control.

laws are generally implemented using microprocessors or digital signal processors. Unfortunately, the stability of robot digital controllers has not been generally addressed, so that the traditional approach relies on designing continuous-time controllers, meticulously proving stability, then sampling "fast enough" and holding one's breath.

In practice there are many other problems to be faced in robot controller implementation, including actuator saturation, antiwindup compensation, and so on. See Lewis et al., (1993).

Example 14.7.3 (Performance of Digital CT Controllers)

The performance of digital robot controllers has several idiosyncrasies of which one should be aware. In this example, it is desired for a 2-link robot arm to follow, in each of its joints, sinusoidal trajectories $q_d(t)$ with period of 2 sec.

Digital CT Controller. Using a sample period of T = 20 msec yields the tracking error plots shown in Figure 14.7.7. There the performance is quite good for the specific choice of PD gains. The associated control input for joint 2 is shown in Figure 14.7.8.

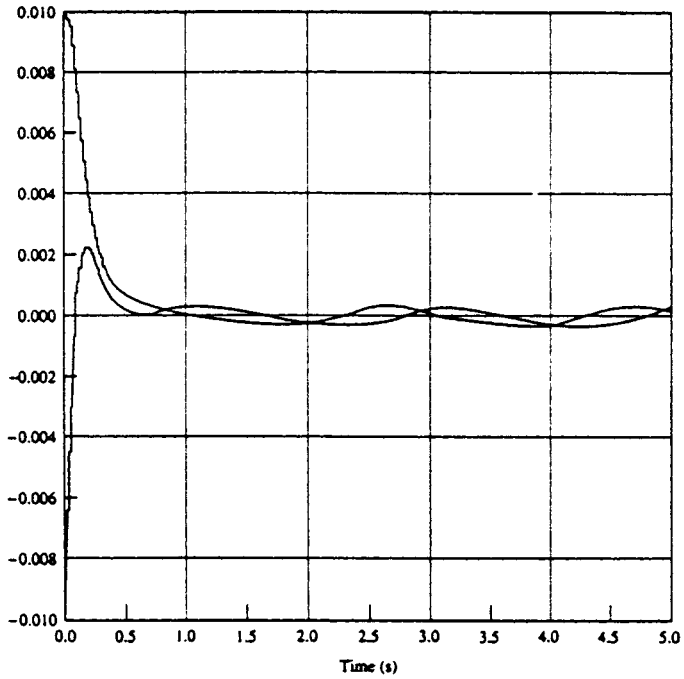


FIGURE 14.7.7 Joint tracking errors using digital computed-torque controller, T = 20 msec.

Limit Cycle of Digital Robot Controller. Unacceptable behavior in digital robot controllers can be due to integrator windup problems, selecting too large a sample period, selecting too small a sample period (so that there is not enough time to perform all control calculations in each period), or the ocurrence of *limit cycles*. If the sample period is selected as T = 100 msec, everything seems acceptable according to Figure 14.7.9, where the tracking errors are somewhat increased but still small. However, Figure 14.7.10 shoves the control torque for link 2, which has now entered a limit cycle-type behavior due to too large a sample period. □

Adaptive and Robust Control

Computed-torque control works very well when all the dynamical terms $M(q)$, $V_m(q,\dot{q})$, $F(\dot{q})$, $G(q)$ are known. In practice, robot manipulator parameters such as friction coefficients are unknown or change with time, and the masses picked up by the arm are often unknown. Moreover, computing nonlinear terms is difficult to do without exotic microprocessor-based hardware. Therefore, in applications simplified CT controllers that do not compute all nonlinear terms are used (e.g., classical joint control). These methods rely on increasing the PD gains to obtain good performance. However, large control signals may result and stability proofs of such controllers are few and far between. Adaptive and robust control techniques are useful in such situations to improve upon the performance of basic PD control

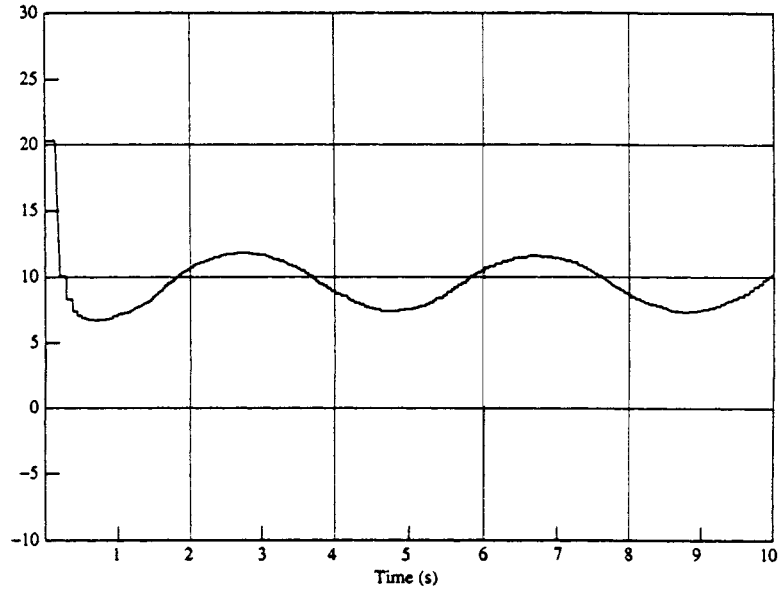


FIGURE 14.7.8 Joint 2 control torque using digital computed-torque controller, T = 20 msec.

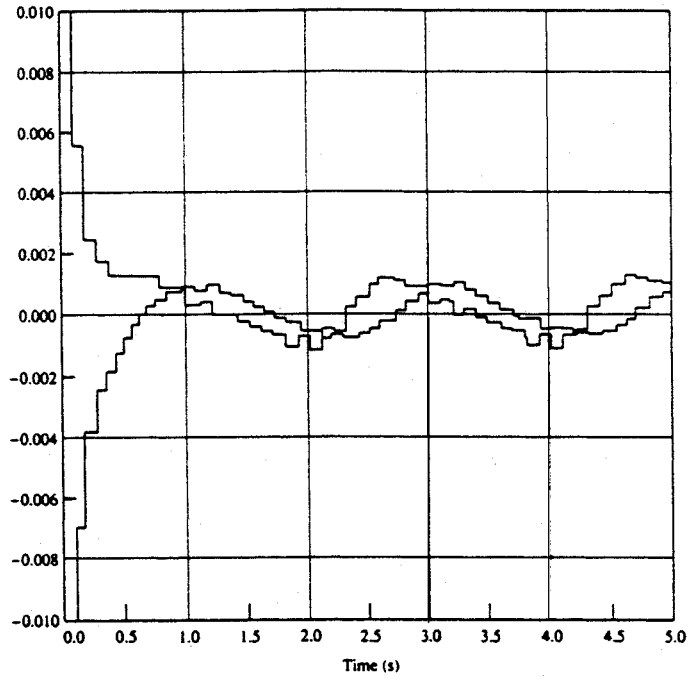


FIGURE 14.7.9 Joint tracking errors using digital computed-torque controller, T = 100 msec.

techniques, providing good performance that can be mathematically proven and so relied upon in applications. Such advanced techniques also extend directly to more complicated control objectives such as force control for grinding, polishing, and so on where straight PD methods are inadequate.

There are many sorts of adaptive and robust controllers (Lewis et al., 1993). A unifying design technique is presented here that extends as well to intelligent control using neural network and fuzzy

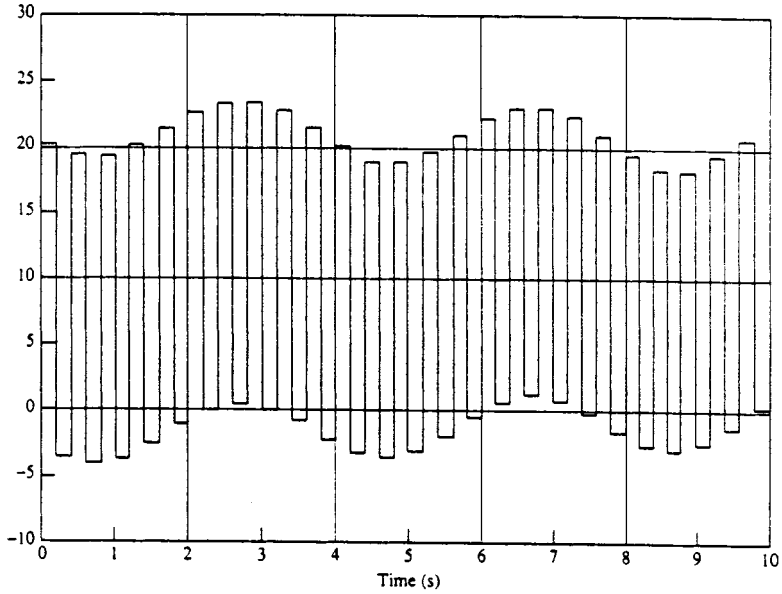


FIGURE 14.7.10 Joint 2 control torque using digital computed-torque controller, T = 100 msec.

logic techniques. Thus, given the desired trajectory $q_d(t)$, define the tracking error and *filtered tracking error* $r(t)$ by

$$e = q_d - q \tag{14.7.21a}$$

$$r = \dot{e} + \Lambda e \tag{14.7.21b}$$

with Λ a positive definite design parameter matrix. Common usage is to select Λ diagonal with large positive entries. Then Equation (14.7.21b) is a stable system so that $e(t)$ *is* bounded as long as the controller guarantees that the filtered error $r(t)$ *is* bounded.

Differentiating Equation (14.7.21b) and invoking Equation (14.7.1), it is seen that the robot dynamics are expressed in terms of the filtered error as

$$M\dot{r} = -V_m r + f(x) + \tau_d - \tau \tag{14.7.22}$$

where the *nonlinear robot function* is defined as

$$f(x) + M(q)\left(\ddot{q}_d + \Lambda\dot{e}\right) + V_m(q,\dot{q})\left(\dot{q}_d + \Lambda e\right) + F(\dot{q}) + G(q) \tag{14.7.23}$$

Vector x contains all the time signals needed to compute $f(\cdot)$ and may be defined, for instance, as $x \equiv [e^T\ \dot{e}^T\ q_d^T\ \dot{q}_d^T\ \ddot{q}_d^T]^T$. It is important to note that $f(x)$ contains all the potentially unknown robot arm parameters, except for the $V_m r$ term in Equation (14.7.22), which cancels out in the proofs.

A general sort of controller is now derived by setting

$$\tau = \hat{f} + K_v r - v(t) \tag{14.7.24}$$

with $\hat{f}$ an *estimate* of $f(x)$, $K_v r = K_v \dot{e} + K_v \Lambda e$ an *outer PD tracking loop*, and $v(t)$ an auxiliary signal to provide robustness in the face of disturbances and modeling errors. The estimate $\hat{f}$ and robustifying

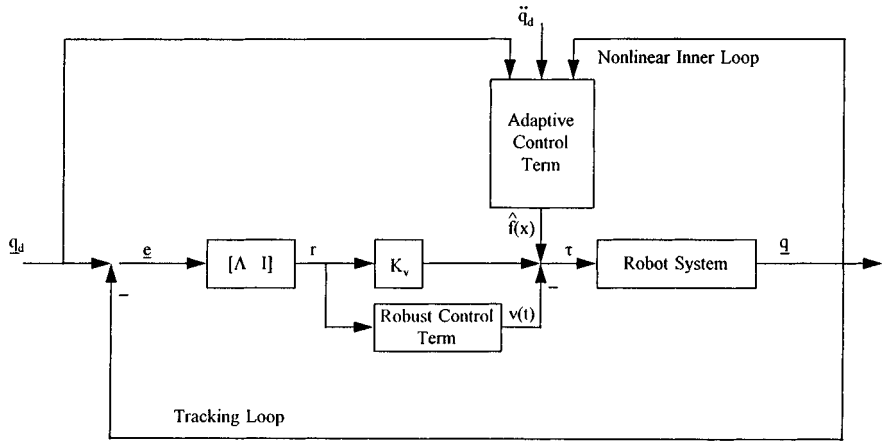


FIGURE 14.7.11 Adaptive filtered error controller.

signal $v(t)$ are defined differently for adaptive control, robust control, neural net control, *fuzzy logic control*, etc. The multiloop control structure implied by this scheme is shown in Figure 14.7.11.

Adaptive Controller

Using nonlinear stability proofs based on Lyapunov or passivity techniques, it can be shown that tracking error stability can be guaranteed by selecting one of a variety of specific controllers. One such is the adaptive controller shown in Figure 14.7.11 and described by the equations

$$\begin{aligned} \tau &= W(w)\hat{\phi} + K_v r \\ \dot{\hat{\phi}} &= \Gamma W^T(w) r \end{aligned} \tag{14.7.25}$$

where Γ is a tuning parameter matrix, generally selected diagonal with positive elements. Matrix $W(x)$ is the (known) regression matrix chosen such that $f(x) = W(x)\phi$, with all the unknown parameters placed into the vector ϕ; $W(x)$ must be computed off-line in a design phase for the specific robot arm to be controlled (Lewis et al., 1993). In the adaptive controller, the second equation represents the internal dynamics of the controller, where the estimate $\hat{\phi}$ of the unknown parameter vector is produced by dynamic on-line tuning. The robot control input $\tau(t)$ *is* then given in terms of $\hat{\phi}$ by the first equation. Though it need not be computed to produce the control inputs, the estimate of the nonlinear function is given by $\hat{f}(x) = W^T(x)\hat{\phi}$.

The adaptive controller shown in Figure 14.7.11 has a multiloop structure with an outer PD tracking loop and an inner nonlinear adaptive loop whose function is to estimate the nonlinear function required for feedback linearization of the robot arm.

Robust Saturation Controller

Another filtered error controller is the robust saturation controller

$$\begin{aligned} \tau &= \hat{f} + K_v r - v \\ v &= \begin{cases} -r\dfrac{F(x)}{\|r\|}, & \|r\| \geq \varepsilon \\ -r\dfrac{F(x)}{\varepsilon}, & \|r\| < \varepsilon \end{cases} \end{aligned} \tag{14.7.26}$$

where $\hat{f}$ is an estimate for $f(x)$ that is not changed on-line — for instance, a PD-gravity-based robust controller would use $\hat{f} = G(q)$, ignoring the other nonlinear terms. In computing the robust control term $v(t)$, ε is a small design parameter, $\|\cdot\|$ denotes the norm, and $F(x)$ is a known function that bounds the uncertainties $\|f - \hat{f}\|$. The intent is that $F(x)$ is a simplified function that can be computed even if the exact value of the complicated nonlinear function $f(x)$ is unknown.

Variable Structure Robust Controller

Another robust controller is the *variable structure robust controller*

$$\tau = \hat{f} + K_v r - v$$
$$v = -\big(F(x) + \eta\big)\mathrm{sgn}(r) \tag{14.7.27}$$

where sgn(·) *is* the signum function and $F(x)$ *is* a known function computed to bound the uncertainties $\|f - \hat{f}\|$. The design parameter η is selected as a small value. This controller takes advantage of the properties of sliding mode or variable structure controllers to provide its robustness.

In adaptive controllers the primary design effort goes into selecting a dynamic estimate $\hat{f}$ that is tuned on-line. By contrast, in robust controllers, the primary design effort goes into selecting the robust term $v(t)$. An advantage of robust controllers is that they have no dynamics, so they are generally simpler to implement. On the other hand, adaptive controllers are somewhat more refined in that the dynamics are learned on-line and less control effort is usually needed. Furthermore, in adaptive control it is necessary to compute the regression matrix $W(x)$, while in robust control it is necessary to compute the bounding function $F(x)$.

Example 14.7.4 (Performance of Adaptive and Robust Robot Controllers)

This example illustrates the sort of performance to be expected from adaptive and robust controllers. In this example, it is desired for a 2-link robot arm to follow, in each of its joints, sinusoidal trajectories $q_d(t)$ with period of 2 sec.

Adaptive Control. In adaptive control, the controller dynamics allow for learning of the unknown parameters, so that the performance improves over time. Typical plots are like those in Figure 14.7.12, where the errors start out large but then converge to zero, and the parameter (mass) estimates converge to constant values.

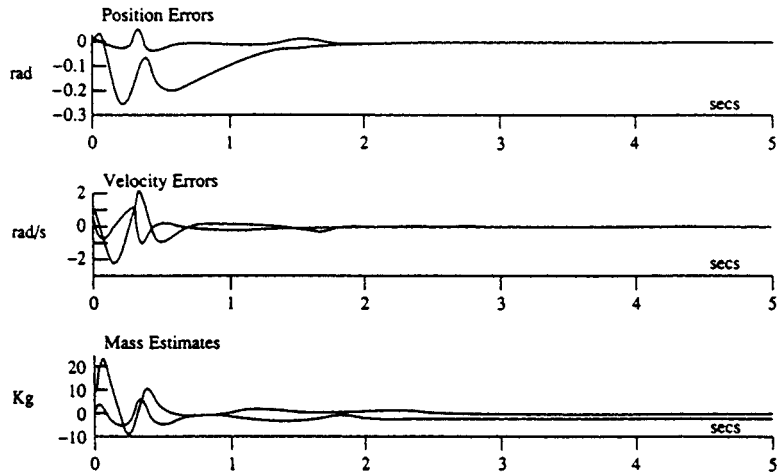


FIGURE 14.7.12 Typical behavior of adaptive controller.

Robust Control. In typical robust controllers, there are no controller dynamics so that the performance does not improve with time. However, with good designs (and large-enough control gains) the errors are bounded so that they are small enough. Typical plots are like those in Figure 14.7.13, where the errors are always small, though nonzero, but do not become smaller with time. □

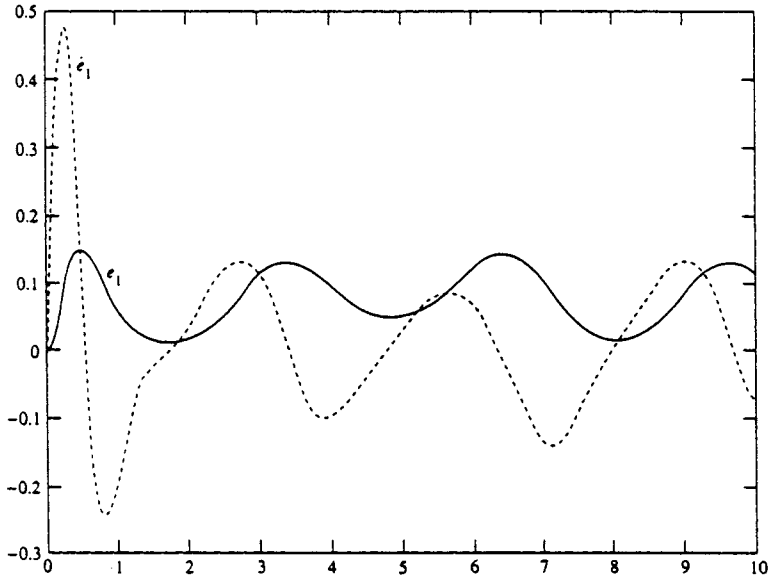


FIGURE 14.7.13 Typical behavior of robust controller.

Learning Control

In many industrial applications robot manipulators are used to perform the same task repeatedly, such as in spray painting, short assembly operations, component insertion, and so on. In such repetitive motion situations, information from one iteration can be recorded and used to improve the performance on the next iteration. This is termed repetitive motion learning control. Using the filtered error approach of the section on adaptive and robust control, it is direct to derive the learning controller of Sadegh et al. for the robot arm in Equation (14.7.1).

Let $\ell = 1, 2, \ldots$ denote the iteration number of the trajectory repetition. Then, using information from the $(\ell - 1)st$ iteration, the controller for the ℓth iteration is given by

$$\begin{aligned} \tau_\ell &= \hat{f}_\ell + K_v r - v \\ v &= -\left(K_p e + K_s \|e\|^2 r\right) \\ \hat{f}_\ell &= \hat{f}_{\ell-1} + K_L r \end{aligned} \tag{14.7.28}$$

where the filtered error is $r = \dot{e} + \Lambda e$, $e = q_d - q$, with $q_d(t)$ the specified trajectory to be followed repeatedly. The gains K_v, K_p, K_s are positive diagonal design matrices, and K_L is a positive diagonal learning gain matrix. The function $\hat{f}_\ell$ is a learning term that uses its value on the previous iteration to improve on an estimate for a nonlinear function appearing in the error analysis.

Control of Flexible-Link and Flexible-Joint Robots

If the robot arm has flexible links, flexible joints, or fast motor dynamics, the control schemes just discussed must be modified. There are two basic cases to consider: *flexible-link robots* and flexible-joint robots; fast motor dynamics can be considered a special case of the latter.

Flexible-Link Robots

In the case of flexible-link robots, the links have significant vibratory modes that cannot be neglected. In this case one may perform an analysis using, for instance, the Bernoulli-Euler model, obtaining an infinite dimensional (partial differential equation) model, which can then be truncated to a finite dimensional (ordinary differential equation) model using, for instance, assumed mode shape techniques. The result is a model such as

$$\begin{aligned} M_{rr}\ddot{q}_r + M_{rf}\ddot{q}_f + V_{rr}\dot{q}_r + V_{rf}\dot{q}_f + F_r(\dot{q}_r) + G_r(q_r) &= B_r\tau \\ M_{fr}\ddot{q}_r + M_{ff}\ddot{q}_f + V_{fr}\dot{q}_r + V_{ff}\dot{q}_f + K_{ff}q_f &= B_f\tau \end{aligned} \tag{14.7.29}$$

which describes the coupling between the rigid modes $q_r(t)$ and the flexible modes $q_f(t)$. In these equations, the quantities M, V, F, G are defined basically as in Equation (14.7.1) and K_{ff} is a matrix of flexible mode stiffness constants.

The complete dynamics are now described by the vector $q = [q_r^T \; q_f^T]^T$. The control objective is to control the link-tip positions to follow a desired trajectory $q_d(t)$ while making small the flexible modes q_f. In the pinned-pinned modes shape method, for instance, the link-tip positions are given by $q_r(t)$. The major problem is that there are now *more degrees of freedom in* $q(t)$ *than control inputs available in* τ. This complicates the control problem greatly; however, a key property is that the matrix B_r is *nonsingular* in flexible-link manipulators.

It can be shown by using a singular perturbation approach, followed by the filtered error approach of the section on adaptive and robust control, that all the basic controllers just described can be used for flexible-link arms if an *additional inner control loop* is added for vibration management. That is, to the control toque $\tau(t)$ generated by Equation (14.7.24), is added the boundary-layer correction (fast) control term, manufactured by an inner loop, given by

$$u_F = -\frac{K_p}{\varepsilon^2}q_f - \frac{K_d}{\varepsilon}\dot{q}_f + \frac{K_p}{\varepsilon^2}\bar{q}_f \tag{14.7.30}$$

where ε is a small parameter (determined according to the time-scale separation imposed by the elements of the stiffness matrix K_{ff}). The slow manifold term $\bar{q}_f$ is a function of the slow control $\bar{u}$ (which is found as before), the variable $q(t)$, and some system parameters. It is possible to avoid measurements of the flexible mode rates $\dot{q}_f$.

Flexible-Joint Robots

The case of flexible-joint robots is in some ways the "dual" problem to that of flexible links. The dynamics of a robot arm driven by motors through rigid joints are given by Equation (14.7.10) for which the controllers described in previous subsections can be used. The dynamics of a robot arm driven by motors through joints with flexibility that is not negligible are given by

$$\begin{aligned} M\ddot{q}_r + V_m\dot{q}_r + F_r(\dot{q}_r) + G_r(q_r) &= K_J(q_f - q_r) \\ J_M\ddot{q}_f + B_M\dot{q}_f + F_M(\dot{q}_f) + K_J(q_f - q_r) &= v \end{aligned} \tag{14.7.31}$$

where $q_r(t)$ is the robot joint variable vector, $q_f(t)$ are the motor angles, and quantities are defined as per the discussions on Equations (14.7.1) and (14.7.10). It is assumed for simplicity that the gear ratio is $R = 1$. The stiffnesses of the joint motor couplings are on the diagonals of the joint stiffness matrix K_J.

The flexible-joint controller problem can be confronted using either a singular perturbation approach (work by M. Spong) or a *backstepping* approach. Using backstepping, it is found that the same basic structure of controller can be used as in Figure 14.7.30, but now the controller has multiple loops, with two adaptive systems required. The extra loop arises since the control input $v(t)$ controls directly the motor angles, which provide indirectly an input into the arm dynamics to control $q_r(t)$, the quantity of actual interest.

Force Control

In many industrial applications it is desired for the robot to exert a prescribed force normal to a given surface while following a prescribed motion trajectory tangential to the surface. This is the case in surface finishing etc. A hybrid position/force controller can be designed by extension of the principles just presented.

The robot dynamics with environmental contact can be described by

$$M(q)\ddot{q} + V_m(q,\dot{q})\dot{q} + F(\dot{q}) + G(q) + \tau_d = \tau + J^T(q)\lambda \tag{14.7.32}$$

where $J(q)$ *is* a Jacobian matrix associated with the contact surface geometry and λ (the so-called "Lagrange multiplier") is a vector of contact forces exerted normal to the surface, described in coordinates relative to the surface.

The prescribed surface can be described by the geometric equation $\phi(y) = 0$, with $y = h(q)$ the Cartesian position of the end of the arm and $h(q)$ the kinematics transformation. The constraint Jacobian matrix $J(q) \equiv \partial\{\phi[h(q)]\}/\partial q$ describes the joint velocities when the arm moves on the surface; in fact, the normal velocity is $J(q)\,\dot{q} = 0$. According to the implicit function theorem, on the surface $\phi(q) = 0$ one may find a function $\gamma(\cdot)$ that $q_2 = \gamma(q_1)$, where the reduced variable $q_1(t)$ corresponds to motion in the plane of the surface and $q_2(t)$ represents dependent variables. The robot, constrained for motion along the surface, satisfies a *reduced-order* dynamics in terms of $q_1(t)$. Defining the extended Jacobian $L(q_1) \equiv [I^T\ \partial\gamma^T/\partial q_1]^T$, the relation of q_1 to the full joint variable q is given via $\dot{q} = L(q_1)\,\dot{q}_1$. For further details see McClamroch and Wang (1988) and Lewis et al. (1993).

The hybrid position/force control problem is to follow a prescribed motion trajectory $q_{1d}(t)$ tangential to the surface while exerting a prescribed contact force $\lambda_d(t)$ normal to the surface.

Define the filtered motion error $r_m = \dot{e}_m + \Lambda e_m$, where $e_m = q_{1d} - q_1$ represents the motion error in the plane of the surface and Λ is a positive diagonal design matrix. Define the force error as $\tilde{\lambda} = \lambda_d - \lambda$, where $\lambda(t)$ is the normal force measured in a coordinate frame attached to the surface. Then a hybrid position/force controller has the structure

$$\tau = \hat{f} + K_v L(q_1) r_m + J^T\left[\lambda_d + K_f\tilde{\lambda}\right] - v \tag{14.7.33}$$

This controller has the basic structure of Figure 14.7.11, but with an inner force control loop. In this controller, the nonlinear function estimate inner loop $\hat{f}$ and the robustifying term $v(t)$ can be selected using any of the techniques mentioned heretofore, including adaptive control, robust control, intelligent control, and so on. A simplified controller that may work in some applications is obtained by setting $\hat{f} = 0$, $v = 0$, and increasing the PD motion gain K_v and force gain K_f.

Teleoperation

In teleoperation, a human operator conducts a task, moving a master robot manipulator and thus defining motion and force commands. The master is connected through a communication channel to a slave robot manipulator in a remote location whose purpose is to mimic the master, thus performing the commanded motions and exerting the commanded forces on its environment. A typical teleoperation system is depicted in Figure 14.7.14. Task performance is enhanced if the human operator has information on the contact force being exerted by the slave manipulator. A convenient way of providing this information is to "reflect" the contact force to the motors on the master so the operator can feel a resistive force indicative of the contact force.

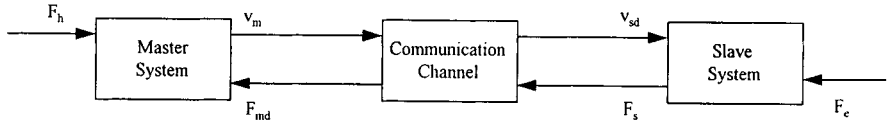


FIGURE 14.7.14 Typical robotic teleoperation system.

To focus on the issues peculiar to teleoperation, one can consider simplified dynamics for the master and slave arms given, respectively, by

$$M_m \dot{v}_m = F_h + \tau_m \tag{14.7.34}$$

$$M_s \dot{v}_s = -F_e + \tau_s \tag{14.7.35}$$

where the human operator input torque is F_h and the contact force exerted by the slave is F_e. In actual systems, one should include the nonlinear coriolis, centripetal, friction, and gravity terms (see Equation (14.7.1)), so that a preliminary feedback linearization (computed torque control) is needed to get the dynamics in this simplified form. Moreover, the Jacobians associated with the force inputs should also be considered (see Equation (14.7.32)).

The control problem is to provide motor control torques τ_m, τ_s so that the slave velocity $v_s = \dot{q}_s$ equals the commanded (master) velocity $v_m = \dot{q}_m$ and the environmental contact force F_e is proportional to the commanded force F_h (there could be a desired force amplification). In Figure 14.7.14, F_s *is* the sensed force resulting from the contact force F_e, the reflected force provided to the master robot is F_{md}, and v_{sd} is the desired velocity command for the slave. A straightforward control scheme for teleoperation is given by

$$\begin{aligned} \tau_m &= -K_m v_m - F_{md} \\ \tau_s &= -K_s v_s + F_s - \alpha_f F_e \end{aligned} \tag{14.7.36}$$

where K_m, K_s are positive master and slave control gains and α_f *is* a positive force gain. The selection of τ_s closes a local force control loop around the slave manipulator.

The key to successful control now lies in the appropriate definition of F_s, F_{md}, and v_{sd}. A naive definition of the sensed force is $F_s = F_e$, the contact force. It has been observed in experiments that this definition is unsuitable and results in instability. Therefore, a *coordinating torque* F_s is defined based on the slave velocity error $e_s(t) \equiv v_{sd}(t) - v_s(t)$ so that

$$\begin{aligned} \dot{\varepsilon} &= e_s = v_{sd} - v_s \\ F_s &= K_p e_s + K_i \varepsilon \end{aligned} \tag{14.7.37}$$

Though it may seem odd to define F_s in terms of the velocity error, it can be shown that, taking into account the impedance relationship of the environment, $F_e = Z_e v_s$, this definition makes the coordinating torque dependent on the contact force F_e. In fact, this definition results in the *passivity* of the slave dynamics referred to the variables (v_s, F_s).

Now, it can be shown that stable teleoperation results if one selects

$$F_{md} = F_s$$
$$v_{sd} = v_m \tag{14.7.38}$$

Unfortunately, if there is any delay in the communications channel, this simple scheme is doomed to failure and results in unstable control. One technique for repairing this problem is to remove the force reflection and let the operator rely on transmitted visual information to infer the contact forces. In practical applications, this can result in the exertion of excessive forces that break tools and fixtures.

It has been shown in Anderson and Spong (1989) that if there is a time delay T in the communications channel, one may modify the controller as shown in Figure 14.7.15 to obtain stable teleoperation regardless of the magnitude of T. In this figure N is a positive scaling factor introduced since the force and velocity signals may differ by orders of magnitude. This modification makes all blocks in the diagram *strictly passive*, so that stability can be shown using circuit analysis techniques. The teleoperation controller with time delay compensation is given by the torques Equation (14.7.36), the coordinating torque in Equation (14.7.37), and the modified reflected force and slave velocity commands given by

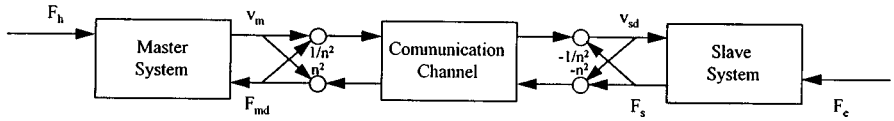


FIGURE 14.7.15 Passive robotic teleoperation system using active control.

$$F_{md}(t) = F_s(t-T) + n^2\left[v_m(t) - v_{sd}(t-T)\right]$$
$$v_{sd}(t) = v_m(t-T) + \frac{1}{n^2}\left[F_{md}(t-T) - F_s(t)\right] \tag{14.7.39}$$

It is noted that in this modified controller, part of the reflected force is derived from the slave velocity error and part of the slave velocity command is derived from a force error term.

14.8 Planning and Intelligent Control

Chen Zhou

The previous section dealt with servo-level joint control of robot manipulators. This section deals with higher-level functions of planning and control, including generation of the prescribed joint trajectories that are required for servo-level control. Robots are designed to accomplish various tasks such as spot welding in assembly, part loading in material handling, or deburring in processing. A task consists of a series of complex motions of various joints and the end effector. The execution of robot tasks is controlled by robot programs. Robot programming can be classified into three levels: (1) joint level, (2) manipulator level, and (3) task level (see Leu, 1985). At the joint and manipulator levels, a task is decomposed into a set of explicit paths. The robot program is essentially a series of move statements to instruct the robot to pass through a sequence of paths. The programming at these two levels involves tedious specification of points, paths, and motions.

The difficulties involved at the first two levels of programming led to research and development for task level programming. At the task level, a robot program is a sequence of goals or objective states of the tasks, such as inserting a peg or deburring an edge. Due to the omission of explicit path and kinematic instructions by the programmer, the robot must know its configurations, its environment and the goal locations, such as the location of materials, as well as the obstacles within the envelope. The robot controller has to construct a set of collision-free paths that are optimized in terms of time, motion, or other control characteristics.

Task planning is a process in which the detailed motion control involved in a task (or a subtask) is determined by software or algorithms. The objectives can include proper grasp configurations, collision-free paths, minimum travel distance, minimum travel time, and avoidance of *singularities*. In a singular configuration, the robot may lose a degree of freedom or lose the ability to provide designed power. For complex tasks, task decomposition can be performed to provide more specific control of the robot. *Path planning* is a process of finding a continuous path from an initial robot configuration to a goal configuration without collision. It is a very important component in task planning.

Task planning includes several special cases. First, errors can occur during execution of a task. Various sensors have been developed to detect error conditions. Since the occurrence of an error is random, there is uncertainty associated with task planning when error and error recovery are concerned. Second, multiple robots are often used in robotic applications. The simplest case involves two arms. An important issue in two-arm task planning and control is the coordination of the two arms. Third, in robotic manufacturing cells, robots must coordinate with other equipment in the cell and the control of the robot can often affect the performance of the entire cell. Therefore, cell control is also discussed in this section. At the end of this section, we also mention artificial intelligence as applied to planning and man–machine interface.

Path Planning

Path planning involves finding a continuous path from an initial robot configuration C_{init} to a goal configuration C_{goal} without collision. Figure 14.8.1 illustrates an example of path planning in two-dimensional space. The small rectangular object represents a *mobile robot* or an end effector of a manipulator, and the other objects represent the obstacles within the working envelope. The dashed line shows the path of the center point of the robot. The four small rectangles show the orientation of the robot at the initial, goal, and two intermediate configurations.

Often, the collision-free path is not unique. In addition to avoiding collision, one can also add requirements for smoother motion, shorter traveling distance, shorter traveling time, or more clearance from the obstacles. Therefore, path planning can also involve optimization with respect to certain performance measures.

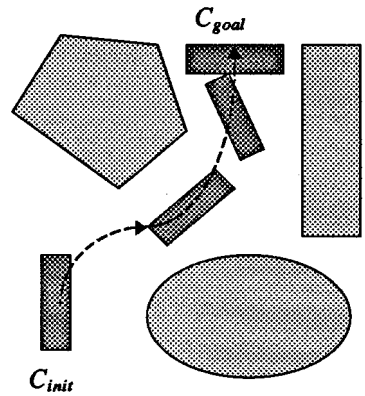


FIGURE 14.8.1 Illustration of path-planning problem.

The abstract model of the path-planning problem can take different forms depending on the application characteristics. It can be in two dimensions or in three dimensions. The concern can be the end effector alone or the entire robot arm. The end effector can be considered as a solid body or an infinitesimal point. Different considerations can have significant implications on the solution methodology and complexity. In this handbook, we will only discuss the simplest cases in which a point end effector in two-dimensional space is concerned. The readers are referred to Latombe (1991) for more complex procedures.

Road Map Approach Based on Visibility Graph

The road map approach is one of the earliest path-planning methods. The obstacles are modeled as polygons. A *visibility graph* is a nondirected graph. The nodes of the graph are the vertices of the polygons, the initial point and the goal point. The links of the graphs are straight-line segments that connect a pair of nodes without intersecting with any obstacles. A reduced visibility graph for the example is shown in Figure 14.8.2. A reduced visibility graph does not contain links that are dominated by other links in terms of distance. The elliptical obstacle is approximated by a hexagon. In the visibility graph, all the paths consisting of successive links that connect C_{init} to C_{goal} represent semicollision-free paths. The coarse line represents one of these paths. The use of the term "semicollision free" is due to the fact the path may actually contact an obstacle. It is clear that the path is not unique. In the example the possible paths can be $C_{init}\,AC_{goal}$, $C_{init}\,BC_{goal}$, or $C_{init}\,CD_{goal}$. Some offer shorter travel distances while others offer smoother paths. This method can be extended to include circular end effector and obstacles which have lines and arcs as boundaries.

Road Map Approach Based on Voronoi Diagram

For the same problem described above, one can create a *Voronoi diagram* based on the vertices and line segments of the obstacles and the working envelope and use this graph to generate a collision-free path. A Voronoi diagram is a diagram that consists of lines having equal distance from the adjacent objects. Obviously, the Voronoi diagram does not touch the obstacles and can provide collision-free paths. A Voronoi diagram in a polygonal space with polygonal obstacles is composed of straight line segments and parabolas. When both adjacent object segments are straight lines or vertices, the segment of the Voronoi diagram is a straight line. When one object segment is a point while the other is a line segment, the segment of Voronoi diagram is a parabola. Two additional links need to be created to connect the C_{init} and C_{goal} to the Voronoi diagram. Any set of links that connects C_{init} and C_{goal} through the diagram represents a collision-free path. Unlike the road map approach based on visibility graph, this approach tends to maximize the clearance between the robot and the obstacles. For the characteristics and creation of the Voronoi diagrams, the reader is referred to Okabe et al. (1992).

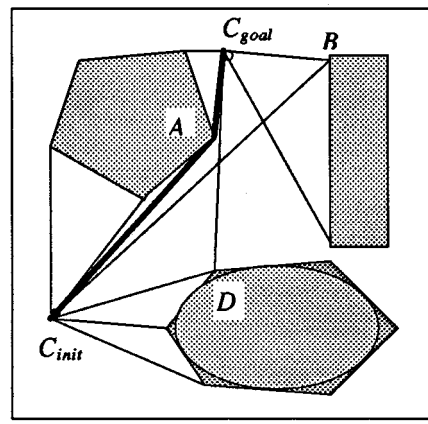


FIGURE 14.8.2 Road map method based on visibility graph.

Cell Decomposition Approach

In the *cell decomposition* approach, the robot free space is decomposed into simple connected geometric shapes, or cells such that a path can be generated between any two points within a cell. When Euclidean distance is used (as when using a Cartesian robot), convex polygonal cells satisfy this requirement. The simplest way to generate cells is the line sweeping method. An example in which the work envelope and the obstacles are polygons is shown in Figure 14.8.3. The two shaded areas are obstacles. In this example, the decomposition is done by sweeping a vertical line across the work envelope. A cell is formed whenever a vertex is encountered by the sweeping line. After decomposition, a *connectivity graph* is constructed. A connectivity graph is a nondirected graph. The nodes of the graph are the cells. If two cells share a common edge, they are connected and a link is drawn between the two nodes. The existence of a collision-free path can be found by searching the connectivity graph to see if there exists a path that connects two nodes containing C_{init} and C_{goal}. If such a path exists, one can construct collision-free paths by determining paths in the cells and connect the paths in adjacent cells.

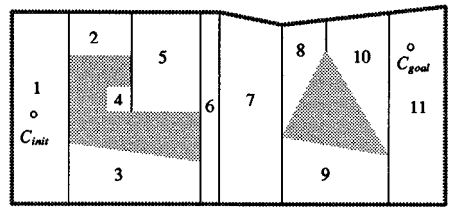
Apparently, the path and decomposition are not unique. One can select different paths and decompositions to optimize other measures. Figure 14.8.3(d) shows another possible decomposition of the same space.

Potential Field Approach

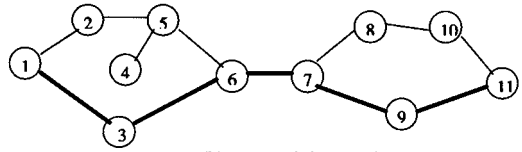
The idea of the *potential field* method is to represent the robot work space as a potential field which has peaks and a valley. The valley is at the goal configuration, while the peaks are at the location of the obstacles. The robot, represented by an article, will roll naturally away from the peaks and toward the valley in such a terrain. Mathematically, this is done by creating an artificial potential field with peaks as obstacles and a valley as the goal, and by using a search procedure to plan a descending path to the valley. The artificial potential field is composed of an attractive potential with its lowest point as the goal configuration, and a repulsive potential for each of the obstacles. An example is shown in Figure 14.8.4. The dimension of the robot envelope is 10 wide and 10 high. Here *ZG* is the attractive potential function, *ZR* is the repulsive potential function for all obstacles, and *Z* is the combined potential function. In this example, the attractive potential function *ZG* is a parabolic well:

$$ZG(x,y) = \frac{A}{2}\left[\left(x - G_x\right)^2 + \left(y - G_y\right)^2\right] \tag{14.8.1}$$

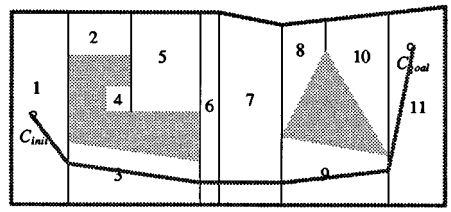
where *A* is a constant used to adjust the magnitude of the attractive field, G_x = 2 and G_y = 1 are the coordinates of C_{goal}. *ZR* represents three cylindrical obstacles with diameters 2, 1, and 1, respectively.



(a) cell decomposition



(b) connectivity graph



(c) a possible collision free path.

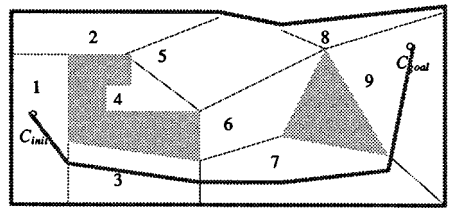


FIGURE 14.8.3 Cell decomposition method.

The center location of the three obstacles are at (2, 5), (7, 7), and (6, 3). Let ZR_i be the obstacle repulsive fields of i^{th} cylindrical obstacle. The repulsive function for i^{th} obstacle is

$$ZR_i(x,y) = \begin{cases} \dfrac{B_i}{2}\left[\dfrac{1}{\sqrt{\left(x-R_{x,i}\right)^2+\left(y-R_{y,i}\right)^2}} - C_i\right] & \text{if } \dfrac{D_i}{2} \ge \sqrt{\left(x-R_{x,i}\right)^2+\left(y-R_{y,i}\right)^2} < D_i \\ Z_i & \text{if } \sqrt{\left(x-R_{x,i}\right)^2+\left(y-R_{y,i}\right)^2} \le \dfrac{D_i}{2} \\ 0 & \text{otherwise} \end{cases} \tag{14.8.2}$$

where B_i is a constant for the adjustment of the height of i^{th} peak, D_i is the diameter of the i^{th} cylindrical obstacle, and $R_{x,i}$ and $R_{y,i}$ are the center coordinates of the cylinder. The characteristic of this function

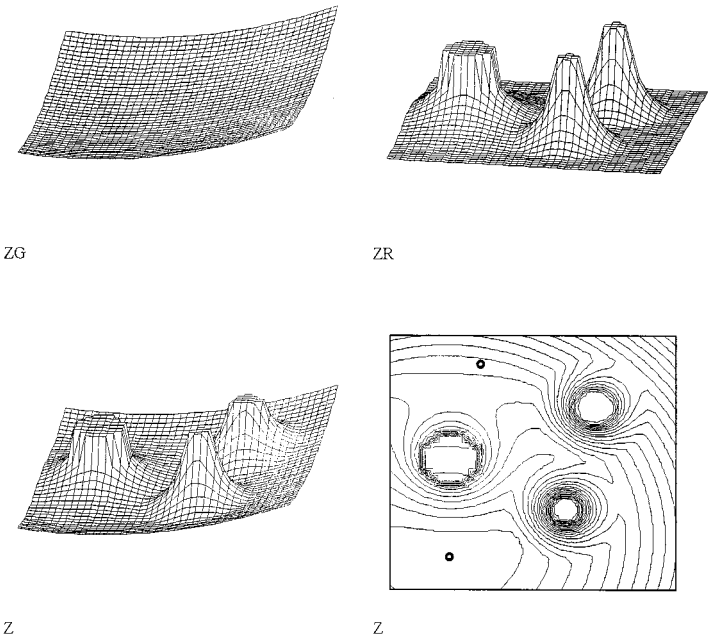


FIGURE 14.8.4 Potential field method example.

is that the potential at any point with distance more than twice the diameter of the closest obstacle is zero. Also, the potential inside the obstacle is a constant and equal to the potential at the boundary of the obstacle, where C_i is a constant to accomplish the former and the Z_i is a constant for the latter. The total potential is the sum of all the terms:

$$Z = ZG + \sum_{i=1}^{3} ZR_i \tag{14.8.3}$$

There are several techniques for potential guided path planning. The simplest is the depth first planning. In depth first planning, a prespecified step δ is predefined. The path will be found iteratively using

$$\begin{aligned} x_{n+1} &= x_n + \delta \frac{\partial Z(x_n, y_n)}{\partial x} \\ y_{n+1} &= y_n + \delta \frac{\partial Z(x_n, y_n)}{\partial y} \end{aligned} \tag{14.8.4}$$

where x_0 and y_0 are at C_{init}. Depth first planning is very fast for certain situations but may cause trapping some local point in others.

Error Detection and Recovery

In the execution of a task, errors can occur. The errors can be classified into several categories: hardware error, software error, and operational error. The hardware errors include errors in mechanical and electrical mechanisms of the robot, such as failure in the drive system or sensing system. Software errors can be bugs in the application program or control software. Timing with cooperative devices can also

be called software error. The operational errors are the errors in the robot environment that are external to the robot system such as jamming of parts or collision with obstacles.

The sensors used in error detection can be classified into several categories: tactile sensors for sensing contact and existence, proximity sensors for sensing location or possible collision, force/torque sensors for sensing collision and jamming, and vision for sensing location, orientation, and existence.

The occurrence of an error normally causes interruption of the normal task execution. Error recovery can be done at three levels. At the lowest level, the task is not resumable. Upon detection of an error, the task is interrupted automatically. The error must be corrected manually and the task must start again manually. At the second level, the task can be resumed. Upon detection of an error, the error can be corrected manually and the task can be continued from the point where the error occurred. At the third level, upon detection of an error, the error will be corrected automatically and the task execution is continued. Figure 14.8.5 shows an example of error and recovery in an insertion task.

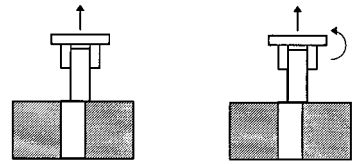


FIGURE 14.8.5 Jamming error detection and recovery.

In this insertion example, a peg is misaligned with the hole. The misalignment can be caused by wrong location of the fixture, wrong positioning of the peg before pickup, etc. The jamming error can be detected by a vertical force sensor in the gripper. At the lowest level, this can be used to trigger an emergency stop and an alarm. An operator can withdraw the gripper, remove the peg, and start the cycle again. At the second level, the operator can correct the problem and continue the insertion after the error condition is corrected. At the third level, additional sensory information is required. For example, additional force sensors can be used to identify the jamming torque. The torque information can be used to redirect the end effector to put the peg in the hole correctly. Since the occurrence of error is random in nature, the incorporation of error recovery introduces uncertainty into task planning. Artificial intelligence methods are often employed in error recovery.

Two-Arm Coordination

Many robotic applications require multiple arms, such as lifting heavy weights in handling or assembling two components that require simultaneous motion. In such applications, a special planning and control issue is the coordination of two arms. Figure 14.8.6 shows an example of two-arm application. In two-arm applications, two arms form a closed chain. Each arm acts as a constraint on the other and can have different contributions to the motion of the part. In terms of constraints, one or both arms can have rigid grasps. Two arms may also be controlled to remain in contact at a point, along a line, or on a surface. In terms of control, two-arm coordination can rely on a master/slave relationship, a push/pull relationship, or other relationships. In the master/slave relationship, one arm is controlled in force mode (master) while the other is controlled in position mode (slave). These constraints and control relationships require different controls from both controllers. Please see Hayati et al. (1989) for more discussion.

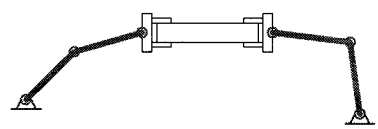


FIGURE 14.8.6 Two-arm coordination.

Workcell Control

Robots are used in many workcells for machine loading/unloading, assembly, part *kitting*, packaging, etc. The task planning associated with workcell control has its unique characteristics. A robot in a workcell often interacts with other resources in the cell such as a machine, a conveyor, a fixture, or a container. These interactions often require exact coordination. The coordination is done based on the clocks or interlocks implemented through *discrete inputs* (DI) and *discrete outputs* (DO). A DO is a binary signal that can be set in one of the two states and a DI is a binary sensing that can detect one of the two possible states from a DO.

In a flexible robotic manufacturing cell, alternative actions often exist for robot control, such as which assembly task to perform first, or which machine to serve first. The ordering of these different alternatives can directly affect the utilization, throughput, and other measures of the cell. Due to its discrete nature, the cell control optimization problem is combinatorial in nature and expands rapidly with problem size. As a result, dispatching rules in scheduling are often employed as rules in rule-based cell controllers.

Additional concerns in cell control relate to two commonly used terms in production system control: *blocking* and *deadlock* (or *locking*). Blocking is a condition in which material cannot be transported to its next location because of a temporary resource unavailability. This can cause the waste of capacity of the current machine. Deadlock is a condition in which two resources mutually require the service of the other but neither can provide the required service at the current state. Therefore, the system will be deadlocked. Examples of blocking and locking are given in Figure 14.8.7.

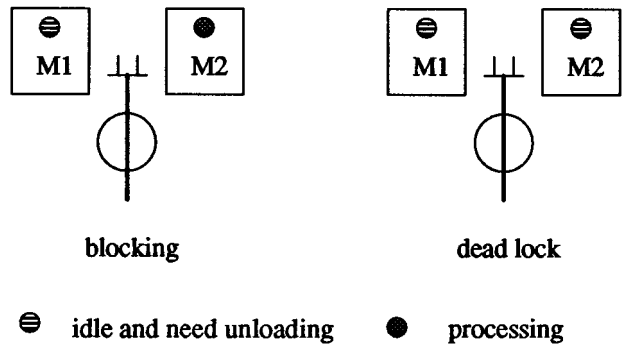


FIGURE 14.8.7 Example of blocking and locking.

In the figure on the left, the part on M1 needs to be transported to M2. However, M2 is in processing state and therefore the part in M1 has to wait for its completion. In the figure on the right, both parts on M1 and M2 are finished and need to be transported to the other machine. Suppose the robot has a single end effector; it is not able to transport either part to the other machine. One possible solution to blocking and deadlock problems is to add buffers. In the example, if a buffer location is added to the cell, the blocking and deadlock can be temporarily resolved. However, the addition of buffers to a cell can increase the cost of the cell and the complexity of the control. The real-time solution to blocking and deadlock problems lies in the identification of possible blocking or deadlocks and prevents them from occurring. Blocking or deadlock avoidance and system optimization are combinatorial in nature. The number of states expands exponentially with the number of states for each resource and number of part types. Therefore, rule-based systems and AI techniques find wide acceptance in cell controls.

Planning and Artificial Intelligence

Artificial intelligence (AI) is a branch of computer science studying the characteristics associated with human intelligence such as reasoning, acquiring knowledge, applying knowledge, and perceiving the environment. Path or task planning is the application of reasoning, knowledge acquisition, and perception

of the environment. Therefore, robot planning is closely related to the study of AI and certain AI techniques can be applied to robot planning and control. Some of the areas of research in AI that apply to robotic planning are problem solving, expert systems, learning, and machine vision.

The methodologies in path planning can be considered as problem solving. In assembly, when multiple components are assembled with precedence requirements, AI techniques can also be applied. Expert systems or rule-based systems solve problems in a discrete space domain based on rules derived from experts or other sources. Finally, machine vision has enjoyed a rapid increase in robotic applications. Machine vision can acquire complex environment information rapidly. Various algorithms can be used to extract useful path planning information such as locations of obstacles, end effectors, tools, etc. and can be used in real-time robot motion control. The reader is referred to Winston (1984) for more AI discussion.

Man-Machine Interface

Robots can commonly be programmed or controlled through teach pendants or a computer. A teach pendant is a small key pad that allows the user to move the robot, record positions, and enter simple programs. Modern robots are also accompanied by programming and control software that runs in microcomputers or workstations. The software environment often includes an editor, menu-driven robot control, and diagnostic utilities. More intelligent robot control programming is commonly supported in this environment than is available through the teach pendant.

Control programs can also be generated off-line. In *off-line programming*, the spatial configuration of the robot and work environment is modeled in the computer. A programmer is presented with a 2D or 3D world model of the robot and its environment graphically. The programmer will specify the locations and paths in this model rather than working with a real robot. Off-line programming has the potential to improve robot productivity and simplify the procedures of creating complex robot programs.

14.9 Design of Robotic Systems

Kok-Meng Lee

For manufacturing in which the manufacturing facility is concerned with similar volumes of production and a wider range of parts, the assembly line/mass production method is often not cost effective. It is often desirable to group equipment units together into workcells that can, in composite, perform an entire family of related operations on the product. The work-in-progress enters the workcell, remains while several functions are performed, and then leaves the workcell.

The individual equipment units that are used in the workcell (for both processing and materials handling) can consist of combinations of manual, semiautomatic, and fully automated equipment. However, in this section, the term "workcell" refers to a grouping of the robot and its peripheral equipment to assemble any of a large variety of products with little or no human intervention, driven by electronically designed data. An assembly robot is a comparatively simple mechanism whose function is to position parts and tools in the space of its work volume accurately. It is a comparatively low-cost machine of high precision of positioning and great reliability. Its simplicity, however, excludes the possibility of human-type actions like form recognition and its prehensile tools are very far from having the number of degrees of liberty a human hand has. If we concede that an assembly robot can by no means compete with a human being in a complex task, we also have to acknowledge that an assembly robot is capable of executing monotonous tasks with consistently high precision, thereby increasing the quality of the product. It can also keep up a fast production line indefinitely. Recognizing this difference between a human and a robot is essential in the design of a robotic system.

The remainder of the section is organized as follows. A set of design considerations for designing an assembly robot workcell is first presented. Layouts for a typical robotic workcell are then discussed. Experience so far has shown that in most instances, it is a feeder that fails in the workcell, not the robot. Feeding methods must be carefully considered when designing a workcell and are discussed at the end of the section.

Workcell Design and Layout

Design Considerations

Assembly systems can be broadly classified as manual, fixed, and flexible systems in relation to the complexity of the product to be assembled and to the production volume as shown in Figure 14.9.1. Flexible robotic workcells are typically used for less complex products at low or medium production volume, while for increasing product complexity, the cells designed for a single-purpose task can be linked into assembly lines. Apart from product volume and complexity, the design of the robotic workcell depends on several factors: namely, number of part types, end-of-arm tooling exchange, as well as product design.

Number of Part Types. A typical workcell consists of a robot and its peripherals made up of part-presentation mechanisms, feeders, conveyor, and end-of-arm tooling. For a small number of part types, parts are presented to the robot by feeders or magazines. As these take up space, only a limited number of different parts can be fed to one robot. In mechanical assembly normally a maximum of five to six different parts can be presented in this way.

To extend the robot's accessibility to a large number of parts, mechanized component feeding systems can be mounted on data-driven carousel conveyors spaced around the robot, each with a fixed dispensing point within reach of the robot gripper. The carousel can accommodate up to several hundred positions onto which magazines, tapes, or other modular dispensing systems can be attached. With multiple programmable carousels, the robot can access several thousand different parts. The application of the mechanized carousel is useful when only a few of each part type from thousands of styles may be used. Other alternatives are (1) kitting, in which all components to be assembled are kitted in a loosely

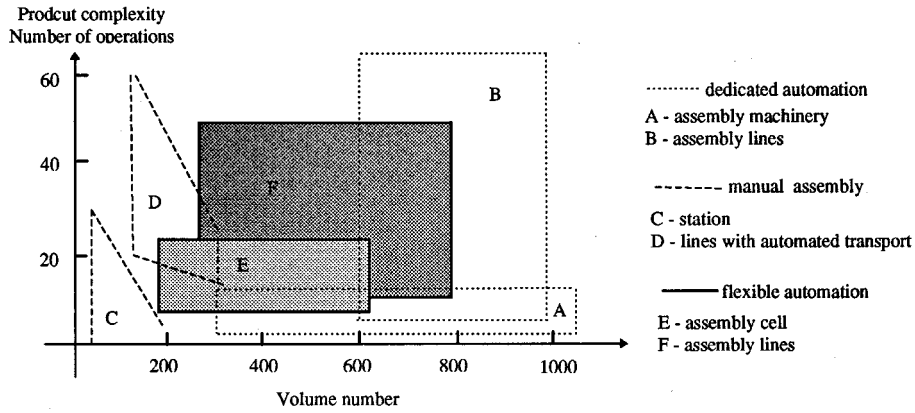


FIGURE 14.9.1 Classification of assembly systems.

palletized waffle pack, and followed by more accurate location using standard machine vision; and (2) the use of accurate totes for robot handling.

End-of-Arm Tooling Exchange. Many systems use different gripper exchange systems in order to cope with different parts. Tool exchanges are often considered as "nonproductive" since they do not contribute to assembly operations. The exchange is serially coupled to the assembly operations. This means that the cycle time increases due to the extra time needed for pickup and drop-off for tool changes as well as travel time between the assembly point and the end-of-arm tooling station. In order to reduce time loss due to the gripper, exchange should be minimized and/or in parallel with other activities, and the distance between pick-up point and assembly point should be very short. This problem could be avoided if a fast-revolving gripper head is used provided that space, weight, and cost of the revolving head do not pose a problem. Alternatively, the pallet carries batch-specific equipment such as grippers, fixtures, and end-of-arm tooling and can be presented to the robot on a conveyer in a similar fashion as the parts.

Product Design. Product design for flexible automation cells includes the following criteria: task operations based on flexible assembly cells for specific product families which must be able to assemble the variants of these product families using programming, fast changeover from one product to another within a flexible assembly cell, and reuse of standard elements for new assembly tasks

In addition to product complexity and volume, two other criteria should be considered in the construction of flexible assembly cells. First, since only a few products are generally suitable for fully automatic assembly, manual working processes are often essential with a large number of products. Flexible assembly cells must be constructed so that manual work stations can be included following ergonomic principles. Second, since the type-specific peripheral costs will increase in relation to the number of individual parts in the product to be assembled, part-specific feeders must be minimized for the economic use of flexible assembly cells.

Workcell Layout

Workcell design and layout in a flexible automation system depend on the nature of the manufacturing processes, the product design, and the material handling system as a whole. The manufacturing systems are classified as electronic product assembly, subassembly of electrical and mechanical components, and kitting cell for large-scale manufacturing.

Electronic Product Assembly. Flexible workcells are commonly used for the assembly of integrated circuit boards (PCB), where a combination of interchangeable part-feeding mechanisms are used to present parts to robots. Since a majority of the processes involved are carried out in the linear, vertical plane, robots of SCARA or gantry construction are best suited for these assembly tasks. The workcell

consists of a robot and its peripherals made up of part-presentation mechanisms, feeders, conveyor, and end-of-arm tooling.

Figure 14.9.2 shows the organization of a typical workcell for assembly of a family of circuit boards (Decelle, 1988), which is a part of the in-line component insertion, inspection, and repair assembly line. Circuit boards to be assembled are secured on panels and flow through the workcell on a conveyor. Each of the circuit boards is characterized by a bar-coded serial number that permits product tracking, data collection, and testing through the assembly. Boards requiring assembly are positioned over an elevator mechanism in the workspace of the robot. The mechanism lifts the board slightly and uses the tooling holes on the panel to locate the circuit board. Two digital signals interface the conveyor to the workcell — one signals the robot that the board is ready for assembly and the other signals the conveyor to index the board to the next workcell. Components are fed to the robot by using feeders. The feeder singles out components to a walking-beam mechanism that transfers parts through lead-cutting, lead-straightening, and lead-verification operations and on to the lead-locating nest for robot pickup. The activities of the robot and its peripherals in a workcell are coordinated by a host computer. The workcell is set up and monitored through the host computer. Through the host computer, the operator provides the workcell the code to be assembled, the components in the feeders, and the configuration of the feeders.

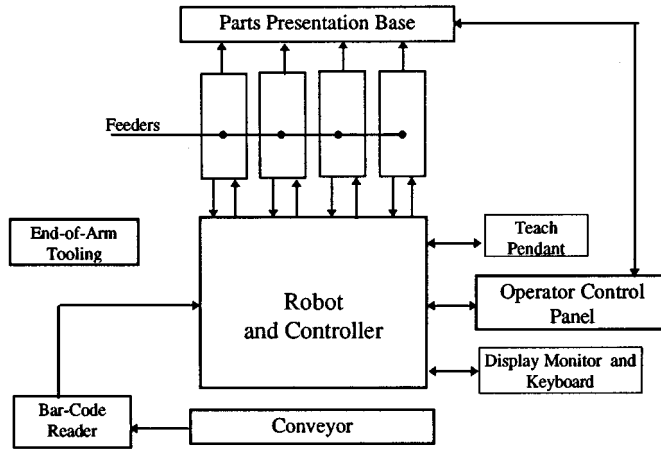


FIGURE 14.9.2 Workcell for elctronic assembly.

Subassembly of Electrical and Mechanical Components. Unlike PCB assembly, packaging and designs of small electrical and mechanical components are generally nonstandardized. Thus, the problem of automated flexible assembly workcells lies in the presentation of parts and the degree of flexibility of assembly of small components often involves both product design and layout considerations extensively.

Figure 14.9.3 shows a self-contained flexible workcell for assembly of small mechanical parts with a circular indexing table. Modular part-feeding equipment such as vibrator feeder bowls and special-purpose trays are placed around the indexing table to feed and to orient small components to the robot. Typical mechanical operations such as riveting, screwing, welding, inserting, pressing, and so on are achieved through quick changeover end-of-arm tooling. The circular indexing table arrangement is advantageous where end-of-arm tooling changes are necessary for handling different parts. It allows changes of end-of-arm tooling to take place while other operations are continuing.

With complex products, assembly in a single flexible assembly cell is not always feasible. In this case, a flexible assembly line can be designed to link self-contained independent workcells (Figure 14.9.3) so that they can be engaged or disengaged as required to allow adaptability in connection with product model change. Alternatively, standard carriers or pallets can be used to present a large number of different part-types to a robot. Each pallet carries a large number of identical parts, unoriented but

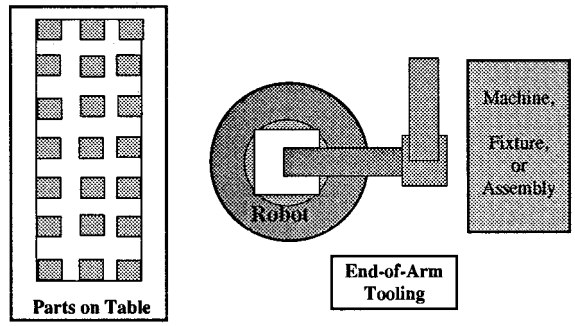


FIGURE 14.9.3 Typical single-purpose workcell.

with the right side up, and placed on a flat board. Standard machine vision was used to detect the orientation of the parts.

Kitting Cell for Large-Scale Manufacturing. In the field of large-scale manufacturing such as automobile manufacturing, engine assembly, and machining processes, where the setup time of specialized tools for each task is excessive, the work is generally distributed into several cycle zones. As an example, actual cutting time (production time) represents a value between 5 and 20% of average machine utilization time that includes nonproductive time accountable by workpiece load/unload, too) change/setting, and workpiece inspect.

To avoid a high level of wear and tear on tools due to constant conversion, the cycle zone is commonly divided into individual operating cells which may be interconnected in series, parallel, or a combination of series and parallel. A typical workcell (Figure 14.9.3) consists of a robot, a part-feeder, an end-of-arm tooling section, and the manufacturing process. The parts are contained in a regularly spaced pallet, which are transported by means of an automated guided vehicle (AGV) or a conveyor to the loading tables and are fed to process by the robot. The most common approach in automated part presentation for machine loading is the use of specially designed pallets for each part family to maintain sufficient position accuracy for a completely preprogrammed robot picking.

In the case of assembly, purchased parts or parts to be processed are kitted onto one kit tray in a single location. Kitting is the process of taking parts from bulk and placing them on a kit tray, which is an organized group of parts. Concentrating the material delivery system and its control to one area is the main benefit of the kitting cell. In addition to efficient use of floor space by eliminating duplicate equipment at each assembly cell, the feeders and tooling are universal — the same equipment is being used all the time for all parts, thus maximizing utilization while minimizing capital expense. The material delivery equipment is eliminated at the assembly cycle times. Also, having all the parts for an assembly on a carrier permits changes in the process route during machine downtime or blockages.

Figure 14.9.4 shows a layout of the kitting cell. An overhead gantry takes bins of parts and dumps them in the appropriate feeders (indicated in Figure 14.9.4 as F1, … F7). The feeders fill the lanes with an initial quantity and replenish them as parts are kitted. The parts come to rest in nests at the end of the feeder lanes. Here the vision system verifies the correct part family, performs some quality checks, and determines the position and orientation for the robot to pick the parts. Should the vision reject the part, the nest will dump the part and a new part will be fed in for an inspection. Using a quick changeover gripper, multiple parts are kitted onto a tray. Once all the parts are on the kit tray, the tray is indexed to the inspection station for verification that all parts are placed. The robot takes the completed kit tray and places it on the assembly conveyor to an idle station, ready to be picked up by an AGV.

Part-Feeding and Transfers

The term “part-feeding” refers here to feeding workpieces from pallets using a preprogrammed robot for subsequent processes such as machining or assembly. The cost to feed parts to a robot for either

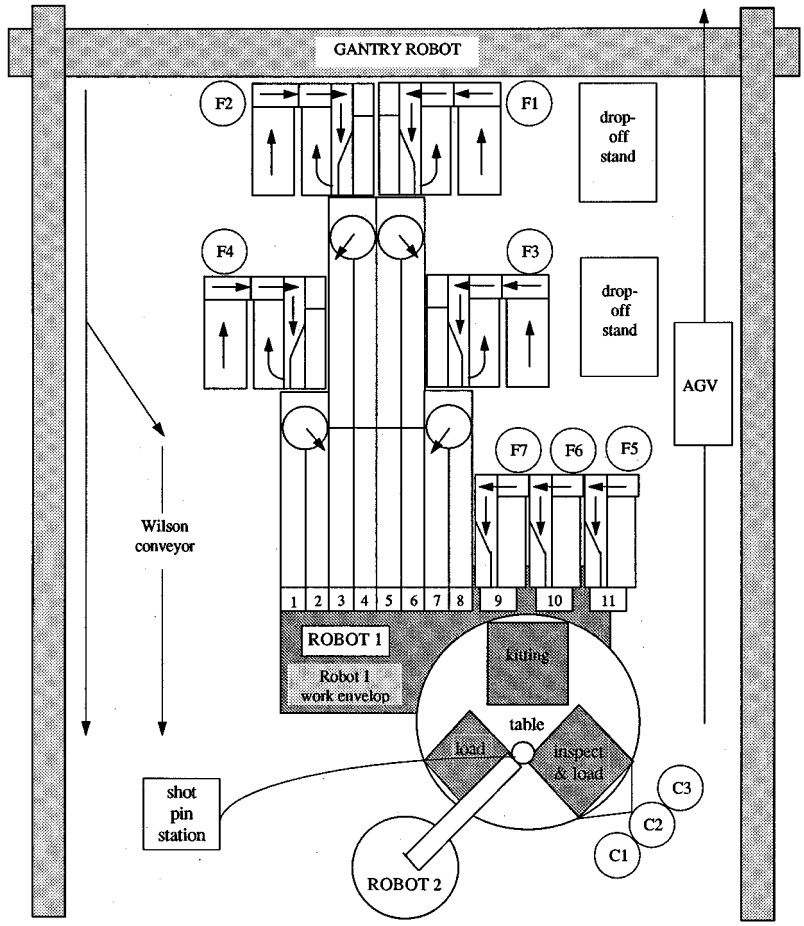


FIGURE 14.9.4 Schematics of the kitting cell.

machine loading or assembly in a flexible manufacturing system (FMS) has often been underestimated, which may comprise as high as two thirds of the overall investment and is usually the source of a large percentage of work stoppages and defects. A general review of existing *mechanical part feeders* can be found in Lee (1991).

The basic kinds of part-feeding may be classified as follows: (1) mechanical feeders which are designed to feed and to orient the parts-dedicated part-feeding apparatus, (2) dimensionally dedicated pallets which are specially designed for each part family to maintain the position/orientation, and (3) machine vision.

Mechanical Feeders

The commonly used mechanical feeders for robotic assembly are bowl feeders, vibratory feeders, and programmable belt feeders. For large volume manufacturing, the employment of the dedicated mechanical part-feeding apparatus may be justified. However, mechanical feeders consume a lot of room around the workcell, often fail due to jamming, and, most significantly, generally require retooling when a component is changed or tool wear is caused by jamming.

Vibratory Bowl Feeders. Vibratory bowl feeders (Boothroyd and Dewhurst, 1985) are most commonly used as mechanical feeders for robotic assembly. The basic component of a bowl feeder consists of a

vibratory bowl, a rotating disk, and an orienting track. Parts to be fed to the robot are separated into a single line and oriented to move to feeding end. These feeders, in general, are not designed to be easily converted to feed new part types. The cost of the bowl feeders can be broadly divided into two parts: special purpose equipment cost and general purpose equipment cost. Typically, changeover would involve replacing the bowl, orientation track, feed track, and escapement, which contribute to special-purpose equipment cost. Only the vibratory drive unit could be reused. This general-purpose portion of the feeder is approximately 30% of the feeder cost.

One way to lower the cost of the bowl feeder per part is to deliver different parts to a robot assembly station using multiple layer vibratory bowl feeders. A multilayer bowl feeder has several bowls mounted in stacked fashion, and in each bowl a different kind of part is stored. The design of multipart vibratory feeders aims at reducing the cost of the vibratory feeders by sharing the general-purpose hardware cost over several parts and by reducing the special-purpose tooling cost. Two basic forms are available: bowl type and in-line type.

To change over this multipart vibratory feeder to other part types, the orienting tracks must be replaced. An effective way to reduce wear is to separate the function of orienting from feeding. The function of the multilayer vibratory feeders is to restrict feeding parts to a separation unit. Parts of several different types are fed but not oriented from a vibratory feeder. In most cases, the workpieces must be held by a mechanical pusher against a pair of orthogonal datum planes on a relatively flat surface with the "right side up." A machine vision system is then used to locate and/or to sort the orientation of the parts using two-dimensional binary images, which is a great deal easier to store and to process.

Vibratory Belt Feeders. In vibratory belt feeders, parts are fed by a vibratory conveyer belt (Boothroyd and Dewhurst, 1985). The principle of the vibratory belt feeder is to produce a vibratory motion on the surface of the brushplate. The motion is obtained by pulling the brushplate sharply down and back and then allowing it to spring up and forward. This action, repeated at high speeds (approximately 3600 times per minute at 60 Hz power supply), produces a definite vibrating movement on the brushplate surface, permitting parts to be conveyed in a smooth and easily controlled manner.

The orienting systems used on these belt feeders may be a mechanical device, an optical sensor, or a vision system. A machine vision system is often used to locate and/or to sort the orientation of the parts using two-dimensional binary images. A line-scan camera is commonly used to create the silhouettes of the workpieces, and in some cases the product designs can be reviewed to simplify the vision algorithm and to reduce the system cost. Since a robot gripper can grasp parts from a queue on the feeder itself, belt feeders do not require any special-purpose tooling for feed track or escapement and thus offer several advantages over the vibratory bowl feeder for robotic assembly

Dimensionally Specific Pallets

One of the most common approaches is the use of specially designed trays or totes for each part family to maintain sufficient accuracy for a completely preprogrammed robot picking. A particular form of these dimensionally precise feeders is known as tape-and-reel for feeding parts of relatively small sizes, which can be placed on tapes of standard width. For some devices that are large, heavy, ceramic, or have fragile leads, tapes are very expensive and impractical.

In general, the dimensionally specific pallets are well suited for large volume production where changes of part types are not frequent. The operational cost of the design-specific pallets includes packaging costs for transport, construction cost for pallet alignment, and engineering cost for new pallet designs.

Vision-Based Flexible Part-Feeding

For flexible manufacturing, where a large variety of product sizes and component types are encountered, the part-feeding system must have the ability to adapt to a changing product design without costly hardware redesign or time-consuming software reengineering. This need has been addressed as a general industrial vision-based bin-picking problem by several authors.

In manufacturing automation applications, the processing speed of acquiring and analyzing an image must be comparable to the speed of execution of the specific task. The attempt to duplicate human

perception by obtaining a three-dimensional detailed image of the part often calls for time-consuming computation and does not necessarily determine the location and orientation of a given part with the accuracy required for successful part acquisition by the robot. Moderate location inaccuracies pose no difficulty for human operators since they use vision, hand-eye coordination, and sense of touch to locate and correctly load the part.

However, if the orientation of the parts can be characterized by the two-dimensional object silhouette, retroreflective materials can be used as a background in generic part presentation (Lee and Li, 1991). Most surfaces on objects exhibit a combination of diffuse and specular reflection. A point on an ideal diffuse-reflecting surface appears equally bright from all viewing directions. Surfaces covered with papers and matte paints may be considered as reasonable approximations. An ideal specular reflector is one that reflects any incident light ray as a single ray in the same plane as the incident ray and the surface normal. The basic principle of the retroreflective vision sensing is to structure the surface reflectance of the pallet or the landmarks so that it is much brighter than objects generally characterized by diffuse or specular surfaces. In practice, a number of nonpredictable factors such as measurement noise, the uniformity of the surface reflectivity, and the uniformity of illumination, which occur on both the object and the background, can be eliminated by a relatively simple technique. If part design can be modified, brightly illuminated retroreflective landmarks can be intentionally created on objects for location tracking. Low cost landmarks could be incorporated in design by using retroreflective liquid paints on existing features. Alternatively, generic landmarks can be constructed by applying solid glass beads on the reflected surface of standard fastening devices such as screw heads.

14.10 Robot Manufacturing Applications

John W. Priest and G. T. Stevens, Jr.

Product Design for Robot Automation

Identifying automation opportunities early in product design is important because product design requirements to facilitate robotic manufacturing are often unique and must be integrated early in the product design process. Some overall manufacturing problems for using robots and some design solutions to resolve these problems are listed in Table 14.10.1.

TABLE 14.10.1 Design Solutions for Robots

Problems in Utilizing Robotics	Design Solutions to Assist Production
Location accuracy and repeatability	Design for vertical assembly; use chamfered edges for mating surfaces; tolerance leeway for mating parts
Part feeding and orientation	Design parts which can be easily fed, provide notches, guide pins, or slots for part orientation; select parts from vendors that will deliver in easy-to-feed packaging
Programming robot and associated equipment	Design simplification; use common parts for different products, part reductions; part families
Application problems with fasteners (screws, washers, and nuts)	Minimize the use of all fasteners; utilize snap fits where possible
Downtime caused by jams and misfeeds due to poor part quality	Select vendors that produce high-quality parts

Rossi highlighted the product designer's role in robotics stating this problem (Rossi, 1985):

> Often designs are made in such a fashion that one cannot access a certain area with a robot. Humans can get around obstacles and operate within those designs easily, but robots cannot because they are not quite as flexible as human beings. I think that this is the single most important item that has kept us from being further along than we are. What happens is that users try to apply a robot to something that's been designed without robotic assembly in mind. They usually run into a problem. Either the robot cannot handle it at all or the users find that they have got to put a lot of additional engineering design into a particular workcell, or perhaps into an end effector, in order to get around the problem. All this does is add to the price tag, and cost is very much in consideration when one is trying to sell these systems. A situation arises where robots are no longer attractive because of all the additional things that need to be done.

In summary, the product must be designed for the manufacturing process and the robot. For more information, the reader should review Boothroyd (1994), Bralla (1986), Priest (1988), and Tanner (1994). Table 14.10.2 shows some design rules for robotic assembly.

TABLE 14.10.2 Design Rules for Robotic Assembly

Product should have a base part on which to build assemblies in a top-down, straight-line motion direction
Base should be stable and facilitate orientation
Parts should be able to be added in layers
Use guide pins, chamfers, and tapers to simplify and self-align the layering of parts
All parts should accommodate handling by a single gripper and be comparable with popular feeding methods
Sufficient access is available for the gripper
Avoid the use of bolt-and-nut assembly
Parts should be able to be pushed or snapped together; when screws are necessary for repair, they should all be the same size
High quality parts are used
Vendors deliver parts that are compatible with the selected part feeder mechanism

Economic Analysis

Economic analyses for robotic applications are similar to those for any manufacturing equipment purchase and usually use minimum annual revenue requirements, present value methods, or break-even analyses. Since robots are a flexible method of automation, a unique aspect of robotics is manufacturing's ability to reuse the robot after its initial production run for other applications in later years. For many companies, this subsequent use of the robot can be shown in the economic evaluation. Some other unique benefits in robotic economic analysis that may be included are improved quality, higher precision, ability to run longer shifts, and reduced floor space. Unfortunately, some unique disadvantages of robot analysis are software integration complexity, inability to respond quickly to product design changes, and process reliability.

In general, there are several situations where robots are more likely to make economic sense. These are

A. Sufficient volume to spread investment costs over many units
   1. High volume
   2. Stable product design
   3. Multishift operations
B. Robot is used on more than one product
   1. Limited number of different products on same production line
C. Part handling problems occur when performed manually
   1. Parts that are very large, heavy, or bulky
   2. Parts that are very fragile or easily damaged
   3. Parts that are extremely small
D. Extremely difficult manufacturing process without using robot or automation
   1. Many processes, especially in electronics, cannot be performed without robots or some type of automation
E. Safety and health concerns of process
   1. Safety and health costs can be significant

The type of data concerning the robot system that is required for an economic analysis is shown in Table 14.10.3.

TABLE 14.10.3 Economic Cost and Savings for Robot Applications

Investment costs
Robot purchase price — for many applications this is a much smaller part of the costs than expected (25 to 45%)
Other equipment (part feeders, conveyors) — this includes the cost of hardware interfaces
Design of end effector, special fixtures, and other equipment — most applications require the design of a unique end effector and special fixtures
Software design and integration — can be a much higher cost than expected due to the complexity of interfacing different equipment controllers
Installation including facility modifications — usually a small cost for robot system
Technical risk — this is the risk of whether the system will perform up to the specifications in areas such as performance, quality, precision, etc.
Operating costs
Training — costs of training operators, engineering and maintenance personnel
Product design changes — cost required to modify the robotic software and hardware when design changes or modifications are made to the product
Operating, utilities, and maintenance — typical costs found for most manufacturing equipment
Savings
Direct labor — labor savings caused by the robotic system
Ergonomic and health — benefits of lower number of job injuries, workers compensation costs, and compliance with OSHA regulations
Quality — improved quality may result due to lower scrap and waste
Precision — robots can often perform tasks at a much higher precision (i.e., lower variability) than manual operations resulting in fewer defects and better product performance

Cost Justification for Robots

In this section an example of a robot justification study is presented. This example uses the discounted cash flow method resulting in the calculation of a rate of return (often referred to as the internal rate of return).

The rate of return, R, is defined by Equation (14.10.1) as

$$0 = \sum_{j=0}^{n} \frac{X_j}{(1+R)^j} = \sum_{j=0}^{n} X_j (P/FR, j) \tag{14.10.1}$$

where

X_j = the net total cash flow for year j
n = number of years of cash flow

Basically, the rate of return, R, is the interest rate that makes the sum of the discounted cash flows equal zero.

The net cash flows, X_j, in Equation (14.10.1) can be defined by Equation (14.10.2):

$$X_j = (G - C)_j - (G - C - D)_j (T) - K + L_j \tag{14.10.2}$$

where

G_j = gross income (savings, revenues) for year j
C_j = the total costs for year j exclusive of book (company depreciation and debt interest)
D_j = tax depreciation for year j
T = tax rate (assumed constant)
K = total installed cost of the project (capital expenditure)
L = salvage value in year j

A numerical example is now presented using the data given in Table 14.10.4. The values in Table 14.10.4 should not be considered representative. They are used only for example purposes.

TABLE 14.10.4 Data for Numerical Example

Project costs	
Purchase price of robot	= $40,000
Cost of end effector	= $10,000
Software integration	= $20,000
Cost of part feeders	= $10,000
Installation cost	= $ 4,000
	$84,000
Yearly operating costs	= $10,000
Yearly savings	
Labor	= $60,000
Quality	= $10,000
Tax rate	= 40%

The first question that must be answered is, what is the capital recovery period (life of the economic study)? In this example 3 years is used. The next question is, what is the yearly tax depreciation? This example uses MACRS (3 years). Therefore, the percentages used to determine the yearly tax depreciation amounts are those given in Table 14.10.5. It should be noted that Table 14.10.5 implies there are 4 years of tax depreciation. This is because the MACRS system uses a half-year depreciation convention. In this example, it is assumed there is sufficient taxable income from other operations that allow the use of the depreciation amount in year 4. In this example, the salvage value is assumed to be zero.

TABLE 14.10.5 MACRS Percentages

Year	Percentage
1	33.33
2	44.45
3	14.81
4	7.42

Using Equation (14.10.2), it is now possible to generate the cash flows shown in Table 14.10.6.

TABLE 14.10.6 Net Cash Flows

EOY	K&L	G	C	D	X
0	$84,000	—	—	—	84,000
1		$70,000	$10,000	$27,997	47,199
2		70,000	10,000	37,338	50,976
3		70,000	10,000	12,440	40,976
4	L = 0	—	—	6,224	2,490

Some sample calculations follow:

$$x_0 = -\$84,000$$

$$x_1 = (70,000 - 10,000) - (70,000 - 10,000 - 27,997)(.40)$$

$$= \$47,199$$

$$x_4 = -(-6224)(.40)$$

$$= \$2,490$$

With the cash flows in Table 14.10.6 the rate of return can be determined using Equation (14.10.1):

$$0 = -84,000 + 47,199(P/FR,1) + 50,935(P/FR,2) + 40,976(P/FR,3) + 2,490(P/FR,4) \qquad (14.10.3)$$

To determine R in Equation (14.10.3), a trial-and-error solution is required. Assuming 20%, the right-hand side of Equation (14.10.3) gives $15,619 and with 25%, it is $–17,262. Therefore, the rate of return, using linear interpolation, is

$$\begin{array}{cc} 20\% & \$15,619 \\ R & 0 \\ 25\% & -17,262 \end{array}$$

$$R = 20 + \frac{15,619}{32,881}(5)$$

$$= 22.38\%$$

This rate of return (22.38%) is now compared to a minimum acceptable (attractive) rate of return (MARR). If R ≥ MARR, the project is acceptable. Otherwise, it is unacceptable. It is pointed out that the definitions of cash flow and MARR are not independent. Also, the omission of debt interest in Equation (14.10.2) does not, necessarily, imply that the initial project cost (capital expenditure) is not

being financed by some combination of debt and equity capital. When total cash flows are used, the debt interest is included in the definition of MARR as shown in Equation (14.10.4).

$$\text{MARR} = k_e(1-c) + k_d(1-T)c \tag{14.10.4}$$

where

k_e = required return for equity capital
k_d = required return for debt capital
T = tax rate
c = debt ratio of "pool of capital" used for current capital expenditures

In practice, it is not uncommon to adjust (increase) k_e and k_d for project risk and uncertainties in the economic climate. There are other definitions of cash flow definitions (equity and operating) with corresponding MARR definitions. A complete discussion of the relationship between cash flow and MARR definitions is given in Stevens (1994).

Assembly

Assembly is projected to be the largest area of growth for robots. Key design goals for robotic assembly are to ensure high-quality parts, minimize the use of fasteners and cables, and provide accessibility so that parts can be easily fed and oriented by automated equipment. Designing to facilitate the use of robotics requires a review of their capabilities. Although assembly robots are often shown as stand-alone equipment, they require considerable amounts of support tooling and auxiliary equipment. These include part feeders, end effectors, special fixturing, and a material handling system. Except in the case of robots with vision or special sensors, parts with which the robot will interact must be precisely located and oriented. This may require additional tooling or special vendor packaging.

Assembly is defined as the combining of two parts into one entity. This combining process may include (1) the use of mechanical fasteners (i.e., screws, snap fits, rivets, etc.); (2) joining processes such as welding, brazing, soldering, etc.; (3) application of adhesives; (4) the simple process of placing two parts together to be joined together later.

Robotic assembly is the use of robots to perform one of these assembly processes. A typical set of tasks for robotic assembly using mechanical fasteners might be

1. Go to location (x_1, y_1, z_1) and grasp part A (assumed to be properly positioned and oriented).
2. Place part A in a fixtured assembly position (x_2, y_2, z_2, including proper orientation).
3. Go to location (x_3, y_3, z_3) and grasp part B (assumed to be properly positioned and oriented).
4. Place part B on part A (x_4, y_4, z_4) including proper orientation.
5. When an additional process is needed, fasten or join part A to part B using process tooling.

As can be seen in this simple list of tasks, developing robotic assembly system focuses on getting the parts to be assembled in the proper position and orientation and the combining process itself. Because of this, the rest of this section will describe the parameters of these two aspects: part feeding and presentation and the combining process.

Part Feeding and Presentation

Robot assembly requires the robot to go to a predefined location and grasp a part. The part may be positioned and oriented or it may not. Since a positioned and oriented part is preferred, part-feeding methods that can perform this task are desired. The most popular types are

1. Vibratory bowl feeders
2. Pallets and trays
3. Specialized feeders
4. Special vendor packages

5. Conveyors

Vibrating bowl feeders are one of the most popular methods due to the large number of parts that it can feed and its cost effectiveness. Pallets, in turn, are popular for many electronic parts and fragile parts where the part cannot withstand the forces found in a vibrating bowl feeder. Specialized feeders are those feeders that are usually designed for a particular type of part. These can include tube feeders, magazine feeders, and slides. Vendors can often provide parts in specialized shipping packages which keep the parts in the proper position and orientation. Finally, when the parts are delivered by conveyor, special fixturing and stops can often be placed on the conveyors to position/orient the part.

When the part is not positioned or oriented, additional sensors must be added to the system. Commonly used sensors are

Machine/robotic vision
Simple sensors such as photodiodes
Tactile/touch sensors

Robotic vision is becoming more popular in assembly as their purchase, software integration, and installation costs continue to decrease. Although most robot manufacturers offer vision systems as an option, they are still an expensive addition to the system. Simple on/off sensors can be used in certain cases when only limited data are required. Tactile sensors can sometimes be used to touch/feel the part to identify specific features of the part or to recognize its location.

Combining Process

After the parts are placed together, the combining process will often require the robot to perform some process operation. This can include an additional equipment such as a fastening gun for a screw, adhesive applicator and pump for a bonding operation, or a solder gun for soldering. Most robot manufacturers can offer equipment for the various types of assembly processes.

For more information on robotic assembly, the reader should review Asfahl (1992), Groover (1986), Klafter et al. (1989), and Sandler (1991).

14.11 Industrial Material Handling and Process Applications of Robots

John M. Fitzgerald

Replacing humans with robots to perform processes has often led to failure. The reason is that the robots are often mechanically capable of the manipulation while being incapable of process planning and control. Thousands of robot installations have failed because replacing the manual method with the automatic method lacked adaptability to process related variation. The human operators had been using their cognitive abilities to do the job. A vast majority of successful robot implementations past and present have a very important common aspect: repeated execution of fixed programs with little or no on-line modification of path or position.

Process robot planning and programming still usually require the efforts of highly skilled technicians. Often, complex programs cost too much and take too long. Continuously controlling and varying path manipulation parameters for real-time process control is difficult. Many processes are not known well enough to describe their control algorithmically. In a few applications sensors are becoming more common for adapting robot plans to changes in the environment. Setup, seam tracking, positioning, conveyor tracking, and now automatic programming for painting and finishing are becoming practical as sensor costs and computation costs continue to decline.

In this section robotic material handling and process applications are presented from an automation system perspective focusing on the robot's manipulation functions. Manipulation is considered a manufacturing material transformation and a transportation process factor. Programming and control are viewed as the means of integrating robot manipulation as part of the manufacturing process. The reader who is interested in a specific application is encouraged to first review the relevant process technology sections of this book.

Implementation of Manufacturing Process Robots

Manipulation as a Process Requirement

The starting point of automation system design is a thorough understanding of the process to be automated. Implementation of a process robot requires a focus on manipulation as a process factor. The pose and path requirements of the process are independent of the manipulator used.

It is useful to conduct a static spatial analysis of manipulation requirements and then examine the mechanical and dynamic requirements when designing or selecting a process robot manipulator.

A spatial description of the relative positions and orientations of the workpiece and tool during processing provides the basis for describing the required manipulation. *Tool poses* are graphed in an appropriate reference frame, usually the frame of the workpiece, or in the case of machine loading, the work holding fixture may be used. Path requirements are secondary for these applications. The path taken does not affect the process. For *continuous path processes* entire paths must be graphed or mapped. If continuous analytical descriptions of the path are not available, a sampling of discrete points along the required path can be used to represent the space occupied by the path. The result in both cases is a Cartesian mapping of spatial requirements of pose and path. A description of the pose and path precision requirements should be included. Next the mechanical and dynamic requirements are defined. Payload and force reactions at each position and along the path must be understood. Other important dynamic requirements such as acceleration and power should be quantified. The manipulation requirements are the basis for design and selection of both the robot arm and the controller.

Manipulation Capability of Process Robots

The basic mechanical capability of the robot mechanism to perform the manipulation work is determined by its mechanical structure, kinematic configuration, and drive mechanism. There are several applications including painting, palletizing, spot welding, and arc welding for which specific types of robot arm

designs have evolved driven by process needs. Although predisposed by design to perform a particular process, these robots have no innate process capability and are not guaranteed to perform in a specific application. Specifications of gross robot performance characteristics such as reach, *repeatability*, accuracy, and payload are usually readily available from their manufacturers. A well-defined set of process manipulation requirements when compared with published robot specifications usually isolates the field of mechanically qualified candidates. It is more difficult to characterize and evaluate a robot's capability for complex motion. The exact working of the robot's trajectory generation software is usually not known by end users and can only be evaluated by indirect testing. Acceleration and load capacity are usually specified, and there are some standard methods for specifying path performance, but the robot's dynamic behavior and performance are difficult to measure. Specific performance testing is usually required to prove manipulability for process robot applications.

Integration of Manipulation Control and Process Control

Achieving manipulator and process control integration depends upon robot programming and external data access. For any given application the required motion execution may be possible, but programming may be too difficult to be practical. Establishing that the robot is capable of coordinated motion can be done by reviewing the specifications or by conducting motion tests. As an illustration of the importance of programmability consider, for example, a situation in which a complex series of twisted curves define a robot tool path. If two robots with identical kinematic structure and joint trajectory generation capability differ in their programming in that one is capable of executing paths following user-defined mathematical functions and the other is only capable of executing paths defined by closely spaced taught poses, the difference in programming effort could easily amount to hundreds of hours. For each application encountered the programming methods must be assessed to determine if the required motion is programmable in a practical sense.

Access by process robot programs to external data is becoming more important. Although most process robots now work without any external process feedback, this is beginning to change rapidly with the development of improved low cost sensor systems and methods. Virtually all robots are capable of discrete digital and analog signal input and output. Most may also be equipped with standard serial and parallel communication capacity. If sensor information is to be used for set-up positioning or real-time path adjustment, the robot controller must have the communication and control to convert data into information that can be used to modify path and position commands. In cases of extreme path complexity, path planning systems external to the robot controller may be needed to create the paths. Testing will verify the ability of the robot controller to accept and execute externally generated motion sequence data.

Industrial Applications of Process Robots

Palletizing and Depalletizing

Many products are packaged in boxes of regular shape and stacked on standard pallets for shipping. Robots are commonly used to palletize and depalletize boxes because they can be programmed to move through the array of box positions layer after layer. Although palletizing is more common than depalletizing, there is no major functional difference in the manipulation requirements. Transport distances of several feet are common. Stack heights usually do not exceed 5 ft. Payload weight can be in excess of 100 lb. When standard servo-driven joint actuators are used accuracy and repeatability will usually be far better than the required box positioning precision.

Palletizing typically requires four axes of controlled motion — three for translation and a fourth for yaw to orient the box. Cylindrical coordinate robots are favored in palletizing because they have large vertical lift and a compact footprint allowing more of the floor area in the workspace for conveyors and pallets. When larger workspace is needed gantry robots must be used. Continuous duty cycles are not uncommon and robot power is important for maximizing throughput. The most technically demanding aspect of system design is the gripper. Vacuum grippers are popular for lifting boxes by their tops, but other more complex gripping methods are sometimes needed. Payloads must be carefully positioned

with respect to the robot's wrist and other links to balance gravitational and dynamic loading. Load shifting during high acceleration moves can result in dropping or mislocating the box.

Palletizing position arrays are usually taught or programmed relative to a corner or keystone box position as a reference so that the entire array can be shifted by redefining that one position. Programs are simple and easily modified to adapt to changes in box dimensions. Monitoring is done by checking the state of discrete proximity and vacuum sensors. A proximity sensor mounted on a gripper will indicate if an object is at an expected location; or the same simple proximity sensor may be used to stop the robot in the correct location to pick up a box from a stack of unknown height when the top of the box is encountered. Vacuum pressure switches are often used to verify acquisition by suction cup. A simple proximity switch can be used to signal the presence of an expected package at the pick-up point. With careful timing and additional sensor inputs, items can be transported to and from moving conveyors.

Packaging

Packaging is often a combination of palletizing and assembly-type actions. A collection of objects which may not be identical are inserted into a box or other container. The robot may also be required to assemble, place dunnage, seal, or mark the package. Insertion may simply require positioning the pack item over the opening of the package and dropping it. Boxes most often are supplied partially assembled, printed, and folded flat. Usually human operators or a special machine will open and prepare the box for packing; rarely will the robot be used for this purpose. Often the robot can be used to place cardboard layer separators, foam, or cardboard holding forms and other protective dunnage in the box. Finally, sealing and marking operations may be performed by the robot. Pack items may require complicated assembly-type motions such as rotations and curved moves to clear other packed items.

Three to six axes of motion may be needed. Packing items with a variety of sizes, shapes, and other varying physical properties into one package have the potential to complicate motion and tooling requirements. Grippers can be designed with multiple functions or they can be designed to be exchanged by the robot at tool storage racks. When material throughput is high, a single robot may be dedicated to each pack item. Simple programming methods are employed such as teach programming. Discrete sensors are useful for monitoring grip status of pack items.

Machine Tending: Loading and Unloading

Forges, stamping process, some machine tools, and molding machines are now commonly tended by robots. Historically these types of machines have been loaded by human operators. Now these jobs are considered to be too arduous and hazardous. An important benefit of robotic machine loading is improved product quality resulting from consistent machine cycles. Robots eliminate the inconsistencies of human-paced loading and as a result the cycle can be precisely repeated. For heated molding, stamping, and forging processes, part formation and release are sensitive to the thermal state of the machine. If a machine is left open for loading for differing amounts of time each cycle, significant cooling variations result in potential sticking and geometric flaws. When robots are used, the process can be tuned to the consistent robot loading cycle.

Machine loading is usually more demanding than other material handling applications because part orientation and placement are critical and may require locating mechanisms such as tooling pins and pads and/or sensor logic to guarantee interface between the robot and the serviced machine. Accuracy is usually not an important factor because the loading stations are permanently located in the robot workspace, but repeatability requirements may be as small as several thousandths of an inch. Payloads can range from a few ounces to several hundred pounds. Grippers for machine loading may also require tooling pins and pads to locate and orient parts and to mate precisely with the machine's part holding fixture. The gripper may dock with the holding fixture and then transfer the part when loading clearances are very tight.

The entire range of robot types, sizes, and configurations is used for machine tending. Articulated arm robots are needed when dexterous manipulation is required to transport parts through the maze of clamps and spindles and other protrusions and obstacles found on some machines or when part orientation

must change for loading. Applications where the robot is dedicated to loading a single part into a single machine in high volume production are not uncommon. Position programming is usually done by teaching. It is common to monitor discrete sensors in the gripper and the loaded machine to insure proper loading before cycling the process machine.

Sorting

Discrete parts are often sorted during production, usually as a condition of transfer to the next production station. The sort characteristics are usually distributed in some unpredictable manner so that individual inspection and handling are required. The diflference between sorting and other transfer or loading robot applications is that the disposition of the part is based on information gained during the sort. The robot must have the programming functions to support multiple preprogrammed path execution triggered by the logical sort outcome conditions.

Part Dipping

Many processes require controlled manipulation of parts temporarily submerged in some working fluid or coating material. Some common part dipping processes are the following.

Investment Casting. Intricately shaped and often delicate wax forms are coated with a slurry of stucco material which cures to form a mold. Later the wax is melted and drained from mold which can then be filled with molten metal. The dipping motion must be carefully controlled to prevent trapping bubbles and distorting the wax shape.

Solder Pretinning. Electrical contact pads and component leads are coated by dipping in molten solder as a preliminary step to assembly and soldering of the connections. A temperature-dependent flux reaction is required to achieve wetting by the solder so the robot must hold the component submerged in molten solder for a precise delay period. Speed of withdrawal is a major process variable for controlling the coating thickness of solder.

Conformal Protective Coating. Some electrical and mechanical components are dipped in liquid polymers to seal out moisture, air, and contamination. The viscosity of the polymer and the speed of insertion into the fluid must be controlled so that flow into small features occurs without trapping bubbles of air. Once submerged the component may be reoriented to several poses and to allow air bubbles to escape.

Quenching. Heat treating is a commonly used method of improving alloy properties. Various fluids are used as cooling baths. Controlling insertion and manipulation is important for control of cooling rate.

Dipping processes require precision of part insertion and withdrawal so velocity and acceleration must be programmable and repeatable. The stirring requirements may require the use of a two- or three-axis wrist in addition to the translation motion axes. Grippers may require special cleaning or cooling capability either on board or at service stations located conveniently in the robot's reach.

Resistance Spot Welding

Robotic spot welding (see Figures 14.11.1 and 14.11.2) is the most pervasive robot application in the automotive industry. Resistance spot welds are formed by tightly clamping steel pieces together with opposing contact electrodes and then passing a large amount of current through the joint, welding the metal while producing a spray of molten sparks along with loud noise. Then the joint is held momentarily until the weld solidifies. Welds are made at discrete positions by moving the robot-mounted gun to pretaught poses. The spot welding process parameters, pressure and temperature, are controlled with the separate gun controller. Weld location and therefore positioning of the gun are critical.

Dexterity, payload, and quickness are critical operational requirements for spot welding robots. Gun pose repeatability is critical for consistently locating weld joints. Access to joint locations is limited because both electrodes must reach the weld site while maintaining clearance between gun frame and workpiece edges. Large articulated arm robots are typically used for most spot welding applications because of the dexterity needed and because the weight of the welding gun and associated robot-mounted

FIGURE 14.11.1 Six-axis articulated arm robots spot weld automobile bodies on a transfer line. (Courtesy of Nachi Robotics, Ltd.)

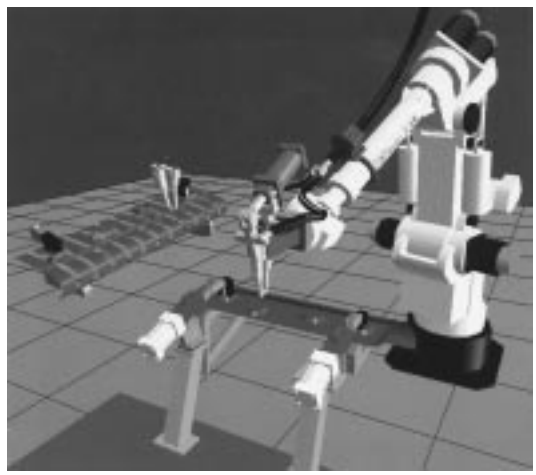


FIGURE 14.11.2 Spot welding operation simulated in off-line programming environment. (Courtesy of Deneb Robotics, Inc.)

apparatus often exceeds 200 lb. Fixed cycle programs are typical which may require several man-months to develop and less than a minute to execute. The robot spot welding path position names, path order, and control logic can be developed off-line, but lack of robot positioning accuracy characteristic of large articulated robots requires the weld positions for each individual robot to be taught by posing and recording them manually. This takes advantage of the robot's repeatability which is often orders of magnitude better than its accuracy. Unfortunately, when a robot that has been teach programmed is replaced by another robot, even an identical model, hours or days of teaching will be required to bring the replacement robot on line. Practical new PC-based calibration methods which eliminate this problem by effectively improving accuracy are now becoming commercially available.

Drilling

Hole drilling is a precision machining process. Most robots cannot hold a drill spindle rigidly enough to overcome the drilling reactions and most robots cannot generally move in a precise enough straight line to feed the drill. Drilling robots use special drilling end effectors which locate and dock onto the work piece or a fixture. The robot wrist and arm must be compliant and forceful enough to hold the drilling end effector firmly into location against the fixture or workpiece. Drilling end effectors have a spindle motor and a feed mechanism which execute a separately controlled drilling cycle while the robot holds the end effector in position.

The robot's only contribution to the process is to move the drilling end effector into its docking or holding place. Drilling robots have been used most successfully in the aerospace industry because airframe structures require thousands of holes to be placed precisely and in complex orientations. Manipulability requirements for drilling are similar to those for spot welding. The drilling end effector weight will tend to be less than a welding gun but tool holding force and reach usually impose the requirement for large robots.

Fastening

Robots are commonly used for applying threaded fasteners in the automobile industry for fastening wheels, and in the electronics industry for screwing components to circuit boards and circuit boards into chassis. Robots are also used for riveting in airframe fabrication.

Fastening is an end effector position-and-hold application. The robot does not follow the threaded fastener as it turns and travels into place; the end effector uses a slide or cylinder for that purpose. Automatic nut runners and screwdrivers and the associated hardware feeding apparatus are broadly available. Since a human is no longer operating the fastening tool other means of process control are needed. Usually fastener angular displacement, longitudinal displacement, and torque can be monitored and correlated with signatures or patterns characterized for specific fastener joints. Manipulator arm and control system requirements are similar to other position-and-hold applications. Very large torque may be encountered. Torsion bars or other static mechanisms are often needed to prevent the arm from being torque loaded.

Inspection

Robot inspection involves relative part/sensor manipulation to compare, measure, or detect a physical characteristic of the objective workpiece. Sensors used in robotic inspection include chemical detectors, computer vision systems, infrared detectors, sonar, laser radar, radiation detectors, capacitive proximity sensors, touch probes, X-ray cameras, particle/photon detectors, thread probes, and go-no go gauges. Robot inspection applications cover the range of manipulation from end effector position-and-hold to continues in-contact path motion. In some cases the kinematic structure of the robot is used as a spatial measuring device by incorporating surface sensors or probes in the last link as robot end effectors (Figure 14.11.3). The forward kinematic solution of the joint angle measurements sampled at a contact pose give the position in Cartesian space of the contact point. If the manipulator is stationary during the measurement then the robot's Cartesian positioning error must be added. If the robot is calibrated the error may be almost as small as the repeatability (0.001 to 0.020 in. for most industrial servo-driven

arms). If the arm is moving while measurements are made significant error may be added because of delay in sampling the manipulator joint positions.

FIGURE 14.11.3 Robots inspect pick-up truck body prior to final assembly. (Courtesy of Fanuc Robotics, N.A.)

Programming considerations are critical because robot inspection often requires data collection at a huge number of discrete positions. When CAD data are available, off-line programming of inspection may be possible, particularly for position and sense-type inspections. Sensor tool pose requirements can be quickly and accurately defined in the CAD environment and the pose data transformed into robot workspace coordinates. When hundreds or thousands of inspection poses are required manual teach programming may be too time consuming and cost prohibitive, especially when a mix of different parts must be inspected.

Paint and Compound Spraying

Paint spraying is a major application in the automotive industry. Painting booths are hazardous because the paint material is often toxic, carcinogenic, and flammable. Human painters often wear required protective clothing and breathing equipment. The paint fan projecting from the arm-mounted spray gun must be manipulated smoothly along paths that are often curved and complex.

Most robots used for painting are especially designed for that purpose. They usually have large reach, small payloads, and repeatability is usually larger than that of other types of robots and may exceed ±0.010 in. Painting robots are typically six DOF articulated arms, often with supplemental axes to pitch the paint gun and to traverse alongside a moving line. The potential for ignition of solvents and suspended particles may require taking precautions to eliminate ignition sources associated with the sparking of motors and other electrical components. Until brushless DC motors became commonly available for robot actuation virtually all painting robots were hydraulic because of the motor sparking problem. Lead-through teaching (also called teach-playback) is typical for painting. Off-line programming of painting is becoming more popular and some special software packages are available. Figure 14.11.4 shows an image from a simulation used in validating painting robot motion programs. Some compound spraying is done with smaller general-purpose robots, for example, spraying of protective coatings in the electronics industry. Circuit boards may require unexpected dexterity in order to point the spraying nozzle correctly to coat board features.

Compound Dispensing

Compound dispensing refers to laying a bead of fluid material on a surface. Application examples include caulking car bodies, sealing windshields, placing solder masking on circuit boards, gluing subassemblies,

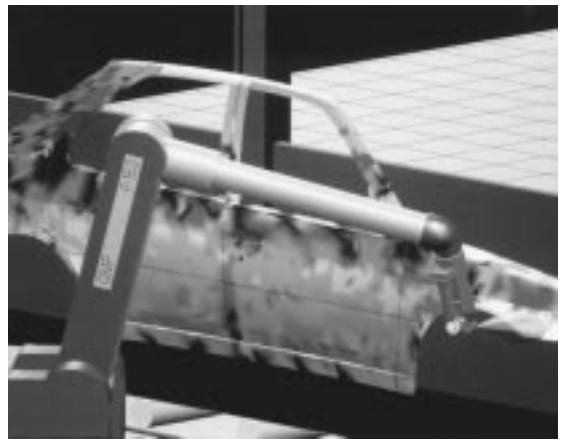


FIGURE 14.11.4 Painting operation simulated in off-line programming environment. Estimated paint thickness is illustrated by shading. (Courtesy of SILMA, Inc.)

solder paste dispensing, and decorating candies and cakes. Precision of placement and amount is critical. Smooth controlled paths are often essential. Position accuracy and tool path velocity accuracy are both important requirements.

All types and configurations of robots are used for dispensing. Many applications require only three DOF. When obstructions must be maneuvered around to gain access to the dispense locations, five or six DOF are needed. Payloads are usually small. Dispense speed may be limited by either the robot's ability to track a path at high speeds or by the dynamics of the dispensing process. In automotive applications the fixed cycle mode of operation is common. A robot program to lay a bead of sealer along the edge of a windshield is a taught path requiring good dynamic path repeatability of the robot. In electronic circuit board fabrication and decorating cakes, each workpiece may have a different dispense pattern; teaching paths are not practical in this situation. Some method of off-line programming must be used.

Cutting

Many engineering materials are produced and supplied as stacked or rolled flat plates or sheets. Further fabrication can require forming and/or cutting these materials into precise shapes. Robots are frequently used to manipulate a variety of cutting tools along paths that are often complex and curved. Many cutting processes are also used to produce fine features such as holes and slots. Common robotically manipulated cutting processes are listed below.

Laser. Molten metal heated by collimated intense light is blown away by a gas jet. Most common use is for thinner metals (0.50 in. or less) and on a variety of other thin materials.

Waterjet. A high velocity water jet is formed by forcing very high pressure water through a small orifice in the range of 0.008 to 0.040 in., which can cut a variety of nonmetals.

Abrasivejet. After a high velocity water jet is formed it passes through an abrasive mixing chamber where abrasive particle are entrained in the jet. A variety of metals and other tough and hard materials can be precisely cut; many materials can be cut with good control up to 1.0 in. thick. Cut thickness in excess of 6.0 in. has been reported.

Plasma Arc. Molten metal heated by an electric arc is blown away by a gas jet. Plasma arc cutting is commonly used to cut patterns in plate steel.

Router. A rotary cutter is most often piloted either on the workpiece or a guide fixture for precise trimming or chamfering edges of plate and sheet material.

Knife. A variety of knife types, some ultrasonically assisted, are employed to cut mostly nonmetals.

The cutting tool path and pose must be precisely controlled to achieve accurately patterned piece parts. The demands on manipulator performance are primarily determined by the interaction of desired part geometry, material thickness, and material properties. Feed rate, tool stand-off, beam, jet, and arc angle are all cutting process control variables which must be adjusted to material characteristics for good results. An extreme case of cutting manipulator performance demand is the combination of thin easily cut material with complex shape, and small geometric tolerance. This requires high speed coordinated motion of five or more axes which must be kept on track. This translates to a requirement for high performance servo-control elements in order to achieve high rates of mechanical response and joint angle position and velocity precision. Some type of advanced programming method such as a CAD/CAM may be required for high part mix applications. When extreme precision and complexity are required, as in many aerospace applications, precision fixtures incorporating tool guides may be used to force the tool path to repeat with near-zero deviation. In the case of contact tools such as routers and wheel knives the end effector must be capable of bearing preload forces applied in excess of the tool reaction forces to eliminate tool bounce-induced path errors.

Equipment and tooling for many of the robotic cutting processes may be complex and expensive. End effectors can easily cost tens of thousands of dollars and require difficult and cumbersome wiring and plumbing. Routing of laser wave guides and high pressure tubing for water jets from power source to robot end effector requires skill and experience in both design and installation.

The majority of cutting robots are three-axis natural Cartesian machines specially designed to cut flat sheet materials. Cutting speed and accuracy performance are aided by their easily calculated kinematics and easily predicted dynamics. Path planning and path generation are also simplified with flat parts. CNC is commonly used, and many automatic nesting and path programming software systems are readily available. There is some use of tool position sensors for part location during setup. In-process sensor-based tracking for edge cutting has been implemented with success, but is rare. Several five-axis gantry-style machines have been implemented for cutting complex aerospace materials including impregnated broadcloth patterns and composite wing skins. Articulated arm robots may be used when less precision is needed, as in trimming automobile carpet or making cut-outs in large plastic moldings.

Arc Welding

Arc welding is a metal joining process that uses intense heat produced by an electric arc between an electrode and the metal parts being welded. The weld pool and arc are always shielded by inert gas or a chemical vapor. In gas metal arc welding (sometimes called metal inert gas welding), which is the most common robotic arc welding, an electrode of filler metal wire is fed through a gun into the weld pool site as the robot manipulates the gun along the weld path. The hazards of arc welding include: intense ultraviolet, visual band and radio frequency radiation, toxic fumes, and noise. The pose (position and orientation) of the welding gun with respect to the joint or seam is a major arc control parameter. The feed rate of the gun is important in control of penetration and other weld characteristics. Unlike spot welding, manipulation is an arc welding process control variable.

Most arc welding robots operate in fixed cycle mode, which means they execute or play back a programmed sequence. If assemblies are presented to the robot with consistent seam geometry, then the path can be taught once, stored, and then executed repeatedly for each assembly. Given that all other relevant process variation is within accepted limits, the system will produce satisfactory output. However, weld seam position and seam shape variations may influence the process, especially in larger assemblies. When there is variation in the upstream sizing, cutting, fit-up, and jigging of weld assemblies, the location and orientation of the weld seam will vary. Also, as the weld progresses, localized thermal expansion and residual stresses can force seam distortion. A range of methods for adapting the robot system to these variations exists, from correcting pretaught path plans at setup time, through actively altering robot motion in “real time.” A sustained high level of academic and commercial research and development effort has resulted in practical methods of automatic weld seam tracking and process control.

Rapid deployment in recent years is a direct result of sensor integration for seam tracking Seam tracking methods correct the path to compensate for errors in location and orientation of the welding tip based on sensor data. Commonly used sensors include mechanical probes, computer vision, laser edge detection and ranging, and arc current and voltage. Because of the extreme environment of the region surrounding an active welding tip, sensors are often housed in protective chambers. Typically the errors are measured and calculated in a convenient reference frame in the three-dimensional workspace of the robot system, the same space in which the tool path is described. In some cases the error is measured by tool-mounted sensors relative to the moving reference frame of the tool. The preprogrammed path is then shifted by mathematical transform in the reference frame of the tool. An important aspect of seam tracking is the use of sensors to detect the sides of the weld channels as boundaries for automatic side-to-side weaving. This is usually done with "through-the-arc" sensing in which arc current is monitored as an indicator of clearance between the welding tip and the channel edge. While tracking in the direction of the seam, transverse motion commands are given so that the tip approaches one edge until the edge is sensed and then the motion is commanded in the direction of the other edge. Weld penetration, filler deposit amount, and weld bead shape can be controlled in-process by variable control of the welding speed or feed rate. Arc welding robot systems which use sensors to adjust the robot path in real time (computation is fast enough to respond to sensor data with useful path adjustments) are among the most advanced or intelligent robotic applications found in practical industrial use.

Robots used for arc welding must be capable of precisely executing taught paths. Motion must be smooth and precisely controlled. Velocity control is important but the speeds required are not high, 2 in./sec, while welding is faster than most applications require. Higher velocities are important to reduce cycle time for applications with lengthy arc-off motion. Welding robots must have good reach and dexterity. Five DOF is required as a minimum and six DOF adds to gun maneuverability. There is normally no forced contact with the weld seam and the welding gun's weight is usually less than 20 lb, so robot payload requirements are light. If real-time path altering is required, the robot's motion generation functions must have programmable interfaces with the sensor systems. The robot's controller must have a means of accepting data and manipulating it for use with high level functions in the robot's native programming language.

Finish Machining

Few material-forming processes produce finished parts. Most machining operations leave burrs and sharp edges. Large aircraft wing skins are milled by three-axis terrace cutting leaving small steps which must be blended to prevent fatiguing stress concentrations. Complex curved surfaces like ship propellers and aircraft landing gear are machined with rounded milling tools which leave a pattern of tool marks which must be ground off. Cast parts require gate and sprue removal and deflashing. Many parts must have their surfaces conditioned for appearance or subsequent plating and coating operations. Stamping and forging of automobile door panels and engine components leave "imperfections" which are finished out by hand. Die cast surfaces of hardware for door handles, faucets, furniture, and appliances are ground and polished. Finishing removes material to reduce waviness, reduce roughness, remove burrs and sharp edges, and to remove flaws. Manipulation is a finish machining process control variable. Tool pose, applied pressure, feed rate, and tool path must be controlled. Smaller parts are finished using fixed-floor or bench-mounted tool stands. For larger work pieces the finishing tool is mounted on the robot. Finish machining is very demanding of the manipulator because continuous path control is required while maintaining contact between part and tool.

Medium- to large-sized robots are usually required for surface finishing because end effector weight and tool reaction force are additive when calculating payloads. A further margin of payload is usually required to offset the fatiguing effects of vibration and cyclic loading from tool reactions. Tool point positioning accuracy is less important than tool orientation and feed rate. In general, higher rates of surface curvature will require greater robot path precision. Six DOF robots are often required. Edge machining tool paths are constrained by burr geometry, finishing tool characteristics, end effector geometry, and part geometry. A single part may have edge features in several directions and orientations.

The robot may require an assortment of different tools and a tool changer to reach all edges. If these measures do not allow access then multiple setups may be needed to present all features for finishing.

Most robotic surface finish machining applications use compliant abrasive processes. Force control is required in some applications to keep tool pressure constant. Force controllers are most often incorporated in the end effector or in the tool stand. Figure 14.11.5 shows a robot equipped with a force-controlled finishing end effector using a servo-controlled pneumatic actuator that is capable of applying consistent tool pressure. Through-the-arm robot force-control is available from some robot manufacturers, but its usefulness is limited to applications requiring slow feed rates because of slow mechanical response.

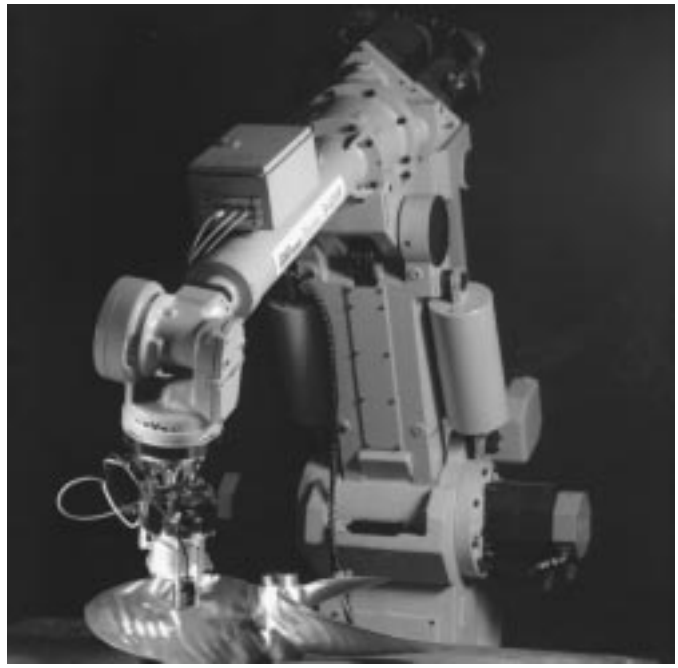


FIGURE 14.11.5 Robot equipped with force-controlled end effector grinds tool marks from ship propellor. (Courtesy of The Automation and Robotics Research Institute, The University of Texas at Arlington.)

Path planning and programming of edge and surface finishing for complex-shaped parts can be very difficult and time consuming. Both tool position and tool pose are critical in obtaining the correct tool contact area. Tedious paths programmed using the teach method require hundreds of hours to develop because of the large number of taught points. Generating the path control sequence is a major problem in manufacturing operations which produce a variety of complex-shaped parts. An example is in polishing large asymmetric-shaped aircraft skin panels. A more difficult automation problem is robotic spot finishing of flaws and other anomalous regions of the part surface when their location and extent are not known before set-up time. The reason is that the paths must be planned, generated, and executed on-line. This type of motion generation system requires part modeling and computational functions not available on most robot controllers. Figure 14.11.6 shows an operator digitizing a part surface to be modeled in preparation for automatic path generation by the PC.

FIGURE 14.11.6 Workpiece surface is digitized and modeled in preparation for automatic tool path generation. (Courtesy of The Automation and Robotics Research Institute, The University of Texas at Arlington.)

14.12 Mobile, Flexible-Link, and Parallel-Link Robots

Kai Liu

This section will discuss nonstandard robots, including mobile robots, lightweight flexible-link robots, and parallel-link robots. These robots are often more suitable than standard serial-link commercial robots for certain applications.

Mobile Robots

Traditionally, standard robots are fixed in position. They are mounted on a rigid base and bolted to the floor so that they can withstand the forces and torques applied when the arm manipulates objects. However, fixed-base robots cannot cope with a large variety of applications in which a robot will operate in large and unstructured domains. A special type of manipulator, that is, a mobile robot, is often required in these applications.

In tomorrow's flexible manufacturing system (FMS) environment, mobile robots will play an important role. They will transport parts from one workstation to others, load and unload parts, remove undesired objects from floors, and so on. In addition to indoor mobile robots, there are some other outdoor occasions where mobile robots may take on heavy responsibilities. Examples include construction automation, military missions, handling of harmful materials, hazardous environments, interplanetary exploration, and so on.

Classifications of Mobile Robots

Mobile robots can be classified by driving mechanism as wheeled mobile robots, legged mobile robots, and treaded mobile robots. Some other types of mobile robots, for instance, the *underwater mobile robots*, the *autonomous aerial mobile vehicle*, and so on, are also available but are not included in this discussion.

Wheeled Mobile Robots. Mobile robots using wheels for locomotion are called *wheeled robots*. Two driving configurations are used in today's wheeled mobile robot — steer-drive and differential-drive. The former uses two driving wheels to make the vehicle move forward and backward. The heading angle is controlled by an independent steering mechanism. Since the driving action is independent of the steering action, the motion control of the vehicle is somewhat simplified. However, due to physical constraints, this configuration cannot turn in a very small radius. This shortcoming makes it less attractive in some industrial applications. Differential-drive configuration mobile robots, on the other hand, have two independent driving wheels positioned at opposite sides of a cart base, arranged parallel to one another. Their speeds can be controlled separately. Thus, by appropriately controlling the speed of each driving wheel, this mechanism is able to drive the vehicle forward and backward, as well as steer its heading angle by differential speed commands. Even though this configuration requires a somewhat more complex control strategy than the steer-drive configuration, its capability of making small-radius turns, even making turns on-the-spot, makes it the first choice in many industrial applications.

Some commercial wheeled mobile robots include the *CyberGuard Autonomous Survelliance Robot* manufactured by Cyberworks Inc., Canada; *B12 Mobile Robot Base* manufactured by Real World Interface, Inc., Dublin, NH; *LabMate Mobile Robot Platform* manufactured by Transitions Research Corporation, Danbury, CT; and *R-20 Mobile Robot* manufactured by Arrick Robotics, Euless, TX.

Legged Mobile Robots. While most mobile robots use wheels for locomotion because of the simplicity of the moving mechanism design and control, some other mobile robots use legs for locomotion. These types of mobile robots are called *legged robots.* The primary advantages of legged robots include their ability to traverse rough terrain with good body stability and minimal ecological damage. In order to maintain good stability, it is sufficient that at any time there are three points in contact with the ground. Therefore, most legged robots use at least four legs, or even six or eight legs. As long as the legged mobile robots' center of gravity is within the triangle formed by the three contact points, stability is

guaranteed. Compared with the wheeled robots, the control of legged robots is much more difficult. Much has been learned about multilegged locomotion from studies of balancing and hopping on a single leg. In particular, biped running can be viewed as successive hopping on alternating legs, since both legs never contact the ground simultaneously. Some examples of legged mobile robots include *ODEX I* manufactured by Odetics.

Treaded Mobile Robots. Another type of mobile robot, the treaded robot, moves much like a tank. An example of the treaded robot is the *ANDROS MARK V* manufactured by REMOTEC, Inc. at Oak Ridge, TN. It is something of a hybrid between a walking and a rolling vehicle. *ANDROS* can ascend/descend 45° stair/slopes by lowering its front and rear auxiliary tracks. It has all-terrain capabilities that are ideal for performing missions in rough outside terrain or in rubble-strewn, damaged buildings.

Sensors and Measurements

To navigate in unknown and unstructured areas, the mobile robot must have the capability of sensing the real world, extracting any useful information from the data acquired, and interpreting the information to understand the environment surrounding it, especially the situation in front of it. Several sensor systems for mobile robot navigation have been reported in the literature (Elfes, 1987). Of these, stereo vision systems and active rangefinding devices are the most used sensor systems. The former extracts range information from pairs of images to build a 3D world map. However, due to the high computational expense — a 3D map may require 1 min to generate — stereo vision systems have not to date been generally used for real-time navigation control. Active rangefinding devices do not suffer from this problem because they can deliver range information directly

Two kinds of rangefinding devices are available: laser rangefinders and ultrasonic range transducers. Even though laser rangefinders can provide fast response with high resolution, a relatively long measurement range, and high measurement precision, the required systems structure and configurations are very complicated, which makes the system itself very expensive. On the other hand, sonar systems are simple and low cost (probably orders of magnitude less expensive than laser-based systems), though the measurements have lower resolution and lower precision.

Determining range by means of sonar systems is a simple process. A short burst of ultrasonic sound is first transmitted by an ultrasonic range transducer, then an echo is expected to be received by the same transducer. If in a reasonable time period no reflected signal is detected, it is assumed that there are no objects in the area of interest. Otherwise, the time for round-trip propagation is determined and the distances to any objects are calculated. The transducer yields a 3-dB full angle beamwidth of 50 KHz at approximately 12 to 15°, depending on the signal frequency and transducer diameter. Thus, to scan the whole area surrounding the mobile robot, at least 24 to 30 transducers, of which the transmit/receive axis lies in the same horizontal plane, are needed.

Vision systems, also sometimes useful in robot sensing, usually consist of one or more video cameras and an image processor. The vision system can provide the richest source of information, which is, in fact, needed in certain applications such as road following, object identification, and so on.

Feedback from rotary and linear actuators used in wheeled and/or legged mobile robots is provided by position sensors and/or velocity sensors. This information is then processed for estimating position and orientation of the mobile robot in world coordinates.

By far the most commonly used position sensor is the *optical encoder*, which uses marks to indicate position. The typical encoder has a track for each binary digit of information. The encoder is mounted on the servo motor. When the motor rotates certain degrees, the absolute rotation position of the axis can be read from the digital output of the encoder. The resolution of the encoder is equal to $(360/2^n)$ degree, where n is the number of tracks. If an 8-track encoder is used, then a 1.4°/step resolution can be attained.

Other types of position sensors used in mobile robot systems include synchros, resolvers, potentiometers, linear variable differential transformers (LVDT), rotary variable differential transformers (RVDT), amplitude-modulated laser radars, and laser interferometers.

Conventional servo design requires that the servo controller include a "velocity term" in its transfer function. Without the velocity term, a servo system will usually exhibit an undamped, resonant behavior and can be highly unstable. In principle, the signal from a joint position sensor can be electronically differentiated to obtain joint velocity. However, if the joint position sensor has a noisy output, differentiating the position sensor signal can effectively magnify the noise sufficiently to make the servo system unstable or unreliable. To overcome this difficulty, several velocity sensors are available for directly measuring the joint velocity. A *DC tachometer* system consists of a voltage meter (or a current meter) and a small DC generator (sometimes called a "speed-measurement generator"). The latter is usually constructed with a permanent-magnet stator and a multipole wound armature. The armature is connected directly to the rotating shaft of the servo motor which is used to drive the manipulator joint. When the small permanent magnet DC generator rotates with the servo motor, its output voltage (when driving a high-impedance load) varies in proportion to the rotation speed of the armature. Voltage output variations can then be translated into speed changes or used as a feedback signal to control the robot arm velocity.

Supplementary position and orientation information can also be supplied by inertial guidance sensors, terrestrial magnetic field sensors, or inertial reference systems (IRS).

Navigation

Autonomous navigation of mobile vehicles has been studied by many researchers. In Elfes (1987), a sonar-based navigation system for an autonomous mobile robot working in unknown and unstructured environments was developed. The workspace is classified into "probably empty regions," "somewhere occupied regions," and "unknown regions" based on the interpretation of the data obtained from the sonar system. In this scheme, as more and more data are received, the first two regions may increase, and the uncertainty of these regions also decreases. It is reported that after a few hundred readings, a sonar map covering a thousand square feet with up to 0.1-ft position accuracy can be made. Another navigation scheme uses a stereo vision system to control a mobile base autonomously operating in a complex, dynamical, and previously unknown environment. A pair of stereo cameras is mounted on the mobile base to generate a symbolic world model. Based on this model, the desired trajectories are specified for the driving motors.

Although the schemes described above work well in specific environments, path planning and navigation control are always separated into two isolated issues. The path planning mechanism designs a smooth path from an initial position to a goal position by providing profiles of position and velocity, or profiles of position and heading angles, in Cartesian space. It assumes that perfect knowledge of the system dynamics and the environments is always available and that the position and orientation of the vehicle are measurable absolutely. After the desired trajectories have been designed, the navigation mechanism will take charge of driving the mobile robot to follow the prescribed trajectory as closely as possible. Even though each mechanism may work well through closed-loop control, the whole navigation system is an open-loop system. Static path planning strategies do not provide the essential adaptability necessary for coping with unexpected events. The success of navigation control depends mostly on the accuracy of absolute measurements of position, velocity, orientation, and their rates of change. All of these must be measured in (or transformed to) Cartesian space. This is a very expensive and difficult job.

Other possible closed-loop navigation control schemes use intelligent control techniques, for instance, fuzzy-logic control. In such a control scheme, the path-planning mechanism and trajectory-following mechanism are often integrated, not separated. The path is planned dynamically and is always up-to-date. All the information that the system needs to know such as "where is the goal (the dock)," "what is the required final orientation (the docking angle)," "what is the present orientation (the present heading angle)," "what is the present distance between the car and the goal," "what is the present distance between the car and any obstacles," "what is the safe turning radius (the minimum curvature radius)," and so on is easily captured through sensing the environment surrounding the car using onboard sensors (e.g., sonar) that yield relative information.

The advantages of such intelligent control strategies are evident. They unite navigation and maneuvering into a single set of algorithms. Full and accurate knowledge of the system dynamics is not required. The only knowledge needed are the correlations between the control actions (acceleration, steering, etc.) and the performance ("behaviors") of the system. The absolute measurement of the position and velocity in Cartesian space is not required. Only information about relative locations is necessary, and this is always available. Tight coupling between sensor data and control actions provides the adaptability necessary for coping with unexpected events. Actually, there is no path planning to be performed; the driving mechanism reacts immediately to perceived sensor data as the mobile robot navigates through the world.

Flexible-Link Robot Manipulators

Most robots used in today's manufacturing systems are rigid-link manipulators. Making the robot links and drives extremely stiff to minimize vibrations allows rigid-link robots to track a desired trajectory with very high degree of accuracy, often using standard classical (PID) control schemes. However, the price paid for this includes heavy manipulators, a low payload-weight-to-arm-weight ratio, high power consumption, and slow response rates to motion control commands.

With the growing demand from industry automation for lower manufacturing costs, higher motion speeds, better performance, and easier transportation and setup, the rigid-link manipulators may, sooner or later, be replaced by some sort of lightweight mechanical structures, such as flexible-link robots. While lightweight flexible manipulators have certain inherent advantages over rigid-link robots, they impose more stringent requirements on system modeling and controller design because of the vibrations of the flexible modes.

Modeling of Flexible-Link Robots

One essential step toward successful control synthesis is to obtain an accurate dynamic model for flexible-link manipulators. The flexible manipulator dynamics can be derived on the basis of a recursive Lagrangian assumed-modes method (Book, 1984). The motion of robots with link flexibility is governed by partial differential equations that must be satisfied inside a given domain defining the flexible structure, and by boundary conditions to be satisfied at points bounding this domain. Therefore, the dynamic model of flexible-link manipulators consists of highly coupled nonlinear integro-partial differential equations. Control of a structure using a formulation based on partial differential equations is extremely difficult. To reduce the complexity of the model, the assumed-modes method is used to produce a set of nonlinear ordinary differential equations based on an orthonormal series expansion of the flexure variables.

A solution to the flexible motion of links is obtained through a truncated modal approximation, under the assumption of small deflections of the links. The dynamic equations of motion for an *n*-degree-of-freedom manipulator with up to *m* flexible links can be written as

$$\mathbf{M}(\mathbf{q},\delta)\begin{bmatrix}\ddot{\mathbf{q}}\\ \ddot{\delta}\end{bmatrix}+\mathbf{D}\left(\mathbf{q},\dot{\mathbf{q}},\delta,\dot{\delta}\right)\begin{bmatrix}\dot{\mathbf{q}}\\ \dot{\delta}\end{bmatrix}+\begin{bmatrix}0 & 0\\ 0 & \mathbf{K}_f\end{bmatrix}\begin{bmatrix}\mathbf{q}\\ \delta\end{bmatrix}+\begin{bmatrix}\mathbf{F}_r(\mathbf{q},\dot{\mathbf{q}})\\ 0\end{bmatrix}+\begin{bmatrix}\mathbf{G}_r(\mathbf{q})\\ 0\end{bmatrix}=\begin{bmatrix}I\\ \mathbf{B}_f\end{bmatrix}\tau$$

where $\mathbf{q} = [q_1\ q_2\ \ldots\ q_n]^T$ is the vector of rigid joint variables, $\delta = [\delta_1\ \delta_2\ \ldots\ \delta_m]^T$ is the vector of deflection variables, $\mathbf{M}(\mathbf{q}, \delta) \in R^{(n+m)\times(n+m)}$ is the inertia matrix, $\mathbf{D}(\mathbf{q},\dot{\mathbf{q}},\delta,\dot{\delta}) \in R^{(n+m)\times(n+m)}$ contains both rigid and flexible coriolis/centripetal terms and terms representing the interactions of the joint rigid variables with the deflections, $\mathbf{K}_f \in R^{m\times m}$ is the stiffness matrix, $\mathbf{F}_r(\mathbf{q},\dot{\mathbf{q}}) \in R^n$ is the friction, and $\mathbf{G}_r(\mathbf{q}) \in R^n$ is the gravity term. The control input is $\tau \in R^n$ and the input matrix $[\mathbf{I}\ \mathbf{B}_f]^T \in R^{(n+m)\times n}$ is generally a function of $(\mathbf{q}, \delta)$ depending on the boundary conditions (e.g., pinned-pinned, pinned-free, clamped-free) chosen by the designer. Note that the coriolis/centripetal matrix $\mathbf{D}(\mathbf{q},\dot{\mathbf{q}},\delta,\dot{\delta})$ can take several different definitions. However, among these definitions, there exists one such that the derivative of the inertia matrix $\mathbf{M}(\mathbf{q}, \delta)$ and the coriolis/centripetal matrix are related in a very particular way, that is, $\dot{\mathbf{M}}(\mathbf{q},\delta) - 2\,\mathbf{D}(\mathbf{q},\dot{\mathbf{q}},\delta,\delta)$ is *skew symmetric*, a property that is often very useful for controls design.

It is important to realize that the rank of the control effectiveness matrix rank ($[\mathbf{I}\ \mathbf{B}_f]^T$) *is* less than ($n + m$), that is, *rank* ($[\mathbf{I}\ \mathbf{B}_f]^T = n < (n + m)$. This means that the number of degrees of freedom is greater than the number of control inputs and is the major source of the problems in controlling flexible-link manipulators.

Control of Flexible-Link Robots

Control of mechanical manipulators to maintain accurate position and velocity is an important problem. Rigid-link manipulators are designed to be mechanically stiff precisely because of the difficulty of controlling flexible members. The major objectives in control of a robotic system with link flexibility are to command the tip of the flexible-link manipulator to move from one position to another as quickly as possible (point-to-point minimum-time control), or to follow preplanned desired trajectories (trajectory-following control) while keeping the oscillations of the flexible modes as small as possible. The inherent large nonlinearities of flexible-link manipulators make their control very difficult. The link flexibility makes the robot arm itself sensitive to external excitation; a small impulse signal may cause the flexible modes to oscillate wildly.

Several conventional control techniques have been studied by robotics researchers for the control of flexible-link manipulators. They fall into different categories. The first category includes some approximate techniques such as linear systems approaches, linear minimum-time control, decentralized approaches, and input-shaping techniques. Conventional control techniques cannot usually obtain very satisfactory results for fast desired motions.

The second category includes some new control approaches that take into account many of the nonlinearities. Among them are variable structure control, adaptive control, and the inverse dynamics approach (where the whole control signal is composed of a causal part and an anticausal part). To minimize residual vibrations, several constraints must be applied for the desired tip trajectories. As pointed out in Kwon and Book (1990), for a specified rigid mode trajectory, there is associated a unique flexible mode trajectory. The interactions between the desired rigid motion and the associated required flexible motion (e.g., the inverse dynamics) are very complicated and parametrically sensitive.

Since the number of independent control inputs is less than the number of the output variables in the case of flexible-link arms, the control problem is characterized as having *reduced control effectiveness.* The so-called "model matching conditions" do not hold, and the conventional control techniques usually used in the control of rigid-link robots (e.g., computed-torque control) cannot be directly applied to the control of flexible-link arm. This problem can be solved by a model-order reduction based on a singular perturbation strategy, in which the rigid modes are treated as slow-state variables, while the flexible modes and their time derivatives are treated as fast-state variables. Another approach is the "reduced-order computed torque scheme" that first removes the nonlinearities that are in the range of the control input matrix, then uses PD state-feedback loop to convert the flexible system to a set of uncoupled point-mass-like systems.

A third category of controllers includes various intelligent control schemes such as neural networks (NN) (Lewis et al., 1995) and fuzzy logic control (FLC). Many of the drawbacks mentioned above can be overcome by either FLC or NN control if these are used in conjunction with sound control engineering design practices (e.g., singular perturbation and/or feedback linearization techniques). The reason is obvious: FLC and NN are *model-free* control schemes applicable to a wide range of dynamical systems that are ill understood and ill defined. The primary reason for this model-free characteristic is the 'universal approximation property, shared by NN and FLC. Thus, by careful design, effective control actions can be generated without extended analyses based on a precise, explicit mathematical function.

Parallel-Link Robots

By far the most widely used commercial robots are the serial-link manipulators, whose links and joints alternate with one another in an open kinematic chain. This serially connected configuration is similar to that of the human arm, with each link connecting only to two neighboring links through either prismatic

or revolute joints, except for the last link which attaches to the end effector and the robot base which attaches to the floor. The advantage of the serial chain structural arrangement is that it provides a large work volume and dexterous manipulability; however, it suffers from a lack of rigidity and from accumulated actuator errors. Especially at high speed and high dynamic loading operating conditions, the serial-link manipulators show poor dynamic performance. To improve the dynamic performance and achieve high precision operations, the robot links must be made with high rigidity, which results in heavy robots with low force-output-to-manipulator-weight ratio. On the other hand, if the links can be arranged parallel to one another in a closed kinematic chain structure such that the major force components add together, then high precision operations and high force-output-to-manipulator-weight ratios can be achieved.

The Stewart Platform

The most popular and successful parallel mechanical structure is the so-called Stewart platform, which was first proposed by Stewart (1965) in 1965. As a manufacturing manipulator, the Stewart platform has two fundamental characteristics which set it apart from machine tools and industrial robots — it is a *closed kinematic system* with *parallel links.* The Stewart platform link ends are simply supported, making the manipulator system far more rigid in proportion to size and weight than any serial link robot. Furthermore, the links of the Stewart platform are arranged so that the major force components of the six actuators add together, yielding a force-output-to-manipulator-weight ratio more than one order of magnitude greater than most industrial robots.

The original Stewart platform was designed for an aircraft simulator and consisted of six linear hydraulic actuators acting in parallel between the base and the upper platform, as shown in Figure 14.12.1. All the links are connected both at the base and at the upper platform. Thus, by changing the length of each link, the position and orientation of the upper platform are able to be controlled.

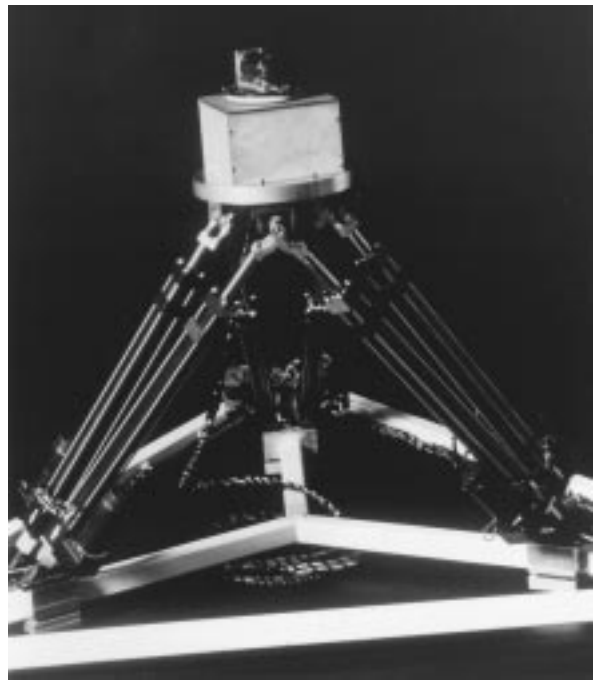


FIGURE 14.12.1 A six-degree-of-freedom *Stewart platform manipulator* developed at the Automation & Robotics Research Institute, The University of Texas at Arlington.

Some nice features of the original Stewart platform include:

- The manipulator design has six degrees of freedom, three for position and three for orientation.
- The six linear actuators are driven by six motors with each motor reacting on the base to avoid interaction between motors. Actually the manipulator can move when five of the jacks are locked merely by adjustment of the remaining jack.
- To achieve the maximum performance for a given power source, each motor operates directly on the same load (the upper platform). This makes a high payload-to-structure-weight ratio that at certain points of the workspace amounts to nearly six times the lifting capability of each individual actuator.
- Having low friction losses, with a powerful hydraulic system the manipulator can respond to commands very quickly.

It is interesting to note that the Stewart platform was not the first parallel link mechanical structure used in industry. As early as the 1950s, McGough devised a similar device for studying tire-to-ground forces and movements. The system had been in operation since 1955, but was never made known to the public until 1965 when Stewart published his journal article.

Advantages and Problems of the Stewart Platform

The Stewart platform appears simple and refined to the point of elegance. The performance mentioned above can be achieved using relatively inexpensive commercially available servo-actuator technology. The Stewart platform uses a closed kinematic chain which is structurally extremely strong and rigid, and is capable of distributing loads throughout the system. The actuator errors are not cumulative, allowing for high precision operations. However, the same closed kinematic structure that provides mechanical stiffness also complicates the forward kinematics analysis. This problem is an impediment to the derivation of dynamic equations and hence control schemes for real-time trajectory generation, which is necessary for industrial application of the manipulator (e.g., surface finishing applications).

It is known that in the case of fully parallel structures, the inverse kinematics (that is, solving for the corresponding lengths of the links given the position and orientation of the upper platform in Cartesian space) is relatively straightforward. However, the forward kinematics analysis for the fully parallel mechanism (e.g., given the length of each link, solve for the position and orientation of the upper platform in Cartesian space) is very challenging. The reason is that the kinematic equations are highly coupled and highly nonlinear.

Much effort has been devoted to finding an efficient algorithm for computing an accurate kinematic solution. To solve for the Cartesian position of the upper platform in terms of the given link lengths, thirty (30) nonlinear algebraic equations must be solved simultaneously, or polynomials of order 16 in a single variable must be solved. Due to the time-consuming nature of these procedures, it is difficult to compute the kinematic solutions on-line in real time. In Liu et al. (1993), a simplified algorithm was proposed which required to solve for only three (3) simultaneous nonlinear algebraic equations. Since the Stewart platform requires complex kinematics computations for trajectory following control, it is difficult to achieve real-time control capable of supporting high bandwidth motion.

The common feature of the forward kinematics analyses mentioned above is that there is no explicit analytical solution. Even for Liu et al.'s algorithm, it is still required to solve three nonlinear algebraic equations by numerical methods. Since there is no explicit expression available for the forward kinematics, deriving the Jacobian matrix and dynamic equations directly in link space and studying the singularity become impossible.

It is known that the Jacobian provides a transformation path which allows a two-way transformation between the link space and Cartesian space. If the Jacobian is not singular, then velocity in link space can be uniquely transformed to the corresponding velocity in Cartesian space. Particularly, if there is no movement in link space, then there is no movement in Cartesian space, so that the Stewart platform will remain rigidly fixed. However, at singular configurations, the transformation path from link space

to Cartesian space is blocked. In this case, even though there is no movement in link space, the upper platform can lose rigidity, still possibly moving along some directions. In other words, at singularities, the Stewart platform may gain extra degrees of freedom. The problem becomes even worse in that, in this situation, forces or torques in Cartesian space cannot be transformed to link space, that is, at singular positions the Stewart platform cannot be controlled to move in all directions and cannot exert force in all directions. From the applications viewpoint, investigating the conditions under which there will be singularities is important.

Thus, while the parallel link manipulators afford structural advantages, they also present severe difflculties for controller design. The control problems associated with such structures are not easy, as the systems do not satisfy most of the assumptions made in the controls literature (e.g., linearity in the parameters and feedback linearizability). Therefore, most existing control algorithms do not work well.

Manufacturing Applications of the Stewart Platform

Since proposed by Stewart (1965) in 1965, various applications of the Stewart platform have been investigated for use as aircraft simulators, as robot wrists, in mechanized assembly, and in active vibration control. As a manufacturing manipulator, the Stewart platform has great potential in automating many light machining applications such as surface finishing, edge finishing, routing, and profile milling. New manipulator applications to manufacturing processes requiring high force and power output such as combined assembly pressing are also possible. There are several light machining applications that a Stewart platform manipulator would perform with less set-up complexity and tooling cost than a serial link robot or a standard machine tool. In Figure 14.12.2 is shown a Stewart platform developed as a surface finishing milling machine.

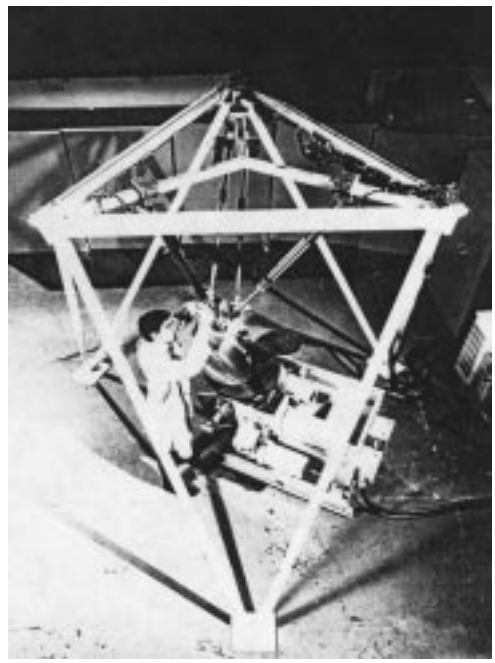


FIGURE 14.12.2 Stewart platform automated surface finishing cell.

Many applications of robotic manipulators to high precision routing can be found in the aerospace industry. A common application is trimming wing skin edges. The precision is usually achieved through the employment of expensive templates that provide a precise guide bearing for the tool to follow as it is held in the naturally compliant grip of a robotic manipulator. The major advantage for using a Stewart platform as a routing machine is that it can follow the contours of many aerospace parts without the

need for a tool guide; it is stiff enough to track the part precisely while withstanding the router cutting reactions. The cost of this contouring capability as compared to a standard five- or six-axis routing machine would be much lower.

The Stewart platform would be superior to any serial link robot as a drilling head manipulator. Virtually all applications of drilling robots in aerospace manufacturing require the use of expensive and complex end effectors of part jigs to compensate for the inaccuracy and lack of stiffness of the robots. Drilling jigs for some parts can cost as much as ten times more than the robot itself. Special end effectors are often used to apply preloads to prevent the drill from "walking" and chattering. The Stewart platform could perform many drilling tasks unaided by special tooling because of its stiffness and precision.

The industrial robot has generally not been considered to be a good milling manipulator. The Stewart platform could potentially perform contour milling of some materials with near-machine-tool accuracy. A Stewart platform milling machine with stiffness and dexterity characteristics intermediate between a large serial link robot and a five-axis mill could be built at or below the cost of a commercial serial link robot. A Stewart platform milling machine successfully used for industrial applications is shown in Figure 14.12.2.

When a direct contact tool like a grinder is used it is important to control both the tool position and the forces involved so that the substrate is not damaged. For example, when grinding mold scale a very aggressive tool may be needed, and the normal force applied can be as large as 40 to 50 lbs so long as the penetration into the surface is precisely controlled. The reactions in the surface tangent plane can be very large and could cause oscillations if not held rigidly. A Stewart platform with a constant force suspension for its tool could apply very large force with very high stiffness in one direction while being compliant and forceful in the normal direction.

Defining Terms

Accuracy. The degree to which the actual and commanded positions (of, e.g., a robot manipulator) correspond for computed as opposed to taught positions.

Adaptive Control. A large class of control algorithms where the controller has its own internal dynamics and so is capable of learning the unknown dynamics of the robot arm, thus improving performance over time.

AML. A Manufacturing Language — a robot programming language.

APT. Automatic Programming of Tools — a robot programming language.

Cell Decomposition. An approach to path planning where the obstacles are modeled as polygons and the free space is decomposed into cells such that a straight line path can be generated between any two points in a cell.

Compliance. The inverse of "stiffness" — useful in end effectors tooling whenever a robot must interact with rigid constraints in the environment.

Computed-Torque Control. An important and large class of robot arm controller algorithms that relies on subtracting out some or most of the dynamical nonlinearities using feedforward compensation terms including, e.g., gravity, friction, coriolis, and desired acceleration feedforward.

End Effector: Portion of robot (typically at end of chain of links) designed to contact world:

- **Compound.** A cluster of multiple end effectors and tooling mounted on the robot wrist.
- **Active.** An end effector with sensing and servo control of the grasp forces and/or finger motions.
- **Prehensile.** An end effector that holds parts between fingertips or encircled by fingers.
- **Vacuum.** A nonprehensile end effector that uses suction cups to hold parts.
- **Dextrous.** A hand with the ability to manipulate parts in the fingers and actively control grasp forces.

Feedback Linearization. A modern approach to robot arm control that formalizes computed-torque control mathematically, allowing formal proofs of stability and design of advanced algorithms using Lyapunov and other techniques.

Flexible-Link Robot. Lightweight mechanical structures where vibration and flexibility of the links must be taken into account in controller design. They possess favorable features including lower manufacturing costs, higher motion speeds, better performance, and easier transportation and setup.

Force Control. A class of algorithms allowing control over the force applied by a robot arm, often in a direction normal to a prescribed surface while the position trajectory is controlled in the plane of the surface.

Forward Kinematics. Identification of task coordinates given configuration.

Fuzzy Logic Control. A multilevel logic controller, which is different from the conventional dual (two-level) logic in which only two values (true and false) may be assigned to each state variable. Fuzzy logic controllers have advantages in being robust to disturbances and not requiring an explicit mathematical model for the design process. They consist of three parts: the fuzzifier, the rulebase, and the defuzzifier.

Grasp Isotropy. A measure of how uniformly forces and motions can be controlled in different directions.

IGES. International Graphics Exchange Specification — a data exchange standard.

Inverse Kinematics. Identification of possible configurations given task coordinates.

I/O Device. Input/output device — a port through which external information is connected to a computer. I/O devices may be A/D, which converts analog signals to digital, D/A, which converts digital signals to analog, or binary, which passes digital signals.

Joint Variables. Scalars specifying position of each joint — one for each degree of freedom.

Kitting. The process of taking parts from bulk and placing them on a *kit tray,* which is an organized group of all parts required to assemble a single product or subassembly.

Learning Control. A class of control algorithms for repetitive motion applications (e.g., spray painting) where information on the errors during one run is used to improve performance during the next run.

Linearity in the Parameters. A property of the robot arm dynamics, important in controller design, where the nonlinearities are linear in the unknown parameters such as unknown masses and friction coefficients.

Manipulator Jacobian. Matrix relating joint velocities to task coordinate velocities - configuration dependent.

Mechanical Part Feeders. Mechanical devices for feeding parts to a robot with a specified frequency and orientation. They are classified as vibratory bowl feeders, vibratory belt feeders, and programmable belt feeders.

Mobile Robot. A special type of manipulator which is not bolted to the floor but can move. Based on different driving mechanisms, mobile robots can be further classified as wheeled mobile robots, legged mobile robots, treaded mobile robots, underwater mobile robots, and aerial vehicles.

Path Planning. The process of finding a continuous path from an initial robot configuration to a goal configuration without collision.

PD-Gravity Control. A special case of computed-torque control where there is a PD outer control loop plus a gravity compensation inner control loop that makes the DC values of the tracking errors equal to zero.

Pinch Grasp. A grasp in which a part is clamped between fingertips.

Pixel. Picture element — one point of an image matrix in image processing terminology.

Prismatic Joint. Sliding robot joint which produces relative translation of the connected links.

Redundant Manipulator. Manipulator for which the number of joint variables is greater than the number of task coordinates.

Remote-Center Compliance (RCC). A compliant wrist or end effector designed so that task-related forces and moments produce deflections with a one-to-one correspondence (i.e., without side effects). This property simplifies programming of assembly and related tasks.

Revolute Joint. Rotary robot joint producing relative rotation of the connected links.

Robot Axis. A direction of travel or rotation usually associated with a degree of freedom of motion.

Robot Joint. A mechanism which connects the structural links of a robot manipulator together while allowing relative motion.

Robot Link. The rigid structural elements of a robot manipulator that are joined to form and arm.

Robust Control. A large class of control algorithms where the controller is generally nondynamic, but contains information on the maximum possible modeling uncertainties so that the tracking errors are kept small, often at the expense of large control effort. The tracking performance does not improve over time so the errors never go to zero.

SCARA. Selectively compliant assembly robot arm.

SET. (Specification for Exchange of Text) — a data exchange standard.

Singularity. Configuration for which the manipulator jacobian has less than full rank.

Skew Symmetry. A property of the dynamics of rigid-link robot arms, important in controller design, stating that $\dot{M} - {}^1/_2 V_m$ is skew symmetric, with M the inertia matrix and V_m the coriolis/centripetal matrix. This is equivalent to stating that the internal forces do no work.

Stewart Platform Manipulator. A special type of parallel-link robot consisting of six identical linear actuators in parallel, an upper platform, and a base. One end of each actuator connects to the base, and the other to the upper platform with two- or three-degrees-of-freedom joints. This manipulator has a greater force-to-weight ratio and finer positioning accuracy than any commercial serial-link robot.

Task Coordinates. Variables in a frame most suited to describing the task to be performed by manipulator.

VDAFS. (Virtual Data Acquisition and File Specification) — a data exchange standard.

Visibility Graph. A road map approach to path planning where the obstacles are modeled as polygons. The visibility graph has nodes given by the vertices of the polygons, the initial point, and the goal point. The links are straight line segments connecting the nodes without intersecting any obstacles.

Voronoi Diagram. A road map approach to path planning where the obstacles are modeled as polygons. The Voronoi diagram consists of line having an equal distance from adjacent obstacles; it is composed of straight lines and parabolas.

Wrap Grasp. A grasp in which fingers envelope a part, to sustain greater loads.

References

Anderson, R.J. and Spong, M.W. 1989. Bilateral control of teleoperators with time delay. *IEEE Trans. Robotics Automation.* 34(5):494–501.

Asfahl, C.R. 1992. *Robotics and Manufacturing Automation.* 2nd ed. John Wiley & Sons, New York.

Ballard, D.H. and Brown, C.M. 1982. *Computer Vision.* Prentice-Hall, Englewood Cliffs, NJ.

Book, W.J. 1984. Recursive Lagrangian dynamics of flexible manipulator arms. *Int. J. Robotics Res.* 3(3):87–101.

Boothroyd, G. and Dewhurst, P. 1985. Part presentation costs in robot assembly. *Assembly Automation,* 138–146.

Boothroyd, G., Dewhurst, P., and Knight, W. 1994. *Product Design for Manufacture and Assembly.* Marcel Dekker, New York.

Bottema, O. and Roth, B. 1979. *Theoretical Kinematics,* North Holland, Amsterdam.

Bralla, J.G. (ed.). 1986. *Handbook of Product Design for Manufacturing,* McGraw-Hill, New York, 7-75, 7-100.

Craig, J. 1985. *Adaptive Control of Mechanical Manipulators.* Addison-Wesley, Reading, MA.

Craig, J.J. 1989. *Introduction to Robotics: Mechanics and Control.* Addison-Wesley, Reading, MA.

Critchlow, A.J. 1985. *Introduction to Robotics.* Macmillan, New York.

Decelle, L.S. 1988. Design of a robotic workstation for component insertions. *ATEJT Tech. J.* 67(2):15–22.

Denavit, J. and Hartenberg, R.S. 1955. A kinematic notation for lower-pair mechanisms based on matrices, *J. Appl. Mech.* 22:215–221,

Duffy, J. 1980. *Analysis of Mechanisms and Robot Manipulators,* John Wiley & Sons, New York.

Elfes, A. 1987. Sonar-based real-world mapping and navigation. *IEEE J. Robotics Automation.* RA-3(3):249–265.

Fraden, J. 1993. *AIP Handbook Of Modern Sensors, Physics, Design, and Applications.* American Institute of Physics, New York.

Fu, K.S., Gonzalez, R.C., and Lee, C.S.G. 1987. *Robotics.* McGraw-Hill, New York.

Fuller, J.L. 1991. *Robotics: Introduction, Programming, and Projects.* Macmillan, New York.

GMF Robotics Training and Documentation Department. 1985. *Paint Processing: Concepts and Practices.* GMF Robotics Corporation, Troy, MI.

Groover, M.K., Weiss, M., Nagel, R.N., and Odrey, N.G. 1986. *Industrial Robotics: Technology, Programming, and Applications.* McGraw-Hill, New York.

Gruver, W.A., Soroka, B.I., and Craig, J.J. 1984. Industrial robot programming languages: a comparative evaluation. *IEEE Trans. Syst. Man Cybernetics.* SMC-14(4).

Hayati, S., Tso, K., and Lee, T. 1989. Dual arm coordination and control. *Robotics.* 5(4):333–344.

Hollerbach, J.M., Hunter, I.W., and Ballantyne, J. 1992. A comparative analysis of actuator technologies for robotics. In *Robotics Review 2,* O. Khatib, J. Craig, and T. Lozano-Perez, (eds.). MIT Press, Cambridge, MA, 299–342.

Hollis, R.L., Allan, A.P., and Salcudean, S. 1988. A six degree-of-freedom magnetically levitated variable compliance fine motion wrist. In *Robotics Research, the 4th Int. Symp.,* R. Bolles and B. Roth., (eds.). MIT Press, Cambridge, MA, 65–73.

Jacobsen, S., Wood, J., Knutti, D.F., and Biggers, K.B. 1984. The Utah/M.I.T. dextrous hand: work in progress. *Int. J. Robotics Res.* 3(4):Winter.

Jamshidi, M., Lumia, R., Mullins, J., and Shahinpoor, M. *Robotics and Manufacturing: Recent Trends in Research, Education, and Applications*, Vol. 4. ASME Press, New York.

Klafter, R.D., Chmielewski, T.A., and Negin, M. 1989. *Robotic Engineering: An Integrated Approach.* Prentice-Hall, Englewood Cliffs, NJ.

Latombe, J.C. 1991. *Robot Motion Planning.* Kluwer Academic Publishers, Amsterdam.

Lee, K.-M. 1991. Flexible part-feeding system for machine loading and assembly. I. A state-of-the-art survey. II. A cost-effective solution. *Int. J. Prod. Economics.* 25:141–168.

Lee, K.-M. and Blenis, R. 1994. Design concept and prototype development of a flexible integrated vision system, *J. Robotic Syst.*, 11(5):387–398.

Lee, K.-M. and Li, D. 1991. Retroreflective vision sensing for generic part presentation. *J. Robotic Syst.* 8(1):55–73.

Leu, M.C. 1985. Robotics software systems. *Rob. Comput. Integr. Manuf.* 2(1):1–12.

Lewis, F.L., Abdallah, C.T., and Dawson, D.M. 1993. *Control of Robot Manipulators.* Macmillan, New York.

Lewis, F.L., Liu, K., and Yesildirek, A. 1995. Neural net robot controller with guaranteed tracking performance. *IEEE Trans. Neural Networks.* 6(3):703–715.

Liu, K., Fitzgerald, J.M., and Fewis, F.L. 1993. Kinematic analysis of a Stewart platform manipulator. *IEEE Trans. Ind. Electronics.* 40(2):282–293.

Lozano-Perez, T. 1983. Robot programming. *Proc. IEEE.* 71(7):821–841.

McClamroch, N.H. and Wang, D. 1988. Feedback stabilization and tracking of constrained robots. *IEEE Trans. Automat. Control.* 33:419–426.

Mujtaba, M.S. 1982. The AL robot programming language. *Comput. Eng.* (2):77–86.

Nichols, H.R. and Lee, M.H. 1989. A survey of robot tactile sensing technology. *Int. J. Robotics Res.* 8(3):3–30.

Nomura, H. and Middle, J.E. 1994. *Sensors and Control Systems in Arc Welding.* Chapman and Hall, 2-6 Boundary Row, London SE1 8HN, UK.

Okabe, A., Boots, B., and Sugihara, K. 1992. *Spatial Tessellations, Concepts and Application of Voronoi Diagrams.* John Wiley & Sons, New York.

Pertin-Trocaz, J. 1989. Grasping: a state of the art. In *The Robotics Review 1*, O. Khatib, J. Craig, and T. Lozano-Perez, Eds. MIT Press, Cambridge, MA, 71–98.

Priest, J.W. 1988. *Engineering Design for Producibility and Reliability.* Marcel Dekker, New York.

Rossi, M. 1985. Dialogues. *Manuf. Eng.* October: 41:24.

Sandler, B.Z. 1991. *Robotics, Designing the Mechanisms for Automated Machinery.* Prentice-Hall, Englewood Cliffs, NJ.

Shimano, B.E., Geschke, C.C., and Spalding, C.H., III. 1984. Val-II: a new robot control system for automatic manufacturing. *Proc. Int. Conf. Robotics.* March 13–15:278–292.

Siciliano, B. 1990. Kinematic control of redundant robot manipulators: a tutorial. *J. Intelligent Robotic Syst.* 3(3):201–210.

SILMA Incorporated. 1992. *SILMA CimStation Robotics Technical Overview.* SILMA Incorporated, 1601 Saratoga-Sunnyvale Road, Cupertino, CA.

Slotine, J.-J. 1988. Putting physics in control: the example of robotics. *Control Syst. Mag.* 8 (December):12–15.

Snyder, W.E. 1985. *Industrial Robots: Computer Interfacing and Control.* Prentice-Hall, Englewood Cliffs, NJ.

Spong, M.W. and Vidyasagar, M. 1989. *Robot Dynamics and Control.* John Wiley & Sons, New York.

Stauffer, R.N. 1984. Robotic assembly. *Robotics Today.* October.

Stevens, G.T. 1994. *The Economic Analysis of Capital Expenditures for Managers and Engineers.* Ginn Press, Needham Heights. MA.

Stewart, D. 1965. A platform with six degrees of freedom. *Proc. Inst. Mech. Engr.* (*London)* 180(15):371–386.

Tanner, W.R. 1994. Product design and production planning. In *CRC Handbook for Robotics.* CRC Press, Boca Raton, FL, 537.

Taylor, R.H., Summers, P.D., and Meyers, J.M. 1982. AML: a manufacturing language. *Int. J. Robotics Res.* (1):19–41.

Tsai, R.Y. and Lenz, R.K. 1989. A new technique for fully autonomous and efficient 3D robotics hand/eye calibration. *IEEE Trans. Robotics Automation.* 5(3):345–358.

Tzou, H.S. and Fukuda, T. 1992. *Precision Sensors, Actuators, and Systems.* Kluwer Academic, Publishers, Amsterdam.

Winston, P.H. 1984. *Artificial Intelligence.* Addison-Wesley, Reading, MA.

Wright, P.K. and Cutkosky, M.R. 1985. Design of grippers. In *The Handbook of Industrial Robotics,* S. Nof, (ed.). John Wiley & Sons, New York, chap. *2*.4.

Additional Reading

For a less mathematical treatment of robotics, including topics in manufacturing and programming, see the book by Fuller (1991). For further reading on information flow and computer science aspects of robotics, see the chapter on "Robotics" in the *CRC Handbook of Computer Science Engineering.* More details on manufacturing and industrial robot applications are found in Asfahl (1985) and Groover et al. (1986). For more on dynamics and control of robot manipulators one may examine books by Lewis et al. (1993) or Spong and Vidyasagar (1989). Robotics including topics in control, vision processing, and programming aspects is discussed in Fu et al. (1987). A constant source of articles on all aspects of robotics is the *IEEE Transactions on Robotics and Automation.*

Mish, K.D. and Mello, J. “Computer-Aided Engineering”
Mechanical Engineering Handbook
Ed. Frank Kreith
Boca Raton: CRC Press LLC, 1999

Computer-Aided Engineering

Kyran D. Mish
California State University, Chico

Joseph Mello
Aerojet Corporation

15.1 Introduction

The revolution in computer technology that has taken place over the last few decades has changed the face of all engineering disciplines. Fortunately, computers appropriate for solving complex mechanical engineering problems have evolved from complicated, rare, and expensive mainframes or supercomputers of yesterday to simple, common, and inexpensive desktop microcomputers widely available today. During this period, the application of improved computer technology has revolutionized the profession of mechanical engineering and enabled practicing engineers to solve problems routinely that a few decades ago would have resisted any available solution technique. However, many engineers possess only a limited grasp of important topics such as computational mechanics, computer science, and advanced computer-aided design principles. Thus, this chapter the present fundamentals of these computational

topics, in order to assist the practicing mechanical engineer in maintaining an appropriate high-level understanding of these emerging engineering tools.

Definition of Terms (CAD/CAM/CAE)

First, it is necessary to define a few terms related to the integration of the computer into various aspects of engineering practice:

- Computer-aided engineering (CAE) is a broadly inclusive term that includes application of the computer for general tasks encountered in mechanical engineering, including analysis, design, and production.
- Computer-aided manufacture (CAM) is the topic concerned with the integration of the computer into the manufacturing process, including such tasks as controlling real-time production devices on the factory floor. This subject is treated in Section 13.
- Computer-aided design (CAD) is a general term outlining areas where computer technology is used to speed up or eliminate efforts by engineers within the design/analysis cycle. This topic is also covered in Section 11.

The acronym CAD is also used as an abbreviation for computer-aided drafting. Since drafting is a key component of the mechanical engineering design process, this more narrow meaning will be integrated into the general notion of computer-aided design practice.

Overview of Computer-Aided Engineering

There are many important components of integrating the computer into mechanical engineering. In the past few decades, integrating the computer into mechanical engineering practice was largely a matter of generating input data for computational analyses and subsequent examination of the output in order to verify the results and to examine the response of the computer model. Today there are a wide variety of integration schemes used in engineering practice, ranging from artificial intelligence applications oriented toward removing the engineer from the design/analysis cycle, to “engineer in the loop” simulations intended to offload to the computer the quantitative tedium of design and analysis in order to permit the engineer to concentrate instead on qualitative issues of professional judgment. In the near future, these existing computer integration techniques will be combined with virtual reality technology and improved models for human/computer interaction. In order to use these new developments, engineers must be appropriately schooled in the fundamental tenets of modern computer-aided engineering.

Goals of Computer-Aided Engineering Chapter

This chapter is designed to aid practicing mechanical engineers in understanding the scope of applications of computer technology to their profession. To avoid premature obsolescence of the material presented here, a high-level orientation is utilized. Details of computational techniques are relegated to a catalog of appropriate references, and high-level “practical” concepts are emphasized. Low-level details (such as computer performance issues) change rapidly relative to high-level concerns (such as general classifications of computer function, or qualitative measures of computer solution quality). Thus, concentrating on a technical overview of computer-aided engineering will permit this chapter to stay up-to-date over the long term. In addition, this approach provides the practicing mechanical engineer with the fundamental principles required to accommodate the relentless pace of change that characterizes the world of modern computation.

In addition to providing engineers with basic principles of computer science, computational mechanics, computer intelligence, and examples of the application of CAD/CAM principles, this chapter also provides insight into such simple but commonly neglected issues as:

- What are the best ways to identify and judge computer applications for integration and use in a practical engineering setting?
- What common pitfalls can be encountered using general-purpose CAE software, and how can these problems be avoided?
- What are the most important characteristics of engineering computation, and how do engineering computer problems differ from computational methods used in other fields?

The development of material in this chapter is oriented toward presenting the practical *how* of effectively using engineering software as opposed to the mathematical *why* of identifying the cause of pathological computer results. This chapter therefore first provides the practicing mechanical engineer with an overview of the field of computer-aided engineering, and then presents that material in a manner less likely to become obsolete due to the rapid pace of change in this important engineering discipline. Finally, in order to avoid overlap with other chapters, many details (for example, mathematical definitions) are relegated to other chapters, or to appropriate references. Thus this chapter mainly provides an overview of the broad subject of computer-aided engineering, leaving the individual component subjects to other chapters or to outside references.

15.2 Computer Programming and Computer Architecture

Few other aspects of engineering practice change as rapidly as the use of the computer. Fifty years ago, mechanical engineers analyzed and designed automobiles that are not radically different from the automobiles of today, but practicing engineers of that era had no means of electronic calculation. Twenty-five years ago, mechanical engineers designed high-performance aircraft and space vehicles that are similar to those currently under development, but the computers of that more recent era had less power than a typical microprocessor available today in an automobile's braking system. Ten years ago, an advanced supercomputer appropriate for high-performance engineering computation cost several million dollars, required custom construction, installation, and support, and could perform between 10 million and 100 million floating-point operations per second. Today, one can buy such computational power for a few thousand dollars at a retail electronics store. Given this background of astonishing change in the performance, size, and cost of computers, it is clear that the applications of computers to the practice of mechanical engineering can be expected to advance at an equally breathtaking rate.

This section presents fundamental principles from the viewpoint of what can be expected between today and the beginning of the next century. The emphasis is on basic principles and fundamental definitions.

Software Engineering Overview

The modern field of software engineering is an outgrowth of what used to be called computer programming. Like many other engineering disciplines, software engineering gains much of its knowledge base from the study of failures in the specification, design, and implementation of computer programs. Unlike most engineering fields, however, the accepted codes of practice for software engineering are not yet universally adhered to and many preventable (and expensive) software failures still occur.

History of the Software Engineering Discipline

Software engineering is barely two decades old. Early efforts in this field are based on viewing the process of designing and implementing computer programs as analogous to designing and constructing physical objects. One of the most influential works demonstrating this "manufacturing tool" approach to the conceptualization of software is *Software Tools* (Kernighan and Plauger, 1976). In this view, large software projects are decomposed into reusable individual components in the same manner that complex machinery can be broken up into smaller reusable parts. The notion that computer programs could be constructed from components that can then be reused in new configurations is a theme that runs

throughout the history of the software engineering field. Modularity alone is not a panacea, however, in that programming effort contained within modules must be readable so as to facilitate verification, maintenance, and extension. One of the most influential references on creating readable and reliable code is *The Elements of Programming Style* (Kernighan and Plauger, 1978).

Emphasizing a modular and reusable approach to software components represents one of the primary threads of software engineering, and another important related thread arises in parallel to the idea of "manufacture" of software as an administrative process. After the failure of several large software projects in the 1960s and 1970s, early practitioners of software engineering observed that many of these software failures resulted from poor management of the software development process. Chronicles of typical failures and the lessons to be learned are detailed in Fred Brooks' *The Mythical Man-Month* (Brooks, 1975), and Philip Metzger's *Managing a Programming Project* (Metzger, 1983). Brooks' conclusions are generally considered among the most fundamental principles of the field. For example, his observation that adding manpower to a late software project makes it later is generally accorded the status of an axiom, and is widely termed "Brooks's law." More recent efforts in software management have concentrated on constructive solutions to past failures, as exemplified by Watts Humphrey's *Managing the Software Process* (Humphrey, 1989).

Understanding of the local (poor design) and global (poor management) causes of software failure has led to the modern study of software engineering. Obstacles to modularization and reuse of software tools were identified and proposed remedies suggested and tested. Administrative and large-scale architectural failures have been documented, digested, and their appropriate solutions published. The result is a substantial body of knowledge concerning such important topics as specification and design of computer programs, implementation details (such as choice of programming language), and usability of software (e.g., how the human user relates to the computer program). This collection of knowledge forms the basis for the modern practice of software engineering (see Yourdon, 1982).

Standard Models for Program Design and Implementation

A simple conceptual model of a program design and implementation methodology is presented by Yourdon (1993). This model is termed the "waterfall model" because its associated schematic overview bears a resemblance to a waterfall. The fundamental steps in the waterfall model (see Figure 15.2.1) proceed consecutively as follows:

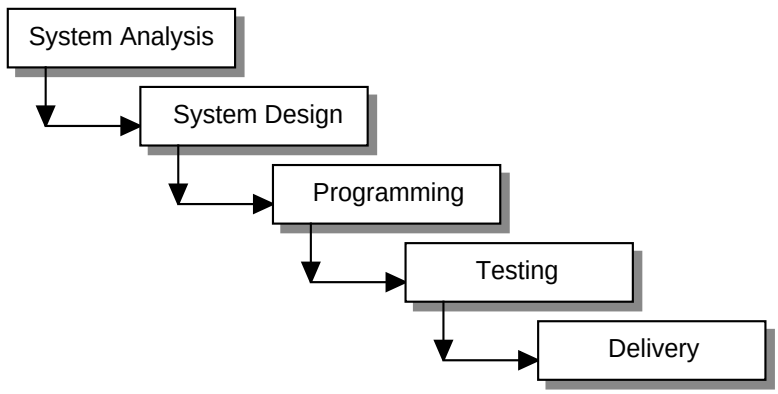


FIGURE 15.2.1 Waterfall model for software development.

- *System Analysis* — the requirements of the software system are analyzed and enumerated as a set of program specifications.
- *System Design* — the specifications are translated into precise details of implementation, including decomposition of the software system into relevant modules, and the implementation details required within this modular structure.

- *Programming* — the proposed design is implemented in a computer programming language (or collection of languages) according to the structure imposed in the system design phase.
- *Testing* — the resulting computer program is tested to insure that no errors ("bugs") are present, and that it is sufficiently efficient in terms of speed, required resources (e.g., physical memory), and usability.
- *Delivery* — the completed program is delivered to the customer.

This approach for software development works well for the small tasks, but for larger mechanical engineering projects there are many problems with this model. Probably the most important difficulty is that large projects may take years of development effort, and the customer is forced to wait nearly until the end of the project before any results are seen. If the project's needs change during the implementation period (or more commonly, if those needs were not specified with sufficient accuracy in the first place), then the final product will be inadequate, and required changes will have to be specified before those modifications can be "percolated" back through the waterfall life cycle again. Another important difficulty with this scheme occurs because quality-control issues (e.g., testing) are delayed until the project is nearly finished. In practice, preventative quality-assurance measures work better at reducing the likelihood that errors will occur in the program.

To avoid the inherent difficulties of the waterfall design, restructure this sequential design model into a spiral form that emphasizes more rapid incremental deliveries of programming function. These modified techniques are termed "incremental" or "spiral" development schemes (see Figure 15.2.2). The emphasis on quick delivery of incremental programming function permits seeing results sooner and, if necessary, making changes.

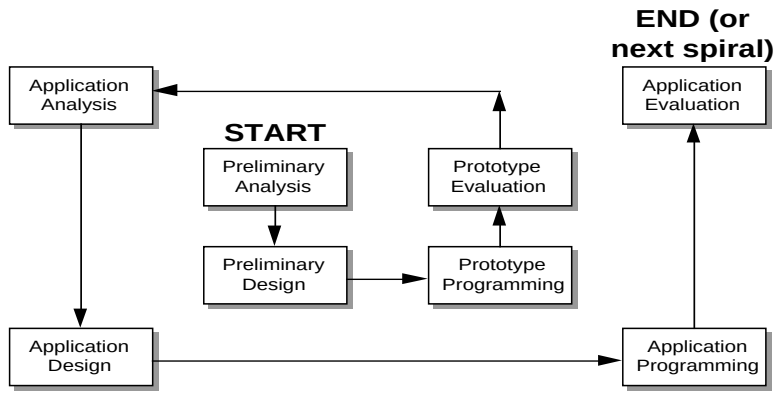


FIGURE 15.2.2 Spiral model for software development.

One important interpretation of the spiral development model is that the initial implementation of the program (i.e., the first pass through the spiral) can often be constructed as a simple prototype of the desired final application. This prototype is often implemented with limited function, in order to simplify the initial development process, so that little or no code developed in the prototypical implementation ends up being reused in the final commercial version. This approach has a rich history in practical software development, where it is termed "throwing the first one away" (after a similarly titled chapter in Brooks, [1975]). This widely used scheme uses the prototype only for purposes of demonstration, education (i.e., to learn the location of potential development pitfalls), and marketing. Once the prototype has been constructed and approved, design and implementation strategies more similar to the traditional waterfall approach are used for commercial implementation.

An obvious argument against spiral development models that incorporate a disposable prototype is that the cost of development is paid twice: once for the prototype and once again for the final implementation. Current remedies for this inherent problem include "rapid application development" (RAD)

methods, which emphasize the use of computer-assisted software engineering (CASE) tools to permit substantial reuse of the prototype in the final version. With this approach, the programmer uses specially designed computer applications that build the programs out of existing components from a software library. Program components that are especially mundane or difficult to implement (such as those commonly associated with managing a program's graphical user interface) often represent excellent candidates for automation using RAD or CASE tools. This area of software design and implementation will undoubtedly become even more important in the future, as the ability of computer programs to create other computer applications improves with time.

Computer Languages

Regardless of what future developments occur in the field of CASE tools, most current computer programming efforts are carried out by human programmers using high-level programming languages. These programming languages abstract low-level details of computer function (such as the processes of performing numerical operations or allocating computer resources), thus allowing the programmer to concentrate on high-level concepts appropriate for software design and implementation. Figure 15.2.3 diagrams the various interrelationships among abstractions of data, abstraction of instructions, and combinations of abstracted data and instructions. In general, more expressive languages are to be found along the diagonal line, but there are many instances in which increased abstraction is not necessarily a primary goal (for instance, performing fast matrix operations, which can often be done efficiently in a procedural language).

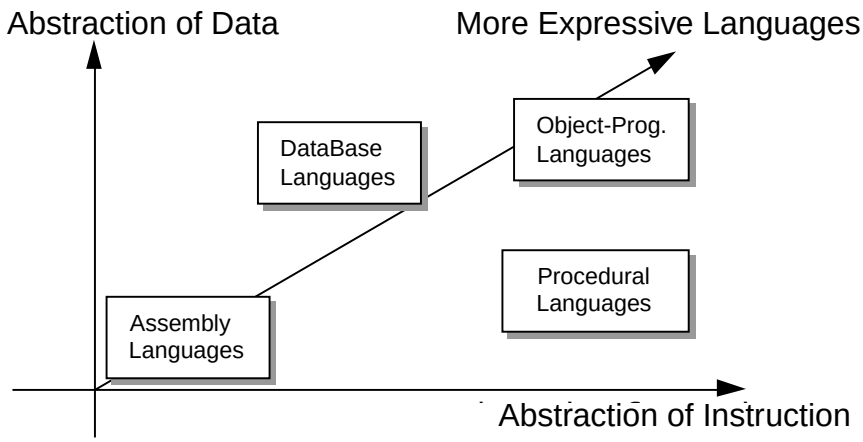


FIGURE 15.2.3 Language abstraction classification.

In addition, high-level programming languages provide a means for making computer programs *portable*, in that they may be moved from one type of computer to another with relative ease. The combination of portability of computer languages and adherence to standard principles of software design enables programmers to produce applications that run on a wide variety of computer hardware platforms. In addition, programs that are designed to be portable across different computers available today are generally easy to migrate to new computers of tomorrow, which insulates the longer-term software development process from short-term advances in computer architecture.

Perhaps the best example of how good program design and high-level language adherence work together to provide longevity for computer software is the UNIX operating system. This common operating system was designed and implemented in the 1970s using the C programming language, and runs with only minor modifications on virtually every computer available today. This operating system and its associated suite of software tools provide an excellent demonstration of how good software

design, when implemented in a portable programming language, can result in a software product that exhibits a lifespan much greater than that associated with particular hardware platform.

Computer Language Classification

Computer programs consist of two primary components: data and instructions. Instructions reflect actions taken by the computer program, and are comparable to verbs in natural human languages. In a similar manner, data reflect the objects acted on by computer instructions, in analogy with nouns in natural languages. Just as one would not attempt to construct sentences with only nouns or verbs, both data and instructions are necessary components of computer programs. However, programming languages are classified according to the relative importance of each of these components, and the resulting characterization of programming languages as "data-centered" (object-oriented) or "instruction-centered" (procedural) has major ramifications toward design, implementation, and maintenance of computer software.

Older programming languages (such as FORTRAN and COBOL) reflect a preeminent role for instructions. Languages that abstract the instruction component of programs permit the programmer to recursively decompose the overall computational task into logically separate subtasks. This successive decomposition of overall function into component subtasks highlights the role of the individual task as an encapsulated set of programming instructions. These encapsulated individual tasks are termed "procedures" (e.g., the SUBROUTINE structure in FORTRAN), and this "instruction first" approach to program design and construction is thus termed *procedural programming*. In procedural programming, the emphasis on decomposition of function obscures the important fact that the data component of the program is inherited from the procedural design, which places the data in a secondary role.

A more recent alternative model for programming involves elevating the data component of the program to a more preeminent position. This newer approach is collectively termed *object* programming, or *object-oriented* programming. Where procedural models abstract programming function via encapsulation of instructions into subroutines, object programming models bind the procedures to the data on which they operate. The design of the program is initially oriented toward modeling the natural data of the system in terms of its behaviors, and once this data component has been specified, the procedures that act upon the data are defined in terms of their functional operations on these preeminent data structures. Object programming models have been very successful in the important task of the creation of reusable code, and are thus valuable for settings (like implementation of a graphical user interface, where consistency among applications is desired) where there is a natural need for considerable code reuse.

The choice of language in computer programming is difficult because proponents of various languages and programming models favoring particular approaches often castigate those who advocate otherwise. Fortunately, there are few other branches of engineering that are so heavily politicized. It is important to look beyond the dogmatic issues surrounding language choice toward the important pragmatic goals of creating readable, verifiable, extensible, and reusable code.

Finally, it is important to recognize that the classification of programming languages into procedural and object-programming models is not precise. Regardless of whether data or instructions are given relative precedence, computer programs need both, and successful software design demands that the requirements of each component be investigated carefully. Although object-programming models relegate procedures to a secondary role to data structures, much of the final effort in writing an object-oriented program still involves design, creation, and testing of the procedures (which are termed "class methods" in object-programming models) that act upon the data. Similarly, it is possible to gain many object-programming advantages while using strictly procedural languages. In fact, some of the most successful languages utilized in current practice (e.g., the C++ programming language) are completely suitable for use as either procedural or object-oriented programming languages.

Procedural Programming Models

The most commonly used procedural programming languages in current use include Ada, C, FORTRAN, Basic, and Pascal. These procedural programming languages are characterized by a natural modular

structure based on programming function, which results in similar design methods being used for each. Since procedural languages are primarily oriented toward encapsulating programming function, each language has a rich set of control structures (e.g., looping, logical testing, etc.) that permits an appropriate level of control over the execution of the various procedural functions. Beyond this natural similarity, most procedural languages exhibit vast differences based on the expressiveness of the language, the range and extensibility of native data types, the facilities available for implementing modularity in the component procedures, and the run-time execution characteristics of the language.

Design Principles for Procedural Programming Models. The fundamental design principle for procedural programming is based on the concept of "divide and conquer," and is termed *functional decomposition* or *top-down design*. The overall effort of the program is successively decomposed into smaller logically separate subtasks, until each remaining subtask is sufficiently limited in scope so as to admit implementation within a procedure of appropriately small size. The overall process is diagrammed in Figure 15.2.4.

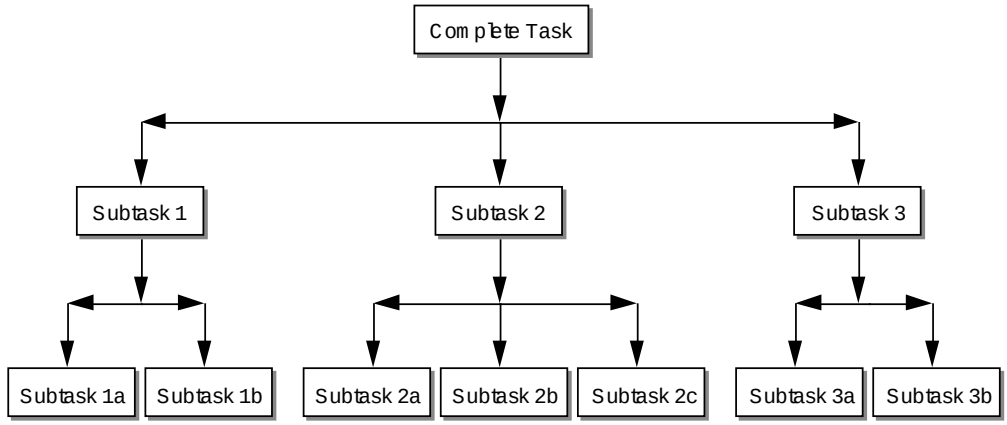


FIGURE 15.2.4 Sample functional decomposition.

The basic motivation behind functional decomposition is that the human mind is incapable of understanding an entire large computer program unless it is effectively abstracted into smaller "black boxes," each of which is simple enough so that its complete implementation can be grasped by the programmer. The issue of exactly how large an individual procedure can be before it becomes too large to understand is not an easy question to answer, but rules of thumb range from around ten up to a few hundred lines of executable statements. While it is generally accepted that extremely long procedures (more than a few hundred lines) have been empirically shown to be difficult to comprehend, there is still a lively debate in the software engineering community about how short procedures can become before an increased error rate (e.g., average number of errors per line of code) becomes apparent.

Although procedural programming languages possess the common characteristic of encapsulating programming function into separate granular modules, there are few similarities beyond this basic architectural resemblance. The various procedural languages often exhibit substantial differences in expressiveness, in the inhibition of practices associated with common language errors, and in the run-time characteristics of the resulting computer program.

Expressiveness represents the range of features permitted by the language that can be used by the programmer to implement the particular programming design. There is considerable evidence from the study of linguistics to support the notion that the more expressive the language, the wider the range of thoughts that can be entertained using this language. This important postulate is termed the "Sapir-Whorf hypothesis." While the Sapir-Whorf hypothesis is considered controversial in the setting of natural human languages, in the vastly simpler setting of programming languages, this phenomenon has been commonly

observed in practice, where use of more expressive high-level languages has been correlated with overall programmer productivity. Expressiveness in a computer language generally consists of permitting more natural control structures for guiding the execution of the program, as well as permitting a wide range of data representations appropriate for natural abstraction of data.

Sample Procedural Programming Languages. Some procedural languages (including FORTRAN and Basic) permit only a limited set of control structures for looping, branching, and logical testing. For instance, before the FORTRAN-77 standard was promulgated, FORTRAN had no way of expressing the standard logical "if-then-else" statement. The FORTRAN-77 specification still does not permit standard nondeterministic looping structures, such as "do while" and "repeat until". The standard Basic language suffers from similar limitations and is further impeded by the fact that most implementations of Basic are interpreted (each line of code is sequentially translated and then executed) instead of compiled (where the entire program's code is translated first and executed subsequently). Interpreted languages such as Basic are often incredibly inefficient, especially on problems that involve substantial looping, as the overhead of retranslating each line of code cannot be amortized in the same manner available to compiled languages. Finally, because Basic is limited in its expressiveness, many implementations of Basic extend the language to permit a greater range of statements or data types. While language extension facilitates programmer expression, it generally compromises portability, as different nonstandard dialects of the extended language generally develop on different computer platforms. The extreme case is illustrated by Microsoft's Visual Basic language, which is completely tied to Microsoft applications and operating systems software (and thus inherently nonportable), but so useful in its extensions to the original Basic language that it has become the de facto scripting language for Microsoft applications and operating systems.

Ada and Pascal are very expressive languages, permitting a rich set of control structures and a simple extension of the set of permitted data types. In addition, Ada and some Pascal dialects force the programmer to implement certain forms of data modularity that are specifically designed to aid in the implementation of procedural programs. In a similar vein, standard Pascal is so strongly typed that it forces the programmer to avoid certain common practices (such as misrepresenting the type of a data structure passed to a procedure, which is a widespread and useful practice in FORTRAN) that are associated with common errors in program implementation. In theory, Pascal's strict approach to representing data structures and its rich set of control structures ought to make it an attractive language for engineering programming. In practice, its lack of features for arithmetic calculation and its strict rules on data representation make it fairly difficult to use for numeric computation. Ada is a more recent language that is based on Pascal, but remedies many of Pascal's deficiencies. Ada is a popular language in mechanical engineering applications, as it is mandated for use on many Department of Defense programming projects.

C and FORTRAN are among the most common procedural languages used in large-scale mechanical engineering software applications. Both are weakly typed compiled languages with a rich set of available mathematical operations. C permits a considerable range of expressive control structures and extensible data structures. In addition, C is extremely portable and generally compiles and runs quickly, as the language's features are closely tuned to the instruction sets of modern microprocessors used in current generations of computers. The original C language specification was replaced in 1988 by a new ANSI standard, and this current language specification adds some features (such as type checking on arguments passed to procedures, a facility that aids greatly in preventing common programming errors) that resemble those found in Pascal, but do not seem to compromise the overall utility of the original C language standard.

FORTRAN's current implementation (FORTRAN-90) adds user-defined extensible data structures and a more expressive instruction set to the FORTRAN-77 standard, but the FORTRAN-90 standard has so far been slow to gain acceptance in the programming community. FORTRAN-77 compilers are still common, and the problems associated with this version of the language (e.g., minimal resources for abstracting data, limited control structures for expressing program flow) still compromise the archi-

tecture of FORTRAN programs. However, FORTRAN retains a rich set of intrinsic numeric operations, so it is still a good choice for its original goal of "Formula Translation" (where the language derives its name). In addition, FORTRAN programs often execute very rapidly relative to other procedural languages, so for programs that emphasize rapid mathematical performance, FORTRAN is still a good language choice. Finally, many FORTRAN-callable libraries of mathematical operations commonly encountered in engineering applications are available, and this ability to leverage existing procedural libraries makes FORTRAN an excellent choice for many mechanical engineering applications.

Advantages and Disadvantages of Procedural Programming. Procedural programming has inherent advantages and disadvantages. One of the most important advantages of some procedural languages (notably FORTRAN) is the existence of many complete libraries of procedures for solving complex tasks. For example, there are many standard libraries for linear algebra (e.g., LINPACK, EISPACK, LAPACK) or general scientific numerical computation (e.g., IMSL) available in FORTRAN-callable form. Reuse of modules from these existing libraries permits programmers to reduce development costs substantially for a wide variety of engineering applications. Under most current portable operating systems, multiple-language integration is relatively straightforward, so high-quality FORTRAN-callable libraries can be called from C programs, and vice-versa. The development of standard procedural libraries is largely responsible for the present proliferation of useful computer applications in mechanical engineering.

Another important advantage of procedural languages is that many important computational tasks (such as translating mathematical models into analogous computer codes) are naturally converted from the underlying mathematical algorithm (which is generally a sequence of instructions, and hence amenable to encapsulation within a procedure) into an associated modular procedure. As long as the data used within a program do not become unduly complex, procedural languages permit easy implementation of many of the standard methods used in engineering analysis and design.

Perhaps the biggest disadvantage of procedural models is that they are harder to reuse than competitive object-programming models. These obstacles to code reuse arise from the fact that data are modeled in procedural programming as an afterthought to the simulation of instructions. In order to reuse procedures between two different programs, the programmer must force the representation of data to be identical across the different computer applications. In procedural programming, the goal of code reuse requires standardization of data structures across different computer programs, regardless of whether or not such standardization is natural (or even desired) by those disparate computer programs. Object programming models are an attractive alternative to procedural programming schemes in large part because these newer programming methods successfully avoid such unwarranted data standardization.

Object Programming Models

Object-programming models place modeling of data structures in a more preeminent position, and then bind to the data structures the procedures that manipulate the data. This relegation of procedures (which are termed "methods" in object programming) to a more secondary role facilitates a degree of code reuse substantially better than is feasible with conventional procedural programming languages. Object programming languages employ aggregate data types (consisting of various data fields, as well as the associated methods that manipulate the data) that are termed *classes*, and these classes serve as templates for creation of *objects* that represent specific instances of the class type. Objects thus form the representative granularity found in object-oriented programming models, and interactions among objects during program execution are represented by messages that are passed among the various objects present. Each message sent by one object to another tells the receiving object what to do, and the details of exactly how the receiving object accomplishes the associated task are generally private to the class. This latter issue of privacy regarding implementation details of object methods leads to an independence among objects that is one of the main reasons that object-programming schemes facilitate the desired goal of code reuse.

One of the most important limitations of procedural languages is abstraction. High-level languages such as FORTRAN or C permit considerable abstraction of instructions, especially when compared to the machine and assembly languages they are designed to supplant. Unfortunately, these languages do not support similar levels of abstraction of data. For example, although FORTRAN-77 supports several different numeric types (e.g., INTEGER, REAL, DOUBLE PRECISION), the only derived types available for extending these simple numeric representations are given by vectors and multidimensional arrays. Unless the data of a problem are easily represented in one of these tabular forms, they cannot be easily abstracted in FORTRAN. To some extent, experienced programmers can create new user-defined data types in C using structures and typedefs, but effectively abstracting these derived data types requires considerable self-discipline on the part of the programmer.

Object-oriented programming languages avoid these pitfalls of procedural languages by using classes as templates for abstraction of both instructions and data. By binding the instructions and the data for classes together, the programmer can abstract both components of a programming model simultaneously, and this increased level of abstraction results in a radically new programming model. For instance, a natural class for finite-element modeling would be the class of finite-element mesh objects. A mesh object (which could easily be composed internally of node and element objects) makes it possible for the object programmer to hide all the details of mesh representation from the rest of the program. A procedural programming model would require standardization of the mesh to consist of (for example):

- A list of nodes, each associated with individual nodal coordinates given in 1D, 2D, or 3D (depending on the geometry of the model used)
- A list of elements, each with a given number of associated nodes
- A list of element characteristics, such as material properties or applied loads

In this representation, each procedure that manipulates any of the mesh data must know all of the details of how these data have been standardized. In particular, each routine must know whether a 1D, 2D, or 3D finite-element analysis is being performed, and pertinent details of the analysis (e.g., is the problem being modeled thermal conduction or mechanical deformation?) are also spread throughout the code by the standardization of data into predefined formats. The sum of these constraints is to require the programmer to recode substantial components of a procedural program every time a major modification is desired.

In the setting of objects, the finite-element mesh object would store its particular geometric implementation internally, so that the rest of the program would be insulated from the effects of changes in that representation. Rather than calling a procedure to generate a mesh by passing predefined lists of nodes, elements, and element characteristics, an object-oriented approach to mesh generation would employ sending a message such as “discretize yourself” to the mesh object. This object would then create its internal representation of the mesh (perhaps using default values created by earlier messages) and store this information privately. Alternatively, the object-oriented program might later send a “solve yourself” message to the mesh object and then a “report your results in tabular form” message for generating output. In each case, the rest of the program has no need to know the particular details of how the mesh object is generating, storing, or calculating results. Only the internal procedures local to the class (i.e., the class methods) generally need to know this private data, which are used locally to implement the functions that act on the class.

This hiding of internal function within an object is termed *encapsulation*, and object programming models permit simultaneous encapsulation of both data and instructions via appropriate abstraction. In this setting, encapsulation permits the programmer to concentrate on creating data and procedures naturally, instead of forcing either component into predefined formats such as floating-point arrays (for data) or predefined subroutine libraries (for instructions). Data and instruction abstraction of this form are thus useful additions to the similar (but less flexible) features available in procedural languages. If these new features constituted the only improvements available from object-programming models, then

they would offer only slight advantages over traditional procedural programming. There are many other advantages present in object-programming models.

The most important advantages of object programming occur because of the existence of *class hierarchies*. These hierarchies permit new objects to be created from others by concentrating only on the differences between the object's behaviors. For instance, a finite-element mesh for a rod lying in three dimensions can be derived from a one-dimensional mesh by adding two additional coordinates at each node. An object programmer could take an existing one-dimensional mesh class and derive a three-dimensional version using only very simple steps:

- Adding internal (private) representation for the additional coordinate data
- Overriding the existing discretization method to generate the remaining coordinates when the "discretize yourself" message is sent
- Overriding some low-level calculations in class methods pertinent to performing local element calculations using the new coordinate representation

Note that all of these steps are private to the mesh object, so that no other part of the program needs to be changed to implement this major modification to the problem statement. In practice, the added details are implemented via the creation of a derived class, where the additional coordinates and the modified methods are created. When messages appropriate to the new class are sent, the derived object created from the new class will handle only the modified data and instructions, and the parent object (the original mesh object) will take care of the rest of the processing. This characteristic of object-programming models is termed *inheritance*, as the individual derived ("child") objects inherit their behavior from the parent class. When the changes required by the modification to the program's specifications are small, the resulting programming effort is generally simple. When the changes are large (such as generalizing a one-dimensional problem to a fully three-dimensional one), it is still often feasible to make only minor modifications to the program to implement the new features.

One of the most important rationales for using object-programming methods arises from the desire to provide a consistent user-interface across diverse programs. Existing standardized graphical interface models (such as the Motif interface available on OSF/UNIX, or the Microsoft Windows interface used on Windows and Windows NT) place a premium on a consistent "look and feel" across different applications. Since managing the user-interface commonly constitutes much of the programming effort required to implement interactive engineering applications, it is advantageous to consolidate all of the code required to implement the standard graphical user-interface into a class library and allow the programmer to derive new objects pertinent to the application at hand.

One such class library is the Microsoft Foundation Classes, which implement the Windows interface via a class hierarchy requiring around a hundred thousand lines of existing C++ source code. Programmers using class libraries such as these can often generate full-featured graphics applications by writing only a few hundred or a few thousand lines of code (notably, for reading and storing data in files, for drawing content into windows, and for relevant calculations). In fact, it is relatively easy to graft graphical user interfaces onto existing procedural programs (such as old FORTRAN applications) by wrapping a C++ user-interface layer from an existing class library around the existing procedural code, and by recycling relevant procedures as class methods in the new object-oriented setting. This "interface wrapper" approach to recycling old procedural programs is one of many standard techniques used in *reengineering* of existing legacy applications (Barton and Nackman, 1994).

One other important characteristic of many object-programming languages is *polymorphism*. Polymorphism (Latin for "many forms") refers to the ability of a single message to spawn different behaviors in various objects. The precise meaning of polymorphism depends upon the run-time characteristics of the particular object-programming language used, but it is an important practical feature in any object-programming language.

Object-Oriented Design Principles. Because object-oriented programming is a relatively new discipline of software engineering (when compared to procedural programming), one cannot yet identify the best

design schemes among the various competing object-oriented design principles. For this reason (and to avoid prejudging the future), this section treats the subject of object-oriented design in less detail than procedural design methods.

The fundamental tenet of object-oriented program design is that the programming objects should be chosen to model any real-world objects present in the system to be analyzed and simulated. For example, in a thermal analysis of a microprocessor, one might identify such natural physical objects as "heat sink," "thermocouple," and "fan." In general, the nature of the physical objects in a mechanical system is stable over long periods of time, so they make natural candidates for programming objects, as their specifications are least likely to vary, and thus they will require minimal modifications to the basic program design.

The next step in performing an object-oriented design is to model the behaviors of the various objects identified within the system. For example, a fan object can "turn on" and "turn off", or might vary in intensity over a normalized range of values (e.g., 0.0 = off, 1.0 = high speed), and this behavior will form the basis for the messaging protocols used to inform objects which behaviors they should exhibit. At the same time, any relevant data appropriate to the object (in this case, fan speed, power consumption, requisite operating voltage) should be identified and catalogued. Here, these individual items of data will represent the private data of the fan class, and the behaviors of this class will be used to design class methods.

The final step in specifying an object-oriented design is to examine the various objects for interrelationships that can be exploited in a class hierarchy. In this setting, "heat sink" and "fan" could be considered to be derived from a larger class of "cooling devices" (although in this trivial example, this aggregation is probably unnecessary). Careful identification of hierarchical relationships among the candidate objects will generally result in an arrangement of classes that will permit considerable code reuse through inheritance, and this is one of the primary goals of object-programming design practice.

In practice, there is no "final" step in designing object-oriented programs, as the design process is necessarily more complex and iterative than procedural programming models. In addition, the object-oriented designer must take more care than given here in differentiating the role of classes (which are the static templates for construction of objects) from objects themselves, which are the dynamic realization of specific members of a class created when an object-oriented program executes. Objects are thus specific instances of generic classes, and the process of creating objects at run time (including setting all appropriate default values etc.) is termed *instantiation.*

Sample Object-Oriented Languages. There are not as many successful object-oriented languages as there are procedural languages, because some languages (such as Ada and FORTRAN-90) that possess limited object-oriented features are more properly classified as procedural languages. However, ADA 95 does include excellent facilities for object-oriented programming.

C++ is the most commonly used object-oriented language and was primarily developed at Bell Labs in the same pragmatic vein as its close procedural relative, C. In theory, the C++ language includes both procedural and object-programming models, and thus C++ can be used for either type of programming. In practice, the procedural features on C++ are nearly indistinguishable from those of ANSI C, and hence the phrase "programming in C++" is generally taken to mean "object-programming in C++". C++ is well known as an object-programming language that is not particularly elegant, but that is very popular because of its intimate relation with the C procedural programming language (C++ is a superset of ANSI C) and because of its extensive features. The design goal of maintaining back-compatibility with ANSI C has led to shortcomings in the C++ language implementation, but none of these shortcomings has seriously compromised its popularity. C++ is an efficient compiled language, providing the features of object-programming models without undue loss of performance relative to straight procedural C, and C++ is relatively easy to learn, especially for knowledgeable C programmers. It supports extensive inheritance, polymorphism, and a variety of pragmatic features (such as templates and structured exception handling) that are very useful in the implementation of production-quality code.

An important recent development in object-oriented design is the Java programming language: the popularity of this new language is closely tied to the explosion of interest in the Internet. Java is widely used to provide interactive content on the World-Wide-Web, and it has a syntax very similar to C++, a pervasive object-orientation, and provides portable elements for constructing graphical user interfaces. Java programs can be deployed using interpreted forms over the web (utilizing a "Java Virtual Machine" on the client platform), or by a more conventional (though less portable) compilation on the target computer.

SmallTalk is one of the oldest and most successful object-programming languages available, and was designed at the Xerox Corporation's Palo Alto Research Center (also responsible for the design of modern graphical user interfaces). SmallTalk supports both inheritance (in a more limited form than C++) and polymorphism, and is noted as a highly productive programming environment that is particularly amenable to rapid application development and construction of prototypes. SmallTalk is *not* a compiled language, and while this characteristic aids during the program implementation process, it generally leads to computer programs that are substantially less efficient than those implemented in C++. SmallTalk is generally used in highly portable programming environments that possess a rich library of classes, so that it is very easy to use SmallTalk to assemble portable graphical interactive programs from existing object components.

Eiffel is a newer object-oriented language with similar structure to object-oriented variants of the Pascal procedural programming language. Eiffel is similar in overall function to C++ but is considerably more elegant, as Eiffel does not carry the baggage of backward compatibility with ANSI C. Eiffel has many important features that are commonly implemented in commercial-quality C++ class libraries, including run-time checking for corruption of objects, which is a tremendous aid during the program debugging process. Even with its elegant features, however, Eiffel has not gained the level of acceptance of C++.

There are other object-oriented programming languages that are worth mentioning. The procedural language Ada provides some support for objects, but neither inheritance or polymorphism. FORTRAN-90 is similarly limited in its support for object-programming practices. Object Pascal is a variant of Pascal that grafts SmallTalk-like object orientation onto the Pascal procedural language, and several successful implementations of Object Pascal exist (in particular, the Apple Macintosh microcomputer used Object Pascal calling conventions, and this language was used for most commercial Macintosh application development for many years). For now, none of these languages provides sufficient support for object-oriented programming features (or a large-enough user community) to provide serious competition for C++, SmallTalk, Eiffel, or Java.

Data Base Systems

In procedural programming practice, modeling data are relegated to an inferior role relative to modeling instructions. Before the advent of object-oriented programming languages, which permit a greater degree of data abstraction, problems defined by large or complex data sets required more flexibility for modeling data than traditional procedural programming techniques allowed. To fill this void, specialized data base management systems were developed, and a separate discipline of computer programming (data base management) arose around the practical issues of "data-centered" programming practice. The study of data base management evolved its own terminology and code of application design, and suffered through many of the same problems (such as language standardization to provide cross-platform portability) that had plagued early efforts in procedural programming. The data base management subdiscipline of software engineering is still fundamentally important, but the widespread adoption of object-oriented languages (which permit flexible modeling of data in a more portable manner than that provided by proprietary data base management systems) has led to many of the concepts of data base management becoming incorporated into the framework of object-oriented programming practice.

Technical Overview of Data base Management

Many important engineering software applications are naturally represented as data base applications. Data base applications are generally developed within specialized custom programming environments specific to a particular commercial data base manager, and are usually programmed in a proprietary (and often nonportable) data base language. Regardless of these issues, data base programming is a particular form of computer programming, and so the relevant topics of software engineering, including procedural and object models, portability, reliability, etc., apply equally well to data base programming. Because many of these principles have already been presented in considerable detail, the following sections on design and programming issues for data base systems are kept relatively concise.

Data base applications are very similar to conventional programming applications, but one of the most important differences is in the terminology used. Data base applications have developed a nomenclature specifically defined to dealing with structured and unstructured data, and this terminology must be addressed. Some of the most appropriate terms are enumerated below.

- Table: a logical organized collection of related data
- Record: a collection of data that is associated with a single item (records are generally represented as rows in tabular data base applications)
- Field: an individual item of data in a record (fields are generally represented as columns in a tabular data base)
- Schema: the structure of the data base (schema generally is taken to mean the structure and organization of the tables in a tabular data base)
- Query: a structured question regarding the data stored in the data base (queries are the mechanism for retrieving desired data from the data base system)

There are many other relevant terms for data base management, but these are sufficient for this brief introduction. One important high-level definition used in data base management is *Structured Query Language*, or SQL. SQL is a standard language for creating and modifying data bases, retrieving information from data bases, and adding information to data bases. In theory, SQL provides an ANSI standard relational data base language specification that permits a degree of portability for data base applications. In practice, standard SQL is sufficiently limited in function so that it is commonly extended via proprietary addition of language features (this situation is similar to that of the Basic procedural language, which suffers from many incompatible dialects). The practical effect of these nonstandard extensions is to compromise the portability of some SQL-based data base systems, and additional standardization schemes are presently under development in the data base management industry. One such scheme is Microsoft's *Open data base Connectivity* (ODBC) programming interface, which provides portable data base services for relational and non-relational data base applications.

Classification of Data Base Systems

There are many different types of data base systems in common use. One of the most important initial steps in designing and implementing a data base application is to identify the relevant characteristics of the data in order to choose the most appropriate type of data base for development. Depending upon the structure of the data to be modeled, the data base application developer can select the simplest scheme that provides sufficient capabilities for the problem. Three sample data base structures are presented below: flat-file data bases, relational data bases, and object-oriented data bases.

Flat-File Data Bases. Flat-file data bases represent the simplest conceptual model for data base structure. A flat data base can be idealized as a table with a two-dimensional matrix or grid structure. The individual data base records are represented by the rows of the matrix, and each record's component fields are represented by the columns of the matrix structure, as shown in Figure 15.2.5. Flat-file data bases are thus confined to applications where all records are structurally identical (i.e., have the same configuration of fields) and where the underlying matrix structure naturally represents the data component of the application.

	Field 1: Name	Field 2: Yield Strength	Field 3: Young's Modulus	Field 4: Shear Modulus
Record 1: Material 1	Aluminum	250 MPa	70 GPa	25 GPa
Record 2: Material 2	Magnesium	150 MPa	45 GPa	18 GPa
Record 3: Material 3	Steel	400 MPa	200 GPa	85 GPa

FIGURE 15.2.5 Flat data base example.

The simplicity of a flat-file data base is simultaneously its greatest advantage and worst disadvantage. The main advantage of using a flat-file data base structure is that querying the data base is extremely simple and fast, and the resulting data base is easy to design, implement, and port between particular data base applications. In practice, spreadsheet applications are often used for constructing flat-file data bases, because these packages already implement the requisite tabular structure and include a rich variety of control structures for manipulating the data.

The biggest disadvantage of flat-file data bases is that the extreme simplicity of the flat structure simply does not reflect many important characteristics of representative data base applications. For programs requiring flexibility in data base schema, or complex relationships among individual data fields, flat-file data bases are simply a poor choice, and more complex data base models should be used.

Relational Data Bases. In practice, data base applications often require modeling relationships among various fields that may be contained in separate data files. Applications with these "relational" features are term relational data bases. Relational data base technology is a rapidly evolving field, and this family of data bases is very common in practical data base applications.

Relations provide a way to generalize flat-file data base tables to include additional features, such as variation in the numbers of fields among different records. A schematic of a simple relational data base schema is shown in Figure 15.2.6. Here a data base of material properties is represented by related tables. Note that because of the disparity in number of material constants (i.e., differing numbers of fields for each material record), a flat-file data base would not be suitable for this data base storage scheme.

The material properties tables (containing the lists of material properties) are related to their parent table, which contains overall identification information. These parent-child relationships give relational data bases considerable flexibility in modeling diverse aggregates of data, but also add complexity to the task of storing and retrieving data in the data base. In a flat-file data base system, a simple lookup (similar to indexing into a two-dimensional array) is required to find a particular field. In a complex relational data base, which may exhibit many nested layers of parent-child relations, the task of querying may become very complex and potentially time-consuming. Because of this inherent complexity in storing and retrieving data, the topics of efficient data base organization and of query optimization are essential for careful study before any large-scale relational data base application is undertaken.

Object-Oriented Data Bases. Many of the data base schemes found in relational and flat-file data base systems arose because of the inability to model data effectively in older procedural programming languages like FORTRAN. Commercial relational data base managers combined powerful data-modeling capabilities with new procedural languages (such as SQL or XBase) specifically designed to manipulate data base constructs. Recently, the current proliferation of object-oriented programming languages, with their innate ability to abstract data as effectively as possible with dedicated data base management systems, has led to the development of object-oriented data base systems. These object-oriented data base packages provide extremely powerful features that may ultimately make traditional SQL-based relational data base applications obsolete.

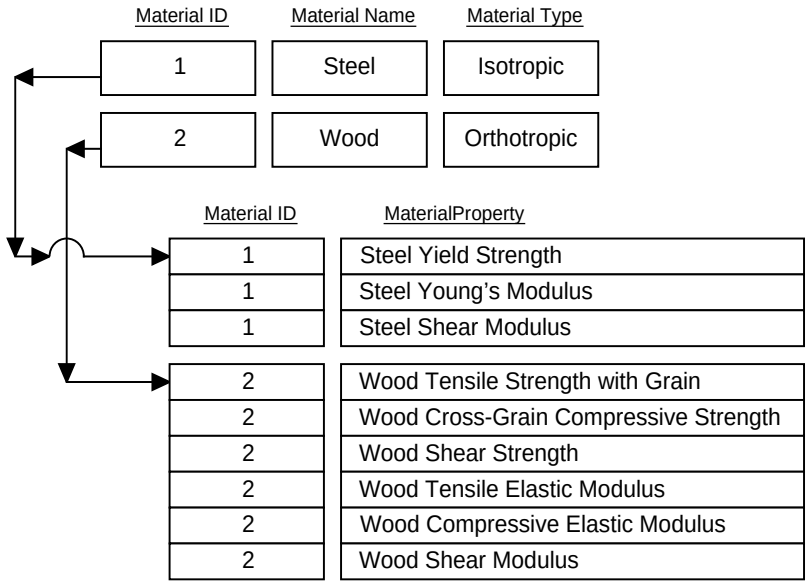


FIGURE 15.2.6 Example relational data base structure.

One interesting example of object-oriented data base technology is the integration of data base technology into a C++ framework. The Microsoft Foundation Class library for C++ provides numerous features formerly requiring custom data base programming that are implemented as C++ class library members. For example, there are extensible data base classes that provide direct support for common data base functions, and there are ODBC (Open Data Base Connectivity, the extension of SQL to generic data base environments) classes allowing the C++ program to access existing relational data bases developed with specialized data base management systems. Given the extensibility of C++ class libraries, this object-oriented approach makes it feasible to gain all of the advantages of proprietary relational data base applications, while preserving the numerous features of working in a standard portable programming language.

Operating System Characteristics

Computer programs depend on low-level resources for execution support, including file services for input/output, graphical display routines, scheduling, and memory management. The software layers that provide these low-level services are collectively termed the computer's *operating system*. Operating systems thus insulate individual programs from the details of the hardware platform where they are executed, and choosing the right operating system can be a critical decision in engineering practice.

Engineering computation is generally identified by three fundamental characteristics:

- Large demand for memory, where extremely large data sets (generally on the order of megabytes or gigabytes) are used, and where all components of these demanded memory resources must be accessible simultaneously (This demand for memory can be contrasted with standard on-line-transaction-processing schemes used in finance and commerce, where there is a similar characteristic of large data sets, but these large financial data models are seldom required to have all components available in memory at the same time.)
- Dependence on floating-point computation, where there are high-precision floating-point representations of numbers (i.e., numbers stored in the binary equivalent of "scientific notation", where storage is divided among sign, mantissa, and exponent, requiring more extensive storage than that required for characters or integer data types)

- Extensive use of graphics in input and display, as graphics is generally characterized as an engineer's "second language," because only the human visual sense has sufficient bandwidth to process the vast amounts of data generally present in engineering computation

While many of these characteristics may be found in other computational settings, the simultaneous presence of all three is a hallmark of computation in science and engineering. Identifying and selecting an operating system that provides appropriate support for these characteristics is thus a fundamentally important problem in the effective development and use of engineering software.

Technical Overview of Operating Systems

A simple and effective way to gain an overview of operating systems theory is to review the classification scheme used to identify various operating systems in terms of the services that they provide. The most common characteristics used for these classifications are enumerated below.

Multitasking. Humans are capable of performing multiple tasks simultaneously, and this characteristic is desirable in a computer operating system as well. Although an individual computer CPU can only process the instructions of one application at a time, it is possible with high-performance CPUs to manage the execution of separate programs concurrently by allocating processing time to each application in sequence. This sequential processing of different applications makes the computer appear to be executing more than one software application at a time. When an operating system is capable of managing the performance of concurrent tasks, it is termed a *multitasking* operating system. Many early operating systems (such as MS/DOS) could only execute a single task at a time and were hence termed *single-tasking* systems. While it is possible to load and store several programs in memory at one time and let the user switch between these programs (a technique sometimes termed *context switching* that is commonly used in MS/DOS applications), the lack of any coherent strategy for allocating resources among the competing programs limits the practical utility of this simple tasking scheme.

A simple generalization of context switching is known as *cooperative multitasking*, and this simple tasking scheme works remarkably well in some settings (in fact, this method is the basis for the popular Microsoft Windows 3.x and Apple Macintosh 7.x operating systems). In a cooperative multitasking setting, the allocation of computer resources is distributed among the competing programs: the individual programs are responsible for giving up resources when they are no longer needed. A comparison to human experience is a meeting attended by well-behaved individuals who readily yield the floor whenever another speaker desires to contribute. Just as this scheme for managing human interaction depends on the number of individuals present (obviously, the more people in the meeting, the more difficult the task of distributed management of interaction) as well as on the level of courtesy demonstrated by the individual speakers (e.g., there is no simple means for making a discourteous speaker yield the floor when someone else wants to speak), the successful use of cooperative multitasking schemes is completely dependent on the number and behavior of the individual software applications that are being managed. Ill-behaved programs (such as a communications application that allocates communications hardware when executed, but refuses to release it when not needed) compromise the effectiveness of cooperative multitasking schemes and may render this simple resource-sharing model completely unusable in many cases.

The obvious solution to managing a meeting of humans is to appoint a chair who is responsible for allocating the prioritized resources of the meeting: the chair decides who will speak and for how long, depending upon scheduling information such as the meeting's agenda. The computer equivalent of this approach is termed *preemptive multitasking* and is a very successful model for managing the allocation of computer resources. Operating systems that use a preemptive multitasking model make use of a *scheduler* subsystem that allocates computer resources (such as CPU time) according to a priority system. Low-priority applications (such as a clock accessory, which can update its display every minute or so without causing serious problems) are generally given appropriately rare access to system resources, while high-priority tasks (such as real-time data acquisition applications used in manufacturing, which cannot tolerate long intervals without access to the operating system's services) are given higher priority.

Of course, the scheduler itself is a software system and generally runs at the highest level of priority available.

Preemptive multitasking operating systems are natural candidates for engineering software, as the intense memory and hardware resources associated with engineering computation require appropriately high-powered operating system support. Virtually all large engineering computers of the present era (e.g., workstations, mainframes, and supercomputers) run operating systems that provide preemptive multitasking, and many microcomputers are now available with similar operating system support.

Multithreading. In the setting of multitasking, the term "task" has some inherent imprecision, and this ambiguity leads to various models for allocation of computer resources among and within applications. In the simplest setting, a task can be identified as an individual software application, so that a multitasking operating system allocates resources sequentially among individual applications. In a more general context, however, individual programs may possess internal granularity in the form of subprocesses that may execute in parallel within an application. These subprocesses are termed "threads," and operating systems that support multiple threads of internal program execution are termed *multithreaded* operating systems.

Examples of multiple threads of execution include programs that support internally concurrent operations such as printing documents while other work is in progress (where a separate thread is spawned to handle the printing process), displaying graphical results while performing other calculations (where a separate thread can be used to paint the screen as data are read or calculated), or generating reports from within a data base application while other queries are performed. In general, multithreading of individual subtasks within an application will be advantageous whenever spawned threads represent components of the application that are complicated enough so that waiting for them to finish (which would be required in a single-threaded environment) will adversely affect the response of the program.

Multiprocessing. One of the most important advantages of separating a program into multiple threads is that this decomposition of programming function permits individual threads to be shared among different processors. Computers with multiple CPUs have been common platforms for performing high-end engineering computation for over a decade (e.g., multiprocessor supercomputer architectures, such as the Cray X/MP and Cray Y/MP models introduced in the 1980s), but the availability of multiple processing units within a single computer has finally gravitated to the realm of low-end microcomputers. The ability of an operating system to support concurrent execution of different program threads on different processors is termed *multiprocessing*. Multiprocessing occurs in two fundamental flavors:

- Symmetric multiprocessing (SMP), where each individual CPU is capable of executing any process, including threads originating within applications or within operating system services
- Asymmetric multiprocessing (ASMP), where different processors are relegated to different tasks, such as running applications or running operating systems services

Asymmetrical processing is commonly implemented using a dual-CPU architecture involving a master/slave relation between the processing units. The master CPU performs the application and some system services, while the slave CPU is relegated to pure system tasks (such as printing, waiting for slow input/output devices, etc.). Asymmetric multiprocessing architectures provide some speed-up of individual programs, but this increased performance is often limited to reducing the wait time required for some system services. Symmetric multiprocessing can produce substantial gains in program execution speed, as long as individual threads do not contend for resources. The ability of a program (or an operating system) to take advantage of multiple CPU resources is termed *scalability*, and scalable operating systems are well positioned to take advantage of current improvements in available multiprocessing hardware platforms.

Virtual Memory. Providing the extensive memory resources required for most engineering software can be an expensive undertaking. Dynamic Random-Access Memory (DRAM) is too expensive to maintain an appropriate supply for every program used in a multitasking environment. In practice, much of the

memory demand in a multitasking setting can be satisfied by caching some of the blocks of data ostensibly stored in main memory to a fast disk storage subsystem. These blocks of data can be reloaded to main memory only when they are absolutely required, and this practice of *paging* memory to and from the disk is termed *virtual memory management.* In most common implementations of virtual memory, the paging scheme provides a level of independence of memory addressing between processes that is carefully implemented so that one process cannot corrupt the memory of another. Such schemes that implement memory protection to prevent interapplication memory corruption are termed *protected virtual memory management.*

Depending on demand for physical memory, virtual memory schemes may be a great help or a hindrance. While there are sophisticated paging algorithms available that are designed to prevent writing needed to memory to disk, in practice, if there are enough different applications competing for memory, the relative disparity in speed of memory vs. disk subsystems may lead to very sluggish performance for applications whose memory resources have been written to the disk subsystem. In addition, multiprocessing architectures place further constraints on virtual memory performance in order to avoid corruption of memory by different threads running on different CPUs. Modern virtual memory management is an active area of research in computer science, but one empirical rule is still true: perhaps the best way to improve the performance of any virtual memory operating system is to add physical ("real") memory!

Networking and Security. One of the most fundamental shifts in computing over the last decade has been the transition from disconnected individual computers to a distributed computing model characterized by networked workstations that support various remote processing models. Most modern operating systems support standard networking protocols that allow easy integration of different computers into local- and wide-area networks, and also permit sharing of resources among computers. Traditional networking functions (such as sharing files between different computers on the same network) have been augmented to encompass remote computing services, including sharing applications between networked computers (which represents a generalization of symmetric multiprocessing architectures from a single computer to a disparate network of connected computers).

Because of the tremendous pace of changes in the field of computer networking, one of the most important features of any network operating system involves adherence to standard networking protocols. Networking standards provide a portable implementation of networking function that effectively abstracts network operations, allowing existing networking applications to survive current and future changes in networking hardware and software. The most common current networking model is one promulgated by the International Standards Organization and termed the *Open Systems Interconnect* (OSI) reference model. The OSI model uses layers (ranging from low-level hardware to high-level application connections) to idealize networking function. Adherence to the OSI model permits operating systems to become insulated from improvements in networking hardware and software, and thus preserves operating system investment in the face of rapid technological improvements in the field of computer networking.

Once an individual computer is connected to a network, a whole host of security issues arise pertaining to accessibility of data across the network. Secure operating systems must satisfy both internal (local to an individual computer) and global (remote access across a network) constraints to ensure that sensitive data can be protected from users who have no right to access it. Since many mechanical engineering applications involve the use of military secrets, adherence to appropriate security models is an essential component of choosing an operating system for individual and networked computers.

There are many aspects to securing computer resources, including some (such as protected virtual memory schemes) that satisfy other relevant computer needs. In the setting of computer security, operating systems are classified according to criteria developed by the Department of Defense (DOD 5200.28-STD, December 1985). These DOD criteria provide for such features as secure logons (i.e., logging into a computer requires a unique user identifier and password), access control structures (which restrict access to computer resources such as files or volumes), and auditing information (which provides

automated record keeping of security resources so as to help prevent and detect unauthorized attempts at gaining access to secure computer resources).

Portability. Some operating systems (for example, MS/DOS, written in Intel 8080 assembly language) are inextricably tied to the characteristics of a particular hardware platform. Given the rapid pace of development in CPU hardware, tying an operating system to a particular family of processors potentially limits the long-term utility of that operating system. Since operating systems are computer software systems, there is no real obstacle to designing and implementing them in accordance with standard practice in software engineering, and in particular, they can be made portable by writing them in high-level languages whenever possible.

A portable operating system generally abstracts the particular characteristics of the underlying hardware platform by relegating all knowledge of these characteristics to a carefully defined module responsible for managing all of the interaction between the low-level (hardware) layer of the operating system and the overlying systems services that do not need to know precise details of low-level function. The module that abstracts the low-level hardware layer is generally termed a *hardware abstraction layer* (HAL), and the presence of a HAL permits an operating system to be ported to various processors with relative ease. Perhaps the most common portable operating systems are UNIX and Windows NT. Both of these operating systems are commonly used in engineering applications, operate on a wide variety of different CPUs, and are almost entirely written in the procedural C language.

Classification of Representative Operating Systems

Several operating systems commonly encountered in engineering practice are classified below in accordance with the definitions presented above. Note that some of these operating systems are presently disappearing from use, some are new systems incorporating the latest advances in operating system design, and some are in the middle of a potentially long life span.

MS/DOS and Windows 3.x. The MS/DOS (Microsoft Disk Operating System) operating system was introduced in the 1980s as a low-level controlling system for the IBM PC and compatibles. Its architecture is closely tailored to that of the Intel 8080 microprocessor, which has been both an advantage (leading to widespread use) and disadvantage (relying on the 8080's arcane memory addressing scheme has prevented MS/DOS from realizing effective virtual memory schemes appropriate for engineering computation). MS/DOS is a single-processor, single-tasking, single-threaded operating system with no native support for virtual memory, networking, or security. Despite these serious shortcomings, MS/DOS has found wide acceptance, primarily because the operating system is so simple that it can be circumvented to provide new and desirable functions. In particular, the simplicity of MS/DOS provides an operating system with little overhead relative to more complex multitasking environments: such low-overhead operating systems are commonly used in realtime applications in mechanical engineering for such tasks as process control, data acquisition, and manufacturing. In these performance-critical environments, the increased overhead of more complex operating systems is often unwarranted, unnecessary, or counterproductive.

Microsoft Windows is an excellent example of how MS/DOS can be patched and extended to provide useful features that were not originally provided. Windows 3.0 and 3.1 provided the first widely used graphical user-interface for computers using the Intel 80 $\times$ 86 processor family, and the Windows subsystem layers, which run on top of MS/DOS, also provided for some limited forms of cooperative multitasking and virtual memory for MS/DOS users. The combination of MS/DOS and Windows 3.1 was an outstanding marketing success. An estimated 40 million computers eventually ran this combination worldwide. Although this operating system had some serious limitations for many engineering applications, it is widely used in the mechanical engineering community.

VAX/VMS. Another successful nonportable operating system that has found wide use in engineering is VAX/VMS, developed by Dave Cutler at Digital Equipment Corporation (DEC) for the VAX family of minicomputers. VMS (Virtual Memory System) was one of the first commercial 32-bit operating systems

that provided a modern interactive computing environment with features such as multitasking, multithreading, multiprocessing, protected virtual memory management, built-in high-speed networking, and robust security. VMS is closely tied to the characteristics of the DEC VAX microprocessor, which has limited its use beyond that platform (in fact, DEC has created a software emulator for its current family of 64-bit workstations that allows them to run VMS without the actual VAX microprocessor hardware). But the VMS architecture and feature set is widely imitated in many popular newer operating systems, and the flexibility of this operating system was one of the main reasons that DEC VAXs became very popular platforms for midrange engineering computation during the 1980s.

Windows NT. Windows NT is a scalable, portable, multitasking, multithreaded operating system that supports OSI network models, high-level DOD security, and protected virtual memory. The primary architect of Windows NT is Dave Cutler (the architect of VAX/VMS), and there are many architectural similarities between these two systems. Windows NT is an object-oriented operating system that supports the client-server operating system topology, and is presently supported on a wide range of high-performance microprocessors commonly used in engineering applications. Windows NT provides a Windows 3.1 subsystem that runs existing Windows 3.1 applications within a more robust and crash-proof computational environment, but NT also provides other user interfaces, including a console interface for textual applications ported from mainframes and MS/DOS, a UNIX-like graphical user interface (provided by implementing common UNIX window management functions on top of NT), and the Macintosh-like interface similar to that introduced in Windows 95 as a replacement for Windows 3.1 in mid-1995.

UNIX. The UNIX operating system was developed during the 1970s at Bell Laboratories to satisfy the need for a flexible and inexpensive operating system that would provide high-end system services (e.g., multitasking, virtual memory) on low-cost computers. Since its initial inception, the UNIX operating system has evolved to become one of the most successful operating environments in history. UNIX provides preemptive multitasking, multithreading (threads are termed “lightweight processes” in most implementations of UNIX), multiprocessing, scalability, protected virtual memory management, and built-in networking. Although there are a variety of competing UNIX implementations, substantial standardization of the UNIX operating system has occurred under the auspices of the Open Software Foundation (OSF), a consortium of computer companies that includes IBM, DEC, and Hewlett-Packard. OSF UNIX is based on IBM’s AIX UNIX implementation and represents one of the most advanced operating systems available today. Another important emerging standard for UNIX is the public-domain version termed Linux: this UNIX variation runs on a wide range of computers, is freely distributed, and has an incredibly diverse feature set, thanks to the legions of programmers around the world who have dedicated their skills to its development, extension, and support. Various versions of UNIX run on virtually every type of computer available, ranging from inexpensive microcomputers through engineering workstations to expensive supercomputers. The ability of the UNIX system to evolve and retain a dominant market position over two decades is a concrete testimonial to the advantages of strict adherence to the principles of software engineering, because nearly all aspects of the UNIX operating system have been designed and implemented according to these principles. In fact, much of the history of software engineering is inextricably bound up with the history of the UNIX operating system.

Parallel Computation

The use of multiple processors and specialized hardware to speed up large-scale calculations has a rich history in engineering computation. Many early mainframe computers of the 1970s and most supercomputers of the 1980s used specialized hardware whose design was influenced by the nature of engineering computation. Vector processors, gather/scatter hardware, and coarse-grain parallel CPU architectures have been used successfully over the past few decades to increase the performance of large-scale computers used in engineering computation. Currently, most of the hardware advances of these past large computers have migrated to the desktop, where they are readily available on microcomputers and

workstations. Understanding the basic principles of these advanced computer architectures is essential to gain efficient utilization of their advantages, and so the following sections present an overview of the fundamentals of this important field.

Technical Overview of Parallel and Vectorized Computation

Parallelism in computer hardware can occur on many levels. The most obvious example was addressed in the setting of operating system services, where scalability over multiple CPUs was presented as a means for a computer's operating system to utilize additional CPU resources. Parallelism achieved by adding additional CPUs to a computer commonly occurs in two variations: a coarse-grained parallelism characterized by a relatively small number of independent CPUs (e.g., typically from 2 to 16 CPUs), and a fine-grained parallelism commonly implemented with substantial numbers of CPUs (typically, from a minimum of around 64 up to many thousands). The former is termed *symmetric multiprocessing*, or SMP, and the latter is referred to as *massively parallel* (MP) computing. Each is frequently encountered in practical engineering computation, although SMP is much more common due to its lower cost and relative simplicity.

It is also possible for parallelization to occur within an individual CPU. On specialized mainframes with attached vector processors (and within the CPUs of most supercomputers), various different machine instructions can be pipelined so that more than one instruction occurs in parallel in a particular arithmetic unit. The practical effect of this internal parallelization in instruction execution is that many common arithmetic operations (such as the multiplication of a vector by a scalar) can be performed by carefully arranging the pipelined calculations to permit impressive performance relative to the computational effort required on a nonpipelined CPU. One of these forms of internal parallelism within the CPU (or attached processor) is termed *vectorization*, as it permits vector operations (i.e., those associated with a list, or vector, of floating-point numbers that are stored contiguously in memory) to be processed at very fast rates.

Such pipelined execution characteristics have now become commonplace even on low-cost microcomputers. In fact, current high-performance microprocessors used in engineering workstations and high-performance microcomputers generally have multiple independent arithmetic units that can operate in parallel. Such internally redundant designs are called *superscalar* architectures and allow the CPU to execute more than one instruction per clock cycle. The cumulative effect of such multiple independent pipelined arithmetic units operating in parallel within an individual CPU is that current microprocessors exhibit astonishing performance on typical engineering problems when compared to large (and expensive) central computers of the 1980s. Engineering problems that required dedicated vector processors in the 1980s are commonly executed faster on low-priced microcomputers today. In addition to current hardware eclipsing older vector processors in performance levels, modern computer language compilers are now commonly tuned to particular microprocessor characteristics in order to provide the same sort of computational performance formerly associated with specialized vectorizing compilers (i.e., compilers that could recognize common vector operations and generate appropriately efficient machine instructions).

Finally, another important form of parallelism has developed in conjunction with high-performance networking schemes. Distributed computing applications are commonly designed to parallelize calculations by dividing up individual threads of execution among disparate computers connected via a high-speed network. While this sort of distributed computing was sometime encountered in the 1980s on high-priced computers (such as DEC VAX clusters, which transparently balanced computational loads over a collection of networked minicomputers), similar distributed computing schemes are now becoming commonplace on microcomputers and workstations connected over local-area networks.

Classification of Parallel Architectures

There are many schemes to characterize parallel computer architectures, including the SMP/MP classification given above. Since computer programs consist of instructions and data, it is possible to further classify parallelization schemes by considering the redundancy (or lack thereof) in these components.

The four possible classifications are single instruction/single data (SISD); single instruction/multiple data (SIMD); multiple instruction/single data (MISD); and multiple instruction/multiple data (MIMD). Of these four, the first pertains to nonparallel computation (such as a standard computer with a single processor). The others include representations of practical schemes for MP (massively parallel) computing, SMP (symmetric multiprocessing) computation, and networked distributed computing.

SIMD computers are particularly simple representatives of massively parallel systems. In order to run at maximum speed, each CPU in a SIMD architecture has to execute the same instructions as its neighbors (here "neighbors" is a flexible term that represents various topological arrangements among small groups of processors). The result is a massively parallel computer where individual processors act in unison: every instruction is executed by many processors, each with its own local data. If the underlying algorithm used can be constrained to fit into this SIMD architecture, the performance obtained can be phenomenal. Many standard numerical schemes were already well organized for SIMD calculations: for example, simple two-dimensional finite-difference calculations for heat conduction involve replacing the value at a node with the average of the node's four neighbors in a north/west/south/east pattern. This simple calculation is relatively easy to implement on a SIMD computer, and so calculations such as these finite-difference molecules result in highly scalable performance. More complex calculations, such as those found in modeling nonlinear problems, are generally much more difficult to implement on SIMD computers and often result in poor performance on these simplified computer architectures.

MISD computers are commonplace today, both in the form of supercomputers and in SMP desktop workstations. In each case, a small set of individual processors shares memory (single data), and each processor operates more or less independently of the others (multiple instructions). This form of parallelism has been aided by operating systems (such as current flavors of UNIX and Windows NT) that support multithreading, allowing the programmer (or the operating system) to distribute threads among the various processors in a typical SMP environment. Given appropriate programmer and operating system support, MISD computers can be highly scalable, so that adding additional processors results in associated decreases in execution time. There are substantial obstacles of increased performance in a MISD environment, including the important issue of contention by different threads for common resources. An example of resource contention occurs when two different processors attempt to access a shared memory address simultaneously: some form of signaling and locking mechanism must be provided to insure that more than one processor cannot simultaneously modify memory. Currently, many vendors offer hardware-based support for resolving contention (and related bottlenecks) in symmetric multiprocessing, and standards are currently evolving in this important area of support.

Probably the most common MIMD example of current interest is in distributed computing performed on networked computers. Since each computer has its own local data and instruction stream (MIMD), this approach combines many of the best features of single-computer implementations of both symmetric multiprocessing and massively parallel architectures. While networks of computers have been used for over a decade in solving many easily parallelized applications (such as classical mathematical problems arising in number theory), using such distributed computer networks to perform general-purpose engineering computations is a more recent development. MIMD networked computation requires a host of instruction synchronization and data replication issues to be resolved (these are the network equivalent of the resource contention problems of SMP architectures), but substantial progress is underway in addressing these bottlenecks to distributed computing. The use of object-oriented models for distributed computing, which permit hiding many of the details of the distribution process via appropriate process abstraction, appears to be an especially important avenue toward the efficient use of distributed networked computing.

Computer Graphics and Visualization

Mechanical engineering is one of the most important sources of applications in computer graphics, and mechanical engineers form a significant market for computer-graphics software. Many of the most important developments in the history of computer graphics, such as the development of spline models

for realistic surface representations, were originally motivated by the particular needs of the mechanical engineering profession. Current topics of importance to researchers in computer graphics, such as the applications of scientific visualization or the use of rational physical-based models in computer animation, are also motivated in large part by the diverse needs or the knowledge base of the mechanical engineering field.

Technical Overview of Computer Graphics

The fundamental motivation for computer graphics and visualization arises from the adage that "a picture is worth a thousand words." The human visual sense is the only sensory apparatus with sufficient bandwidth (i.e., information-carrying capacity) to permit rapid evaluation of the large data sets characteristically associated with problems in science and engineering. Mechanical engineers have historically been aware of the importance of using graphics, and college students in this field have traditionally been required to study engineering graphics as a required course during their first-year programs of study. The commonly cited observation that "graphics is an engineer's second language" is pertinent today, especially in light of the importance of television and other visual arts during the formative years of younger engineers.

There is a rich nomenclature associated with the field of computer graphics (see Foley and VanDam, 1982, for a detailed overview). Some high-level terms commonly used in the field should be defined before more detailed exposition of this important field is attempted; the definitions given below are not intended to be all-inclusive, but instead to be concise enough to permit further review of this topic.

- Visualization: the study of applying computer graphics toward the goal of displaying collections of data in a relevant and informative manner
- Rendering: converting a mathematical model of a scene geometry (or a collection of data values) into a visually meaningful image on the computer
- Virtual reality: technology permitting simultaneous display and control of graphical simulations so that the user interprets the display as the existence of an alternative reality
- Multimedia: the integration of senses besides the visual sense into a computer application or simulation
- Graphical user-interface: a metaphor for human-computer interaction using standardized graphical mechanisms for control, input, and output

One of the main reasons that rapid progress has occurred in computer graphics is that practitioners in the field have learned to leverage the efforts of others by adhering to industry-wide standards. The code of standardization and practice in computer graphics exists on many levels, ranging from low-level coordinate choices developed as a foundation toward abstracting the rendering and display process to high-level format standardizations required for the portable use of multimedia content. The breadth of standardization in computer graphics is beyond the scope of this handbook, but relevant information can be found in the publications of the ACM/SIGGRAPH. This organization is the primary instrument of dissemination and standardization efforts in the computer graphics industry.

Visualization Methods in Mechanical Engineering

The fundamental problem of visualization in science and engineering is the conversion of raw data into an informative visual display. The data may arise from a closed-form mathematical representation, or it may be obtained as an organized collection of data presented in tabular or related formats. The most important initial step in engineering visualization is the recognition of the domain (i.e., the set of input values) and range (i.e., the set of output values) of the data to be represented, so that an appropriate visual display can be synthesized. For example, if the range of the data includes a temporal component, then a time-dependent display scheme such as an animation is often warranted. If the domain is a physical region in space (such as a mechanical object being analyzed), then a common visualization scheme consists of mapping that region onto the display surface as a background for visual results. The details of the particular display choice depend upon the range of the output data: for display of scalars (such

as temperature, or specific components of stress), appropriate visualization schemes include contour maps using lines or color-filled contours, as shown in Figure 15.2.7 (in two dimensions) and in Figure 15.2.8 (in three dimensions, for the same general stress analysis problem). When rendering three-dimensional results as in Figure 15.2.8, it is necessary to provide for removal of hidden surfaces from the display.

For display of vectors, the use of arrows aligned with the local vector field is a common means for displaying this more complex type of range: an example is shown in Figure 15.2.9.

In order to aid in the understanding of complicated visualizations, it is often useful to provide visual cues to aid the viewer in understanding the display presented. In Figure 15.2.10, a contour plot of pressure within a heater fan is shown (at high Reynolds' number), with an inset picture showing the geometry of the fan used in the computational simulation. The combination of virtual display (the pressure contours) and physical display (the fan assembly) aids the viewer in understanding the model geometry and solution response. (Figures 15.2.9 and 15.2.10 are included courtesy of Advanced Scientific Computing Corporation of El Dorado Hills, CA.)

Another useful approach to provide contextual cues to aid in understanding a complex visualization is to overlay the graphical image on top of a physical setting associated with the simulation. For example, geographical simulations (such as models of contaminant transport used for air pollution modeling) can be overlayed on a map of the geographical area under study: viewers will instantly grasp the problem domain in terms of landmarks such as rivers, roads, mountains, and other obvious visual cues. An alternative approach is shown in Figure 15.2.11, where a visualization of a complex reactor simulation is overlayed on a picture of the reactor itself. In this figure, the color contours plot gas temperatures in a model of a chemically reacting Navier-Stokes simulation used for modeling vapor deposition processes occurring in semiconductor manufacture. (The simulation and visualization were both performed at Sandia National Laboratories in Livermore, CA.)

When time-dependent results are to be visualized, there are two standard schemes commonly used. The first approach involves treating the time-dependent behavior of the data as another spatial dimension. This approach is commonly used to graph scalar-valued functions of time t as $y = y(t)$, with the function's variation plotted on the vertical axis and the time on the horizontal axis. It can also be used to draw more complex functions such as the one shown in Figure 15.2.12, where a scalar-valued function of two variables $z = z(x, t)$ is plotted as a three-dimensional surface (in this example, the axial displacement of a bar as a wave moves along its length). Note that the surface is drawn in perspective to aid in the perception of three-dimensional behavior even though it is plotted on a two-dimensional page. Using such depth cues as perspective (as well as others, including hidden-surface removal and shading and realistic illumination models) to aid in realistic display of computer images is the essence of the rendering process.

The other standard approach for visualizing time-dependent data is to treat the temporal variation of the data in a natural form by animating the history of the display. Standard animation techniques were formerly beyond the range of cost and effort for most engineering purposes, as computer-generated animation workstations of the 1980s typically cost over $100,000 and required considerable dedicated expertise for operation. Currently, microcomputers have become excellent platforms for animation and visualization, and appropriate computer programs (known as nonlinear digital video editing software) for generating and composing animations from engineering simulations have become popular on inexpensive microcomputers and low-end engineering workstations.

Multimedia in Mechanical Engineering Practice

Multimedia is an imprecise term whose definition has evolved over the recent past as the performance of computers has improved substantially. Formerly, multimedia was taken to mean integration of different types of display (such as a textual computer used to control a graphical device for display of archived images). The current context of the term includes multiple senses, such as the integration of video and sound, or the control of a visual display by a tactile three-dimensional graphical input device. The former flavor of multimedia naturally draws the analogy with silent movies being replaced by those with a

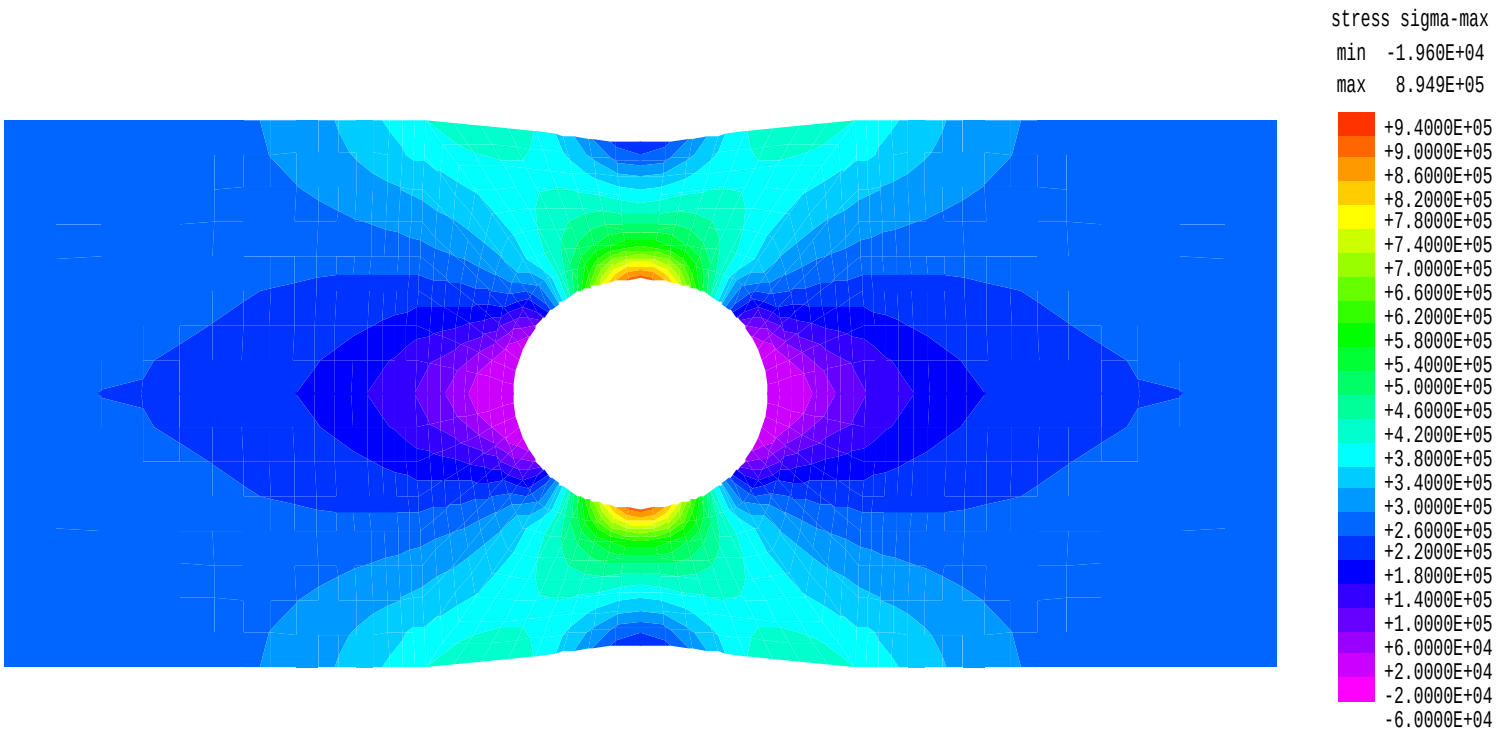


FIGURE 15.2.7 Two-dimensional contour plot for stress analysis.

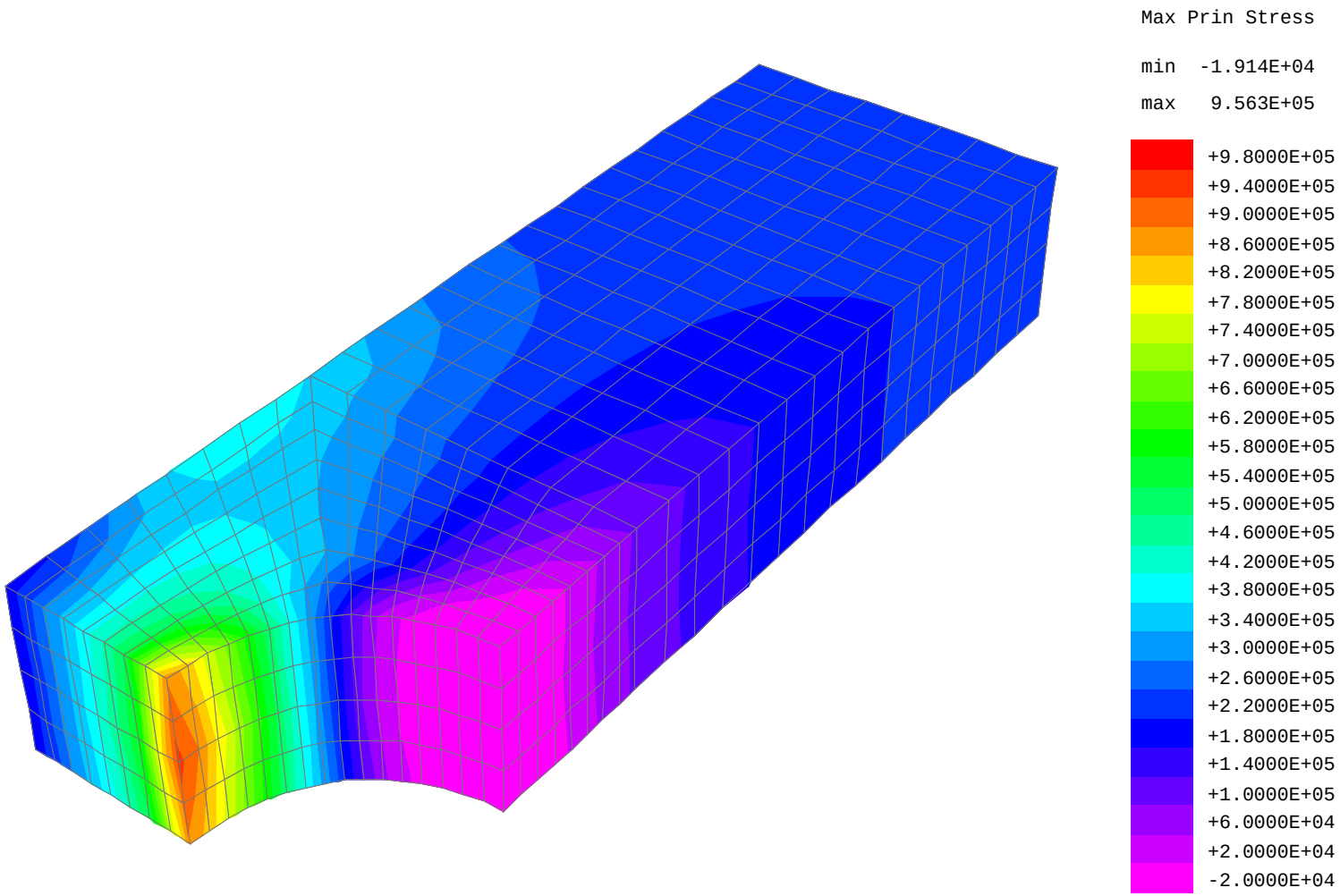


FIGURE 15.2.8 Three-dimensional contour plot for stress analysis.

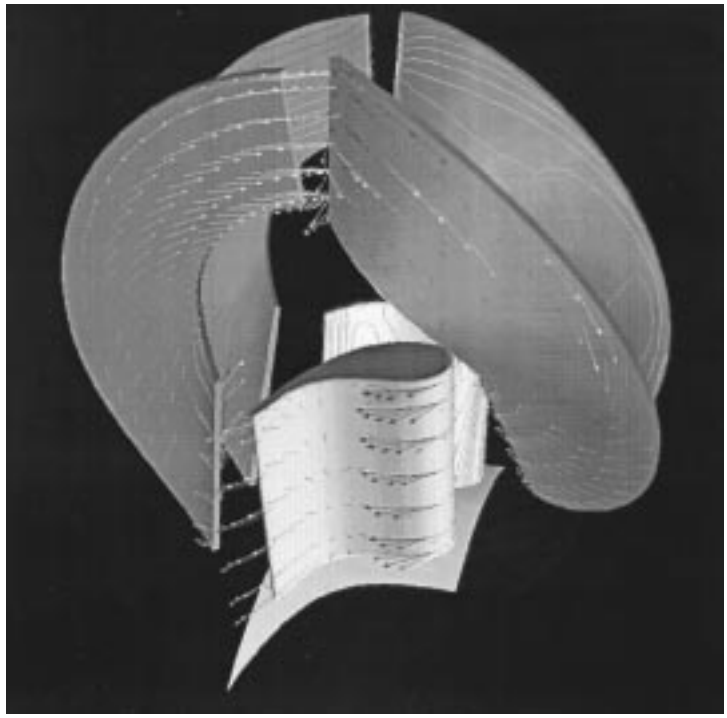


FIGURE 15.2.9 Three-dimensional vector field plot for torque converter simulation.

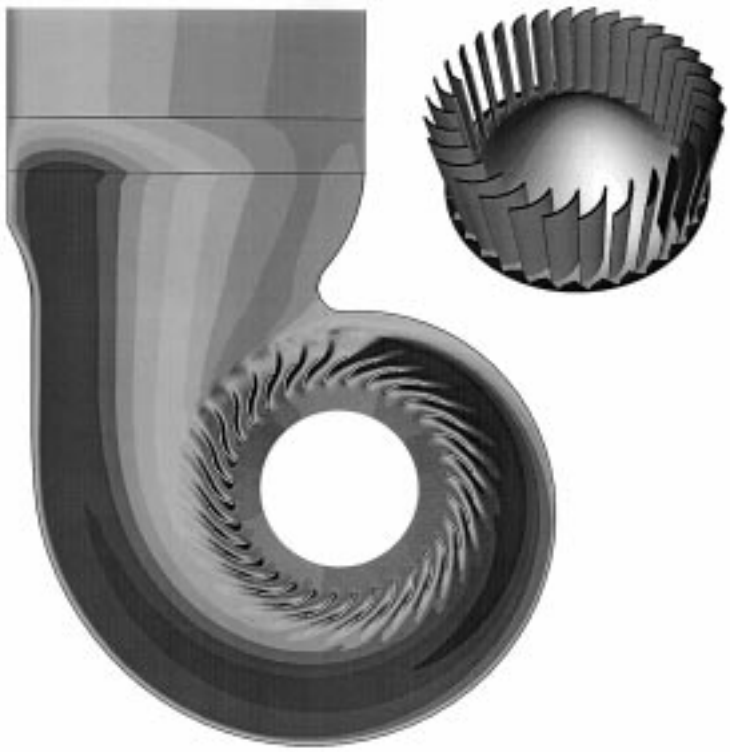


FIGURE 15.2.10 Pressure contour plot with solution geometry cues.

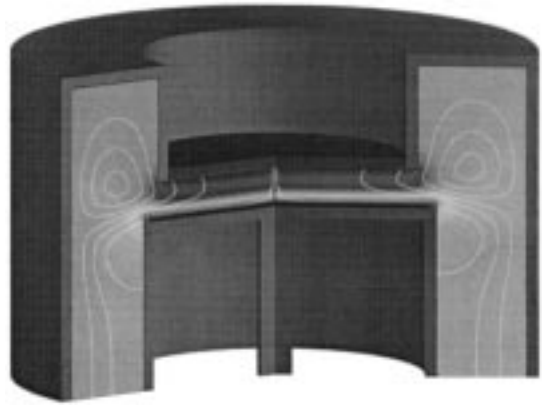


FIGURE 15.2.11 Gas temperature contour plot with reactor background.

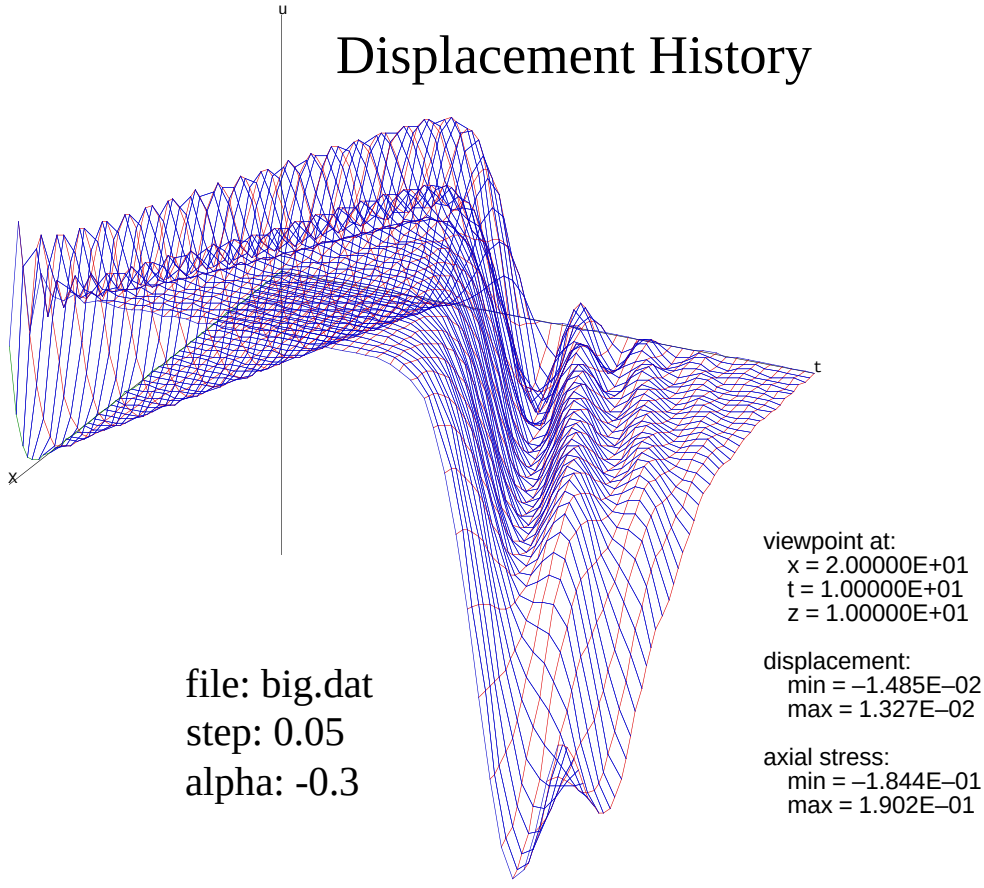


FIGURE 15.2.12 Time-dependent data viewed as space-time surface plot.

sound track: that technological development revolutionized the movie industry in a manner similar to the predicted impact of multimedia on computing practice. The latter example is widely used in virtual reality applications.

Multimedia is presently used in education, training, marketing, and dissemination efforts in the engineering profession (see Keyes [1994] for a detailed exposition of the application of multimedia to all of these fields). Each of these uses is based on two empirical observations: (1) people generally find multimedia presentation of data more interesting, and this increased interest level translates into better retention of data presented; and (2) good use of visualization principles in multimedia presentation provides a better audience understanding of the material presented. The promise of simultaneous increased understanding and improved retention is a strong motivation for using multimedia in the presentation of information.

There are considerable technical issues to address in efficient use of multimedia technology. The most important technical problem is the sheer volume of stored data required to make animations, which are a mainstay of multimedia presentations. A single high-quality color computer workstation image (e.g., one frame of an animation that would typically be displayed at 24 to 30 frames per second to convey the sense of continuous motion) requires from approximately 1 to 4 million bytes of storage. Generating full-screen animations therefore requires anywhere from a few dozen to over a hundred megabytes of data to be processed and displayed every second. Such video bandwidths are rare on computers, and even if they were available, the demand for storage would soon outstrip available supply, as each 10 sec of video potentially requires up to a gigabyte of high-speed storage. Storing sound requires less dramatic capacities, but even low-quality (e.g., AM-radio quality) audio storage requires substantial storage for long durations of audio.

The practical way to overcome these technical limitations to storage and display bandwidth is to develop efficient ways to compress visual and aural redundancies out of the data stream on storage, and then decompress the archived data for editing or display. Such COmpression/DECompression schemes (termed CODECs) are generally used to support multimedia applications ranging from digital animation preparation on microcomputers to video teleconferencing applications. Common CODECs are tailored for individual applications and auxiliary technology (for example, playing video from a CD/ROM device generally requires high compression ratios in order to accommodate the relatively poor data transfer rates obtained with CD/ROM technology), and their performance is measured in terms of compression ratio, which is the ratio of uncompressed image size to compressed image size.

Various CODECs in common use include:

- JPEG (Joint Photographic Experts Group) CODEC. JPEG is a CODEC originally developed for efficient storage of individual images, such as photographic libraries. JPEG is a lossy algorithm (i.e., some information may be lost during compression), but JPEG is so carefully tuned to human perceptual characteristics that it is feasible to achieve compression ratios on the order of 10:1 or more without being able to detect any losses, even after repeated recompressions. Higher compression ratios (on the order of 100:1) can be achieved, but these more efficient rates may introduce visible artifacts of the compression process. JPEG is generally only feasible for animation applications through the use of specialized hardware, but the same hardware can be used for both compression and decompression tasks. JPEG is a very commonly used CODEC in animation applications, as the required hardware is relatively inexpensive, the performance (e.g., speed, image quality) is acceptable, and the algorithm is easy to implement within animation applications (e.g., digital video editing software).

- MPEG (Motion Pictures Experts Group) CODEC. MPEG is a family of CODECs that is becoming popular for display-only animation applications. The various MPEG implementations (MPEG and MPEG-2) require prodigious amounts of computational effort for the compression step, so they are not presently feasible for inexpensive video capture applications. MPEG CODECs routinely provide compression ratios approaching 100:1 with relatively little degradation of image quality. This compression efficiency permits very large and complex animations to be stored and played back using low-technology devices such as CD/ROMs. The fundamental difference between MPEG and JPEG schemes is that with JPEG,each video frame is compressed independently of all the other frames, where MPEG uses adjoining frame information by concentrating on the *differences* between frames to compress the overall animation more efficiently (most animations have only small temporal differences among adjacent frames, and this time-dependent redundancy is ignored by the JPEG algorithm and carefully exploited by the MPEG family). It is expected that the near-term will see many inexpensive implementations of MPEG CODECs for microcomputers, and many proposed video storage standards (e.g., new videodisc technology designed to replace existing VHS video playback devices) rely on MPEG compression schemes for efficient storage of movies on CD/ROMs.

Another important technical issue arising in multimedia applications is the synchronization of time-dependent data streams. For example, multimedia applications using both video and sound must guarantee synchronization of these two data streams in order to maintain audience interest. Probably the most important synchronization method is that provided by the Society of Motion Picture and Television Engineers (SMPTE). SMPTE synchronization depends on a time-code that numbers every frame of video. This time-code information can be used to synchronize audio information to the video source; in fact, this is exactly how commercial movie soundtracks are synchronized with the motion picture. Other time-code mechanisms include RC time-code, which is a frame-indexing scheme that is available on many consumer-grade video products, including camcorders, video recorders, and video editing equipment. RC provides the same sort of synchronization as SMPTE, but with fewer features and at a reduced cost.

Standard microcomputer multimedia programming interfaces generally provide synchronization mechanisms for simultaneous audio and video display. Both Microsoft's Video for Windows system software (used under Windows and Windows NT operating systems) and Apple's QuickTime software (available for Macintoshes, Windows, Windows NT, and UNIX workstations) provide time-code information pertinent for simple synchronization of multimedia. In addition, current versions of QuickTime also provide SMPTE time-code support, in order to simplify the task of high-performance digital video editing on the Macintosh platform.

Graphical User-Interface Design and Implementation

The marriage of human perceptual principles from applied psychology with the technology of interactive computer graphics resulted in the development of graphical user-interfaces (GUIs). GUIs have since evolved into one of the most revolutionary ideas in the history of computing. The basic idea behind a GUI is that visualization can be used to provide a powerful abstraction for the interaction between human and computer, and that basing that graphical abstraction on principles of human perception and ergonomics will result in increased user productivity. Most of the early work on development of GUIs occurred at the Xerox Palo Alto Research Center in California, and the fundamental ideas developed there were first implemented in a consumer setting when Apple Computer released the Macintosh in 1984. Since then, virtually every major vendor of computer has adopted a GUI similar in appearance to that of the Macintosh, and it is clear that using a well-designed GUI yields substantial gains in productivity for most computer-assisted tasks.

The fundamental rules for implementing a GUI within a computer program are simple: the program's interface must be intuitive, consistent, permissive, and responsive (see Apple [1985] for a well-written introduction relating to the widely imitated Macintosh user-interface guidelines). An intuitive program

interface is easy to understand and responds naturally to a user's actions. Implementing an intuitive interface is a similar task to good industrial design, where principles of cognitive ergonomics are used to insure that controls of manufactured products (such as household appliances) are appropriately easy to use. The classic example of an appliance with an intuitive user interface is a toaster (in fact, one of the design goals of the Macintosh development process was to design a computer that would be as easy to use as a toaster), as the controls are naturally placed and behave intuitively. Intuition in computer applications thus implies natural composition, arrangement, and performance of graphical controls, such as buttons to be pressed for execution of commands.

The role of consistency is essential to the goal of improving the productivity of computer users. Until the widespread use of GUIs, most application programs used individual and unique schemes for such mundane tasks as saving files, printing documents, or quitting the program. Unfortunately, there was little standardization of commands for performing these common functions, so the user was faced with a substantial learning curve for each program encountered. With the advent of consistent graphical user-interfaces, the effort of learning most standard commands needs to be learned only once; all other applications will use the same mechanism for performing these common functions. The cost of training is thus primarily amortized over the first program learned, and the resulting productivity gains are often substantial. While the goal of complete consistency between applications (and especially between different types of GUI) has not yet been realized, there has been considerable progress toward this ultimate goal.

Permissiveness places the user in control of the computer, instead of the opposite situation, which was routinely encountered in older textual interfaces. The user should decide what actions to do and when to do them, subject only to the feasibility of those actions. Permissiveness allows users to structure their workflow in a personalized manner, which allows them to seek and find the best individualized way to accomplish computer-assisted tasks. Permissiveness is one of the most difficult attributes for a computer programmer to implement, as it requires enabling and testing a myriad of different execution pathways in order to guarantee that all feasible approaches to using the computer program have been insured to work reliably. The quality control processes of permissive GUI-based programs are often automated (using software that generates random input processes to simulate a variety of user actions) in order to test unexpected command sequences.

Responsiveness is another essential element of ensuring that the user feels in control of the software. In this sense, a responsive GUI-based program provides some appropriate feedback (generally graphical) for every legitimate user-initiated action. This feedback may be as simple as a "beep" when an improper command is issued, or as complicated as rendering a complex picture in response to user-supplied input data. Responsiveness and intuition work together in the control of the program's actions; for example, the pull-down menu command procedures of most current GUI programs permits the user to see clearly (typically, via highlighting of the currently selected menu item) which of a series of intuitive commands are about to be initiated.

Writing good GUI-based software is a difficult process and is especially cumbersome when procedural languages are used. Modern event-driven interactive software is naturally suited to the more asynchronous message-driven architecture available with object-oriented programming models, and it is in the area of GUI-based applications that object-oriented languages have begun to corner the development market. Since managing the GUI requires so much programmer effort (and because the goal of consistency requires that all such code behave in a similar manner), there is a considerable incentive to reuse code as much as possible in this task. The use of class libraries designed for implementing a standard GUI model is a highly effective way to develop and maintain graphical interactive computer programs, and this use is one of the most common applications of class libraries in software engineering.

There are many standard GUI elements, including:

- Menus for permitting the user to initiate commands and exert program control
- Dialogs for field-based input of data by the user

- Windows for graphical display (output) of program data
- Controls (such as buttons or scroll bars) for auxiliary control of the program

Virtual Reality Applications in Engineering

Virtual reality is one of fastest-growing aspects of computer graphics. Efficient and realistic virtual reality applications require incredible amounts of computational effort, so advances in this area have been limited in the past, but the increased computational performance available with current CPUs (and with better support for multiprocessing and parallelization of effort) has made virtual reality solutions feasible for many engineering problems. The primary motivation for virtual reality is the same one as for computer graphics: using the body's senses (and especially the sense of sight) is a very productive way to gain insight into large amounts of data or for simulating difficult (e.g., life-threatening) situations.

There are two common technological approaches to virtual reality: *windowed* virtual reality applications, in which the view into the virtual world is represented by a conventional planar graphical computer display, and *immersive* virtual reality which requires specialized hardware (such as display headsets) to convey a sense of being immersed in the artificial virtual world. While windowed virtual reality applications are common (running on platforms ranging from high-performance computers with color displays to mass-market dedicated computer game machines designed for use with television sets), immersive virtual reality applications are still fairly expensive and rare. In addition, there is considerable anecdotal evidence that attempting to perform immersive virtual reality with equipment insufficiently powerful to update the display instantaneously (i.e., within the interval of around 1/30 sec associated with the persistence of vision) may induce nausea and disorientation in many users.

There are many important types of virtual reality applications in use in engineering and related fields. *Walkthrough* (or perambulatory) virtual reality programs permit the user to move around in the virtual world, where either easily recognizable images (such as architectural renderings of virtual construction) or completely virtual images (such as visualization displays involving surface plots, color contours, streamline realizations, and other standard visualization schemes updated in realtime) are encountered and manipulated. Walkthrough virtual reality applications are typically used to gain new perspective on simulations or proposed constructions. The primary control ceded to the user is the power to move through the virtual world, and perhaps to modify the type of display encountered.

Synthetic experience is another important component of the application of virtual reality. The classic example of this form of virtual reality application is a flight simulator, where an integrated display and a movable cockpit are used to simulate the "look and feel" of flight. Synthetic experience applications permit the user to manipulate the virtual world in exactly the same manner as the real world, and these programs are used for education, training, and professional practice. The basic motivation for this class of virtual reality applications is to permit the user to practice and learn techniques that would be otherwise too expensive or too dangerous to perform by "real" methods as opposed to virtual practice. This approach has been used extensively in military applications (e.g., battle simulators) and is gaining wide use in medicine and other professional fields (e.g., virtual surgery, where surgeons can practice surgical techniques on the computer or even perform real surgery from a remote "virtual" location).

15.3 Computational Mechanics

Computational mechanics is the art and science of simulating mechanics problems on the computer. The rapid development of computational mechanics is an important component of the foundation of the today's high-technology world. Current developments in aerospace, automotive, thermal, biomedical, and electromagnetic engineering applications are in large part due to the success of computational mechanics researchers in finding efficient schemes to simulate physical processes on the computer. Understanding the strengths and weaknesses of the current computational mechanics field is an important task for the practicing mechanical engineer interested in using computer-aided engineering methods.

Computational approximations for mechanics problems are generally constructed by discretizing the domain of the problem (i.e., the underlying mechanical system) into smaller components. Ideally, in the limit as the size of each of these individual discretized components becomes negligibly small, the computational approximation obtained becomes exact. In practice, there is a lower limit on how small a discretization can be used, as round-off error (caused by truncation of numerical representations used on computers) and requisite computational effort eventually conspire to produce inaccurate or excessively expensive results. Other alternative approximation schemes, including semi-analytical methods such as truncated Fourier series representations, are also commonly used to provide numerical solutions for engineering problems. These nondiscretized methods require letting the number of solution functions (such as the trigonometric functions used in Fourier series) approach infinity to gain convergent results and are thus superficially similar to discretization schemes such as finite-element and finite-difference methods, where the number of grid points tends to infinity as the size of the grid spacing decreases.

Computational Solid, Fluid, and Thermal Problems

Before specific computational approximation schemes are presented, it is enlightening to enumerate the wide variety of problems encountered in mechanical engineering practice. In many cases, each individual subdiscipline of mechanics (e.g., solid mechanics) uses different modeling techniques to obtain computational approximations, so that reviewing the various branches of mechanics also serves to foreshadow the spectrum of numerical schemes used in the different branches of computational mechanics.

Solid Mechanics

Solid mechanics is one of oldest and most well-understood branches of mechanics, and its associated computational schemes are similarly well established. The most common application of solid mechanics is *stress analysis*, where the mechanical deformation of a body is studied to determine its induced state of stress and strain. The study of stress analysis formerly included considerable emphasis on construction and instrumentation of physical models using strain gauges and photoelastic coatings, but modern computational schemes for modeling deformation of solids have rendered these traditional laboratory techniques nearly extinct. The most currently popular stress analysis methods are those based on computational approximations using *finite-element* models. This family of computer simulation techniques is one of the most successful branches of modern computational mechanics, and practical implementations of finite-element methods represent a unified computational model for nearly the entire range of solid mechanical response.

Solids are modeled by appealing to three classes of mathematical relationships: equilibrium, compatibility, and stress-strain laws. Equilibrium statements represent mathematical expressions of conservation laws and are generally expressed in terms of the conservation of linear and angular momentum. Compatibility conditions include the kinematical strain-displacement relations of mechanics, where strains are expressed as spatial derivatives of the displacement field. Stress-strain laws are also termed constitutive relations, because they idealize the constitution of the material in order to relate the interplay of stresses and strains in a deformed body. The fundamental mathematical relations for the dynamic response of a three-dimensional isotropic linear-elastic solid body are given below — note that these equations are linear, in that they involve no products of the solutions components of displacement, stress, or strain.

This linearity arises from the assumptions of small deformations, a linear stress-strain law, and the fact that the reference frame for these equations is attached to the solid body. The linear nature of these relations makes them relatively easy to solve for a wide range of problems, and this simplicity of analysis is an important reason why stress analyses based on these relations is such an important part of computational mechanics.

Definitions for Solid Mechanics Equations

$u, v, w = x, y, z$ components of displacement

$\ddot{u}, \ddot{v}, \ddot{w} = x, y, z$ components of acceleration

$\sigma_x, \sigma_y, \sigma_z, \tau_{xy}, \tau_{yz}, \tau_{zx}$ = stress components

$\varepsilon_x, \varepsilon_y, \varepsilon_z, \gamma_{xy}, \gamma_{yz}, \gamma_{zx}$ = engineering strains

g_x, g_y, g_z = gravity body force components

E = Young's Modulus

ν = Poisson's Ratio

ρ = Density (15.3.1)

"Dynamic" Equilibrium Equations for a Solid

$$\rho\ddot{u} = \frac{\partial\sigma_x}{\partial x} + \frac{\partial\tau_{xy}}{\partial y} + \frac{\partial\tau_{zx}}{\partial z} + \rho g_x$$

$$\rho\ddot{v} = \frac{\partial\tau_{xy}}{\partial x} + \frac{\partial\sigma_y}{\partial y} + \frac{\partial\tau_{yz}}{\partial z} + \rho g_y \qquad (15.3.2)$$

$$\rho\ddot{w} = \frac{\partial\tau_{zx}}{\partial x} + \frac{\partial\tau_{yz}}{\partial y} + \frac{\partial\sigma_z}{\partial z} + \rho g_x$$

Compatibility Equations for a Solid

$$\varepsilon_x = \frac{\partial u}{\partial x} \qquad \gamma_{xy} = \frac{1}{2}\left(\frac{\partial v}{\partial x} + \frac{\partial u}{\partial y}\right)$$

$$\varepsilon_y = \frac{\partial v}{\partial x} \qquad \gamma_{yz} = \frac{1}{2}\left(\frac{\partial w}{\partial y} + \frac{\partial v}{\partial z}\right) \qquad (15.3.3)$$

$$\varepsilon_z = \frac{\partial w}{\partial z} \qquad \gamma_{zx} = \frac{1}{2}\left(\frac{\partial u}{\partial z} + \frac{\partial w}{\partial x}\right)$$

Isotropic Linear-Elastic Stress-Strain Relations

$$\varepsilon_x = \frac{1}{E}\left[\sigma_x - \nu\left(\sigma_y + \sigma_z\right)\right] \qquad \gamma_{xy} = \frac{2(1+\nu)}{E}\tau_{xy}$$

$$\varepsilon_y = \frac{1}{E}\left[\sigma_y - \nu\left(\sigma_z + \sigma_x\right)\right] \qquad \gamma_{yz} = \frac{2(1+\nu)}{E}\tau_{yz} \tag{15.3.4}$$

$$\varepsilon_z = \frac{1}{E}\left[\sigma_z - \nu\left(\sigma_x + \sigma_y\right)\right] \qquad \gamma_{zx} = \frac{2(1+\nu)}{E}\tau_{zx}$$

Solids are commonly distinguished from liquids by the ability of solids to maintain their shape without restraint due to enclosing containers. Because of this particular form of mechanical strength, solids provide many simple structural approximations that naturally arise as degenerations of full three-dimensional solid mechanical response. These idealizations of solid mechanical behavior are diagrammed in Figure 15.3.1 and include:

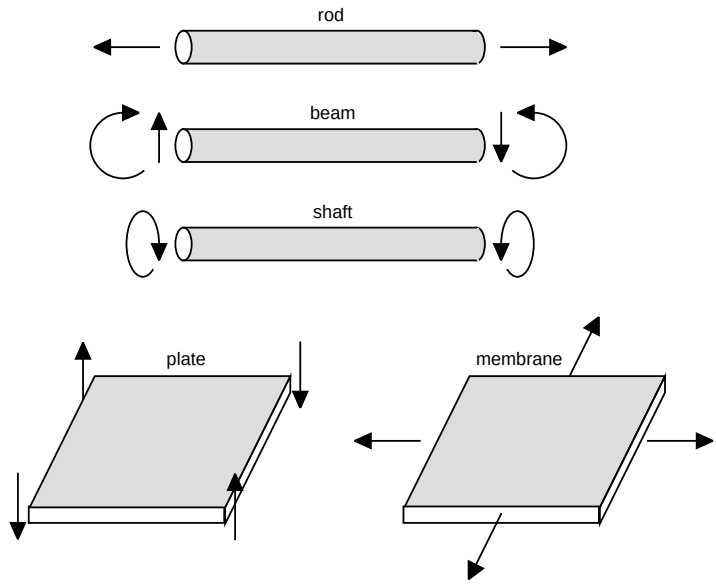


FIGURE 15.3.1 Common idealizations of solid mechanical behavior.

- Rods, with a local one-dimensional geometry and applied loads directed along the natural axis of the rod (rods are also termed spars or struts)
- Beams, which are similar to rods except that beams permit loads to be applied in directions perpendicular to the natural axis
- Shafts, which are also similar to rods, except that they are loaded by torques directed along their natural axis in addition to forces applied in that axial direction
- Membranes, which are locally two-dimensional structures loaded in the plane defined by their geometry
- Plates, which are two-dimensional flat structures loaded by arbitrary forces and couples, including those that excite membrane response and those that bend or twist the natural two-dimensional reference surface of the plate
- Shells, which are generalizations of plates to include curvature of the reference surface in three dimensions

Theories of computational solid mechanics are often categorized by reference to the particular structural idealization being studied. In particular, computational theories for beams, plates, and shells are similar because each is characterized by bending deformation, in which the primary mechanical response may occur at right angles to the natural geometric reference frame for the structure. Problems that include substantial bending deformation often require considerable care in computational modeling in order to avoid inaccurate results.

Another very important aspect of solid mechanics involves nonlinear material and geometric response. Nonlinear material behavior is commonly encountered in computational *inelasticity*, where the plastic, viscous, or nonlinear elastic response of a solid material must be modeled. In many materials (such as most metals) this inelastic behavior in realized after considerable elastic response, and hence it may often be neglected. In other materials (such as rubber or its relatives commonly used in aerospace applications), the nonlinear material properties must be addressed throughout the range of computational simulation. Furthermore, many of the idealizations presented earlier (such as rods and membranes) possess nonlinear geometric behavior that may dominate the mechanical response. One important example of this form of geometric nonlinearity is the buckling of a rod subjected to axial compressive stresses.

Fluid Mechanics

The basic principles of fluid mechanics are exactly the same as those for solid mechanics: the mechanical system under consideration is analyzed using appropriate equilibrium, compatibility, and stress-strain relations. In practice, because of the inability of a fluid to resist shear, fluids are considerably more difficult to characterize than solids, and the effective computational mechanics of fluids is therefore a more difficult and diverse topic. Perhaps the most important difference between common models for solid and fluid mechanics is that the base reference frame for the solution process is generally different. For solids, most analyses proceed by modeling the deformation relative to a reference frame attached to the material body being deformed. Fluids are generally modeled instead by attaching the problem's reference frame to a region in space and analyzing the flow of the fluid as it is transported through this geometric reference frame. The latter scheme is termed an *Eulerian* formulation of the reference frame, and the former is called a *Lagrangian* reference configuration. While the Eulerian world view is more common for analyzing fluids (primarily because tracking the individual motion of all particles of the fluid material is generally intractable in practice), this conceptually simpler approach gives rise to many practical computational difficulties because the underlying conservation principles are most readily posed in a reference frame attached to the material.

One of the most important mathematical models of fluid response are the Navier-Stokes equations. These equations are presented below for a three-dimensional flow field involving a Newtonian fluid (i.e., one readily characterized by the material parameter *viscosity*, which is assumed not to depend upon fluid stress) and an incompressible formulation (see Panton, 1984 or White, 1979). It is worth noting that even in the case of these simplifying assumptions, which correspond to those given above for the mathematical formulation of solid mechanics, the Navier-Stokes equations are nonlinear, in that products of solution quantities occur in the convective terms on the left-hand side of the momentum conservation equations. These nonlinearities arise because of the Eulerian reference frame used to model fluid flow, and cause fluid mechanics problems to be generally much more difficult to solve than analogous solid mechanics problems.

Definitions for Fluid Conservation Equations

u, v, $w = x, y, z$ components of velocity

p = pressure field

g_x, g_y, g_z = gravity body force components (15.3.5)

ρ = density

μ = viscosity

Fluid Momentum Conservation Equations

$$\rho\left(\frac{\partial u}{\partial t}+u\frac{\partial u}{\partial x}+v\frac{\partial u}{\partial y}+w\frac{\partial u}{\partial z}\right)=\rho g_x-\frac{\partial p}{\partial x}+\mu\left(\frac{\partial^2 u}{\partial x^2}+\frac{\partial^2 u}{\partial y^2}+\frac{\partial^2 u}{\partial z^2}\right)$$

$$\rho\left(\frac{\partial v}{\partial t}+u\frac{\partial v}{\partial x}+v\frac{\partial v}{\partial y}+w\frac{\partial v}{\partial z}\right)=\rho g_y-\frac{\partial p}{\partial y}+\mu\left(\frac{\partial^2 v}{\partial x^2}+\frac{\partial^2 v}{\partial y^2}+\frac{\partial^2 v}{\partial z^2}\right) \quad (15.3.6)$$

$$\rho\left(\frac{\partial w}{\partial t}+u\frac{\partial w}{\partial x}+v\frac{\partial w}{\partial y}+w\frac{\partial w}{\partial z}\right)=\rho g_z-\frac{\partial p}{\partial z}+\mu\left(\frac{\partial^2 w}{\partial x^2}+\frac{\partial^2 w}{\partial y^2}+\frac{\partial^2 w}{\partial z^2}\right)$$

Fluid Mass Conservation Equation

$$\frac{\partial}{\partial x}(\rho u)+\frac{\partial}{\partial y}(\rho v)+\frac{\partial}{\partial z}(\rho w)=0 \quad (15.3.7)$$

Fluid mechanical response is generally classified according to a variety of schemes, including whether the flow is internal (confined, as in flow-through ducts) or external (confining, as in the case of modeling flow around an airfoil in aerospace applications), and whether the internal or external flow field contains a fluid-free surface. Free-surface flows are considerably more complex to model using computational methods because the physical extent of the fluid material (and hence of the problem's geometric domain) is unknown *a priori*, and therefore determining the geometry of the mechanical problem must be considered as an essential part of the problem solution process.

One important facet of mechanical engineering practice involves diffusion phenomena, such as the flow of heat through a conductor. Such problems in heat conduction, porous media flow, and other diffusive systems can be readily modeled using standard fluid mechanical models appropriate for diffusion and dispersion. In addition, problems involving combined advection and diffusion (such as mass transport problems encountered in environmental engineering or in manufacturing) are readily modeled using techniques similar to related fluid mechanics models.

One of the most common ways to categorize fluid mechanical behavior arises from consideration of various dimensionless parameters defined by the constitution of the fluid and its flow regime. Representative dimensionless parameters relevant to fluid mechanics include:

- The Reynolds number, which relates the effects of fluid inertia to those of fluid viscosity forces and which naturally arises in a broad range of fluid problems
- The Mach number, which measures the ratio of fluid speed to the speed of sound in the fluid: this parameter is important in compressible flow problems
- The Peclet number, which quantifies the relative effects of fluid transport due to advection and dispersion and is used in fluid mass transport modeling

- The Taylor number, which is used to predict instabilities present in various fluid flow regimes and may be encountered when dealing with the computational ramifications of fluid instability

Unlike solid mechanics, there is no body of computational expertise that provides a generic model for computational approximation of the full range of fluid response. In practice, computational schemes are chosen according to particular aspects of fluid behavior (such as the various classifications and dimensionless parameters given), which substantially limits the range of applicability of any particular computational solution scheme. Many fluid models are solved using finite-element models, as in solid mechanics, but others have largely resisted characterization by such general computational schemes.

Fluid material models are generally more complex than those found in solid mechanics, as fluids exhibit a bewildering array of diversified material responses. Because of the difficulty in modeling many fluid constitutive parameters, many fluid models are analyzed using gross constitutive models that often neglect complicating material behavior (one common example of this case is the case of inviscid flow models, where neglecting the presence of fluid viscosity results in tremendous simplification of the resulting computational approximation). When fluid material response is important, as in the example of modeling transition from laminar to turbulent fluid behavior, the requisite constitutive equations for the fluid become appropriately more complex. While there is not yet a completely unified model for fluid constitutive behavior, each individual subdiscipline of computational fluid mechanics utilizes fluid material models that provide reasonable accuracy for fluid response within the range of the particular problems studied.

Electromagnetics problems are commonly solved by practicing mechanical engineers using techniques originally utilized for computational fluid dynamics. It is feasible to recast much of the computational technology developed for fluids into the setting of electromagnetics problems, because many electromagnetic simulations are qualititatively similar to fluid response. For example, potential flow simulation techniques are readily adapted to the potential formulations of electrostatics and magnetostatics. Alternatively, the full range of fluid response exhibited by the Navier-Stokes equations, which includes transient inertial and dissipative effects, is qualitatively similar to large-scale transient problems involving coupled magnetic and electrical fields. Mechanical engineers interested in performing electromagnetic simulations will commonly find that expertise in fluid computer modeling provides an excellent understanding of methods appropriate for solving electromagnetic problems.

Thermal Analysis

The determination of the static or transient temperature distribution in a physical medium is one of the best-known cases of the mathematical setting known as “scalar potential field” theory. In this case, the scalar potential is the temperature field, and the heat flow in the physical domain occurs in a direction governed by the product of a constitutive tensor known as the *thermal conductivity* and the gradient of the temperature distribution. In simple terms, heat flows downhill, or opposite the direction of the temperature gradient. The governing mathematical relations for transient conductive heat transfer in a three-dimensional solid are given below (see Kreith and Bohn, 1993). Note that this form of the energy conservation law will not apply if the conductivity varies with position, as is common in many mechanical engineering applications: in this case, the second derivatives on the right-hand side of this relation must be generalized to include spatial derivatives of the conductivity term.

Definitions for Thermal Equation

T = temperature field

k = thermal conductivity

s = distributed heat sources (15.3.8)

ρ = density

c = specific heat

Heat Energy Conservation Equation

$$\rho c \frac{\partial T}{\partial t} = -k\left(\frac{\partial^2 T}{\partial x^2} + \frac{\partial^2 T}{\partial y^2} + \frac{\partial^2 T}{\partial z^2}\right) + s \tag{15.3.9}$$

Conductive thermal problems are relatively straightforward to solve, which leads to the incorporation of thermal effects in many commercial solid and fluid mechanics analysis programs. The mathematical quality of these thermal analyses is that they are diffusive phenomena, in that they exhibit smooth solution behavior and that there are few localization effects either in space or time that cause serious computational problems such as those commonly found in transient analyses in solid, fluid, and electromagnetic systems. Convective thermal problems are considerably more difficult to model by computer because they involve transport of heat by a moving fluid, and thus may suffer from the same difficulties inherent in modeling transient fluid response.

Biomechanics

Biomechanics is an increasingly important component of mechanics (see Section 20.3) Computational simulation of biomechanical systems is difficult because many biomechanical problems straddle the boundary between solid and fluid mechanics and the mechanical properties of many organic components are diverse. For example, the computational modeling of head trauma, which is of considerable interest in automotive engineering applications, is complicated by the fact that the brain is a semisolid porous mechanical system, and its motion and deformation are further constrained by the solid enclosure of the skull. While rational finite-element mechanical modeling of head trauma has been underway for over two decades (Shugar, 1994), the problem still resists accurate computational simulation.

Even though biomechanical problems are inherently difficult, they are often encountered in professional practice. The practicing engineer interested in performing accurate computational modeling of complex biomechanical systems should insure that appropriate verification measures (such as computer visualization of model geometry and results) are undertaken in order to minimize the potential for errors caused by the overwhelming difficulties in modeling these important but often intractable mechanical systems.

Coupled Solid/Fluid/Thermal Field Analyses

Many “real-world” problems encountered in mechanical engineering practice defy simple categorization into solid, fluid, or otherwise. These practical problems often include combinations of classes of response, and the efficient solution of such coupled problems requires careful investigation of the individual components of the combined problem. There are several important questions that must be addressed when solving a coupled mechanical problem:

- Is it necessary to solve for all components of the solution response simultaneously (i.e., can the problem components be decoupled)?
- If the solution processes can be decoupled, which aspects of the solution must be found first, and which components can be considered of secondary relevance during the decoupled solution process?

- Are there any mathematical pathologies present in the coupled problem that may compromise the accuracy of the computed solution?

In general, fully coupled solution schemes (in which all components of the solution are found at the same time) will be more accurate over the entire range of solution response and will also require more computational effort. Under some circumstances, one or more components of the solution can be identified as of relatively minor importance (either because of their relative magnitude or their insignificant rate of change with time), and then the solution process can be decoupled in order to simplify the resulting calculations. For example, in many problems involving the coupled mechanical/thermal response of a material, it is common practice to assume that the temperature field can be calculated independently of the state of stress. Furthermore, an additional simplification can be obtained by assuming that the mechanical deformation of the system does not affect the temperature distribution substantially. The analysis of a mechanical system during relatively slow changes in ambient temperature (such as the diurnal variations of temperature encountered outdoors between night and daytime direct sunlight) could be modeled with reasonable accuracy using these assumptions, and the net result of this simplification is that the coupled problem could be solved as a series of uncoupled transient problems. At each time in the solution history, the temperature field would be computed, and this temperature distribution would then be used to compute the mechanical deformation at the same time. As long as the underlying assumptions leading to these simplified decoupled problems are satisfied, this uncoupled approach works well in practice.

In many problems, it is not possible to decouple the component problems, and all aspects of the solution must be determined simultaneously. Some important examples occur in aerospace applications, including the modeling of flutter instabilities in aircraft structures (where the coupled external flow and solid mechanics problems are considered concurrently) and the ablation of rocket engine components (where the modeling of combustion, heat transfer, thermal pyrolysis, mechanical deformation, and porous media flow effects all must be considered simultaneously).

In many coupled problems, there are mathematical solution pathologies present that may corrupt the accuracy of the computer solution. In particular, whenever the solution process involves determination of coupled kinematical solution components (i.e., geometric unknowns such as velocities and displacements) and statical solution parameters (i.e., force-like quantities such as pressure), then the mathematical setting for guaranteeing accuracy of the computer approximation may become considerably more complex. Coupled problems involving both statical and kinematical solution components are termed *mixed* problems, and practicing engineers who must model mixed problems ought to familiarize themselves with an outline of the potential pathologies that can be exhibited by naive application of mixed computational technologies. While a complete characterization of these problems is far beyond the scope of this brief introduction to coupled problems, some hint of the underlying theory and practice is presented in the following sections.

Mathematical Characteristics of Field Problems

Effective computational solution of the problems encountered in mechanical engineering requires at least an introductory knowledge of the underlying mathematics. Without a fundamental understanding of such relevant issues as solvability (the study of the conditions required for a solution to exist), convergence (measuring whether the computed solution is accurate), and complexity (quantifying the effort required to construct a computational solution), the practicing engineer may spend considerable time and resources constructing computational approximations that bear little or no resemblance to the desired “exact” answers. The material presented below is intended to provide an overview of the relevant mathematical content required to evaluate the quality of a computer approximation.

Solvability Conditions for Computational Approximations

Before any physical problem is modeled using a computational approximation, it is essential to ensure that the underlying mathematical model is *well posed*. Well-posed problems can be characterized as

those where a solution exists and where that solution is unique. While solvability concerns (i.e., those relating to whether a mathematical problem can be solved) such as mathematical existence and uniqueness are often considered as mere niceties by practicing engineers, in fact these properties are important aspects of constructing a numerical solution. It is imperative for engineers using numerical approximation schemes to understand that the computer programs generally used are deterministic: they will construct a numerical solution to the problem and will generate results, *regardless of whether the physical solution actually exists.* One of the most important components of the practical use of computational approximations such as finite-element models is the verification that the solution obtained actually provides a reasonable representation of the physical reality. There are a number of important problems that arise in mechanical engineering practice that are *not* well posed, and it is important for the engineer to recognize symptoms of this sort of mathematical pathology.

Existence of a solution is often guaranteed by the underlying physics of the problem. Most practical stress analyses model well-posed physical problems whose analogous mathematical existence is established by appropriate mechanics theorems. For example, any linear-elastic stress analysis of a compressible solid body can be shown to possess at least one solution. Furthermore, if the elastic body is constrained to prevent rigid-body translations and rotations, those same theorems insure that the solution is unique. As long as the engineer uses appropriately convergent computational models (i.e., ones that can be shown to be capable of getting arbitrarily close to exact results in the limit as the discretization is refined), then the unique solution constructed on the computer will represent an appropriate approximation to the unique physical problem being modeled. Analogous existence and uniqueness theorems can be found for a wide variety of problems commonly encountered in mechanical engineering practice.

There are many instances in which the physical problem is not well posed. One important example is when a stress analysis is performed on a problem that is not constrained to prevent rigid-body translations and/or rotations. In this case (which commonly arises during the modeling of vehicles and aircraft) the computational solution obtained will generally represent the "exact" static solution superimposed onto a large rigid body displacement. Depending upon the precision of the computer used (among other details), the computed stresses and strains may be accurate, or they may be contaminated due to truncation errors resulting from the static displacements being overwhelmed by the computed rigid-body motion that is superimposed upon the desired static solution. Proper attention to detail during the modeling process, such as constraining the solution to prevent unwanted rigid body displacement, will generally prevent this class of solvability pathologies from occurring.

Other more complex solvability concerns commonly arise in mechanical engineering problems where there is an underlying constraint applied to components of the solution field. Problems such as incompressible elasticity (where the volume of the mechanical body is invariant under deformation), incompressible flow (where the conservation of mass implies a "continuity" constraint on the velocity field), or electromagnetic field applications (where the magnetic field must be divergence-free) all possess constraints that can lead to serious solvability difficulties during the computational approximation process. In incompressible elasticity (the other mentioned cases admit similar resolutions), two particular pathologies present themselves: if there are no prescribed displacement constraints on the problem, then a solution will exist, but it will not be unique. In this case, the resulting computational approximation will exhibit rigid body motion, and the underlying stress analysis may or may not be accurate, depending upon the precision of the computer and the magnitude of the computed rigid-body motion. On the other hand, if the entire boundary of the incompressible body is subject to a constrained displacement field, then either the overall prescribed displacement of the boundary will satisfy a global incompressibility condition (in which case the solution will exist, the computed displacements will be unique, but the associated stresses may be suspect), or the global incompressibility condition is not satisfied (in which case there is no feasible solution to the underlying physical problem, so the computed solution will be completely spurious and should be ignored).

Because of the difficulty in ascertaining *a priori* whether solvability considerations are germane to the physical problem being modeled, alternative schemes are often desirable for verifying that the computed solution actually represents a reasonable approximation to the physical problem being mod-

eled. One particularly useful computational scheme for verifying solution quality involves the application of computer graphics and visualization to display solution behavior (e.g., stresses and strains) for interactive interpretation of the results by the engineer. Verification of input and output via graphical means involves the development and use of graphical pre- and postprocessors (respectively) designed to permit interactive review of the solution behavior. High-quality interactive pre- and postprocessors for use with a particular computational mechanics application are very desirable characteristics and should be given due consideration during the selection of CAE software for use in mechanical engineering practice. When nonlinear problems (with their accompanying solvability pathologies) are encountered, use of interactive graphics for solution verification is often required to insure the reliability of the computed nonlinear solution.

Convergence, Stability, and Consistency

In any numerical method, the primary issue that must be addressed is that of *convergence*. A convergent method is one that guarantees that a refinement of the discretization will produce generally more accurate results. Ultimately, as the discretization mesh size decreases, the exact solution can be approximated to an arbitrarily small tolerance; as the size of the discretization measure decreases, the answers should converge to the correct solution. A second criterion is that of *accuracy*, which is related to that of convergence. Where convergence addresses the question "does the error go to zero as the discretization step size decreases?," accuracy is concerned with "at any particular size step, how close is the approximation to the solution?," and with "at what *rate* does the error go to zero with step size?." A convergent method in which the error tends to zero as the square of the step size (a *quadratic* convergence rate) will eventually become more accurate than another scheme in which the error and step size decrease at the same rate (a *linear* convergence rate).

Concern especially relevant to time-dependent simulations is that of *stability*: a stable method is one that guarantees that errors introduced at one step cannot grow with successive steps. If a method is unstable, even when the errors introduced at each step are small, they can increase exponentially with time, thus overwhelming the solution. In a stable method this cannot occur, although stability alone does not imply anything about the size of solution errors that may be introduced at each step, or whether errors from different times can grow by accumulation. Many numerical methods for time-dependent approximation are only *conditionally stable*, implying that stability is guaranteed only when the step size is smaller than some critical time scale dictated by the data of the physical problem and by the discretization (in mechanics problems, this time scale is often related to the shortest period of vibration for the structure). Some idea of this critical time scale must be known *a priori* for a conditionally stable method to behave in a robust manner. For this reason, the use of *unconditionally stable* methods is often preferred, since these methods are stable regardless of the step size (although the actual size of the errors introduced at each step may still be large).

Another important issue in numerical approximations for differential equations is *consistency*. A consistent discrete approximation is one that approaches the underlying continuous operator as the size of the discretization tends to zero. All standard finite-difference operators (and their finite-element relatives used for time-dependent approximations) are consistent, because as the size h of the discretization parameter tends to zero, the appropriate derivative is recovered. Inconsistent discretization schemes are rarely (if ever) encountered in practice.

The convergence and stability characteristics of a numerical method are not independent: they are related by the *Lax Equivalence Theorem*, which is alternatively termed "The Fundamental Theorem of Numerical Analysis." In simple terms, the Lax Equivalence Theorem states that *convergence* (which is the desired property, but one which is often difficult to demonstrate directly) is equivalent to the combination of *consistency* and *stability*. Because consistency and stability are relatively easy to determine, the task of finding convergent approximations is often replaced by that of searching for schemes that are simultaneously consistent and stable.

Computational Complexity

Computational complexity is another important characteristic of a numerical scheme. The complexity of an algorithm is measured in terms of the asymptotic rate of growth of computational effort required to perform the algorithm. For example, standard schemes (such as factorization or Gauss Elimination) for solving a full linear system of equations with N rows and columns require computational effort that grows as the cube of N; these algorithms thus possess a *cubic* rate of growth of effort. (Note that standard schemes such as Gauss Elimination also involve terms that grow as the square of N, or linearly with N: as N becomes asymptotically large, these latter terms diminish in size relative to the cube of N, so the algorithm is characterized by the cubic term which ultimately dominates the growth of effort.) Many (or most) standard numerical schemes exhibit some form of polynomial computational complexity, and for large problems, selecting the algorithm with the lowest polynomial exponent (i.e., the slowest asymptotic rate of growth) is generally a good idea.

Many common algorithms encountered in engineering handbooks are not suitable for large-scale engineering computation because their computational realizations exhibit prohibitively large growth rates. For example, many recursive schemes, such as Cramer's rule for finding the solution of a linear system of equations via recursive evaluation of various matrix determinants, exhibit rates of growth that are larger than *any* polynomial measure. Cramer's rule possesses a *factorial* rate of growth of computational effort (i.e., the effort grows as N!), and thus should be avoided completely for N larger than ten. When faced with implementing or using a computational scheme such as this, the practical engineer should consult appropriate references in order to determine whether alternative simpler algorithms are available — in the case of Cramer's rule, it is easy to solve a linear system (or evaluate a determinant) using algorithms that grow cubically, and for large N, there is a vast difference in effort between these latter schemes and a naive implementation of Cramer's rule, as shown in the Table 15.3.1 (note that the values of N presented in the table are extremely small by the standards of typical computational approximations, and hence the relative gain in using lower-order complexity schemes is even more important than the results in the table might indicate).

TABLE 15.3.1 Polynomial and Factorial Growth Rates

N	N^2	N^3	N!
2	4	8	2
5	25	125	120
10	100	1,000	3,628,800
20	400	8,000	2.433×10^{18}
50	2,500	125,000	3.041×10^{64}

In general, the choice of algorithms for computational simulation of mechanics problems should be restricted to methods that are both convergent and at least conditionally stable. In addition, it should be clear that an unconditionally stable method, especially if it has a higher-order convergence rate, is to be desired. Finally, one normally prefers those methods that exhibit the slowest asymptotic rate of growth of computational effort, especially when large problems are to be solved.

Finite-Element Approximations in Mechanical Engineering

Finite-element models are among the most efficient, general, and accurate computational approximation methods available in engineering. Many of the attractive characteristics of finite-element computational models arise from a seemingly unlikely source: instead of representing the mathematical statement of the underlying mechanics as systems of *differential* equations, finite-element models are motivated by considering instead equivalent *integral* formulations, e.g., the basic equations of equilibrium, compatibility, and stress-strain can be replaced by equivalent integral energy formulations such as the Principle

of Virtual Work or the Theorem of Minimum Potential Energy. These latter integral principles form the basis for finite-element computational approximations.

For example, the differential equations for equilibrium of a solid body can be replaced by an integral formulation: the principle of virtual work of elementary mechanics is one particular well-known example of this equivalence. Once identified, the equivalent integral formulation is considerably simpler than the differential form, and this simplicity is even more apparent when mechanics problems are generalized to include such practical effects as point loads, arbitrary geometric constraints, and heterogeneous or anisotropic material properties. Under these more interesting conditions (i.e., in situations that are likely to arise in mechanical engineering practice), the differential equations of mechanics become nearly impossible to write, as the standard calculus symbolism is barely sufficient to express the range of intractable differential behavior. However, the integral formulation requires only marginal modifications under these more general conditions, and this simpler integral approach also leads to appropriate elegant and general computational approximation schemes.

In addition to providing a simpler means to express the fundamental mathematical relations of mechanics, the integral foundation of finite-element models also provides other important features. One important aspect of finite-element models is their computational efficiency; another is that strong and general convergence proofs can be cited for large classes of finite-element models, obviating the need for application of the more complicated Lax Equivalence Theorem to prove convergence and estimate accuracy. Perhaps the most important theoretical characteristic of these integral calculus models is that their general formulation provides a unified computational framework that includes many other practical schemes (such as finite-difference models) as special cases of a general integral theory of computational mechanics. Therefore, careful study of the mathematical underpinnings of finite-element modeling also provides considerable insight into such diverse alternative computational schemes as finite-difference models and boundary-element methods.

Overview of Finite-Element Models

Finite-element models are most commonly utilized in computational solid mechanics, where they have been used successfully on a very general range of practical problems. Integral formulations for solid mechanics (such as the Principle of Virtual Work) are well-established general results, and it is straightforward to develop generic finite-element computer models from these fundamental integral mechanics formulations. Since many problems in other branches of mechanics (such as Stokes Flow in fluid mechanics, which is a direct mathematical analog of the solid mechanical problem of deformation of an incompressible elastic medium) are closely related to standard problems of solid mechanics, finite-element models are rapidly gaining favor in many other branches of mechanics as well. In general, any physical problem that can be expressed in terms of integral energy formulations is a natural candidate for finite-element approximation.

The most common application of finite-element models is in computational solid mechanics. In this area, finite-element models can be made arbitrarily accurate for a wide range of practical problems. Standard stress analysis problems involving moderate deformations of well-characterized materials (such as metals) are generally easy to solve using finite-element technology, and static or time-dependent simulations of these mechanical systems are routinely performed in everyday mechanical engineering practice. More complex problems (such as computational fluid mechanics, coupled problems, or electromagnetic simulations) can readily be solved using finite-element techniques, though the accuracy of the solution in these areas is often limited, not by the finite-element technology available, but by difficulties in determining accurate characterizations for the materials used. For example, in many biomechanical problems, the model for simulating the stress-strain response of the biological material may contain uncertainties on the order of many percent; in cases such as these, it is not the quality of the finite-element application that limits the accuracy of the solution, but instead the inherent uncertainty in dealing with realistic biological materials.

There are still many practical limitations to finite-element modeling in engineering practice. Some of these problems arise from inherent difficulties in developing generic finite-element computer programs,

while others arise from the same generality that makes finite-element models so attractive as computational approximations. Examples of the former include computer programs that do not contain completely robust modern finite-element models for plates and shells, while examples of the latter include finite-element applications that incorporate the latest computational technology but are so difficult to use that inaccuracies result from gross errors in preparing input data. Finally, many practical limitations of finite-element modeling are due to theoretical difficulties inherent in the mathematical formulations of mechanics, and these last pathologies are the most difficult to remove (as well as being problematic for *any* computational scheme and not therefore particular to finite-element techniques).

Classification of Finite-Element Models

There are many classifications possible in computational mechanics, and so choosing an appropriate classification scheme for any general method such as finite-element modeling is necessarily an ambiguous task. The most common classification scheme for categorizing finite-element models is based on consideration of the physical meaning of the finite-element solution, and a simplified version of this scheme can be described by the following list:

- If kinematical solution unknowns (such as displacement or velocity) are determined first, then the model is termed a displacement-based finite-element model.
- If statical finite-element solution parameters are the primary unknowns determined during the computational solution process, then the model is termed a stress-based finite-element model.
- If a combination of kinematical and statical solution unknowns are determined simultaneously, then the computational scheme is called a mixed finite-element model.

Displacement-based finite-element models are the most common schemes used at present, and most commercial-quality finite-element applications contain extensive element libraries consisting primarily of displacement-based families. In a displacement-based model, the analyst creates a finite-element mesh by subdividing the region of interest into nodes and elements, and the program calculates the solution in terms of nodal displacements (or velocities, in the case of flow problems). Once the displacements have been determined, the unknowns in the secondary solution (such as stress and strain) are derived from the finite-element displacement field by using appropriate compatibility and stress-strain relations. Some common two-dimensional finite elements used for displacement-based formulations are catalogued in Figure 15.3.2. The elements shown are successfully used for common problems in stress analysis, simple fluid flow, and thermal analyses, and their generalization to three-dimensional problems is straightforward and computationally efficient.

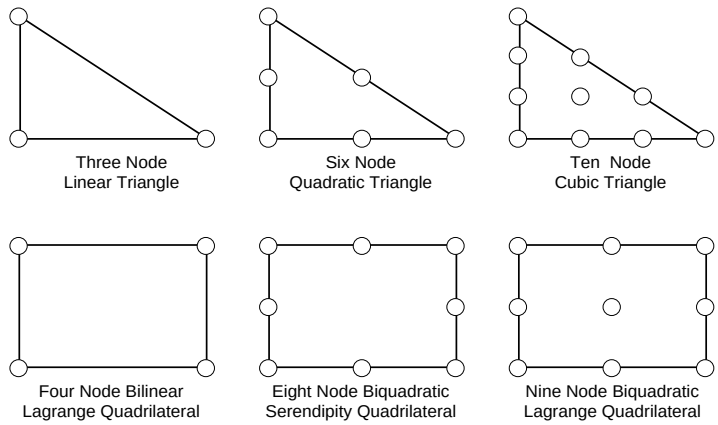


FIGURE 15.3.2 Common two-dimensional finite elements.

Stress-based finite-element schemes are much more rare than displacement models, although stress-based finite elements have some attractive theoretical qualities (such as more direct calculation of stresses and strains, which are the most important currency in practical stress analysis). Many practical stress-based finite-element methods are recast into a form involving determination of nodal approximations for displacement, so that these alternative formulations can be easily integrated into existing displacement-based finite-element programs.

Mixed finite-element schemes commonly arise in problems involving constraints, such as electromagnetics, incompressible deformation or flow, and (in some cases) solid mechanics problems characterized by substantial bending deformation. When using a mixed computational scheme, considerable care must be taken to insure that the resulting approximation is convergent: these subtle mathematical solvability conditions are best approached by a practicing engineer by insuring that the program used automatically applies appropriate consistency conditions (which are commonly presented in the literature for each given field of practice). In short, the best ways to insure that reasonable results are obtained from mixed models are to:

- Know which problems commonly result in mixed formulations.
- Use finite-element computer programs that have already been tested and verified for use on the class of problems of interest.
- Take advantage of auxiliary software (such as graphical visualization applications) to verify the input and output data for all problems.

Note that the potential convergence problems that haunt some mixed formulations in finite-element modeling are generally caused by the underlying mathematical model used to represent the engineering system. In these cases, *any* computational method used may exhibit solution pathologies, and, in fact, the finite-element model (while still containing the pathology) is often the best approach for identifying and rectifying the underlying mathematical problem (this result follows from the fact that the integral framework underlying finite-element approximation permits a more general study of convergence than do competing differential frameworks utilized in finite-difference and finite-volume models).

Finite-element formulations can also be classified into two particular flavors of development: Rayleigh-Ritz and Galerkin. Rayleigh-Ritz finite-element models are developed by applying theorems of minimum potential energy (or some similar integral extremization measure). Galerkin models arise from so-called weak statements of boundary-value problems and are simultaneously more complicated to derive and substantially more general in scope. For problems (such as standard linear stress-analysis problems) where both minimum and equivalent weak statements of the problem exist, the resulting Rayleigh-Ritz and Galerkin finite-element models will yield identical results. For problems where no extremization principle exists (such as high-Reynolds number fluid flow), Galerkin finite-element models (and their relatives) can still be used in practice.

Computational Characteristics of Finite-Element Models

Finite-element models are widely used in mechanical engineering precisely because of their computational efficiency. A brief enumeration of the computational characteristics of finite-element models highlights many reasons why they form such an effective family of computational approximation techniques.

- Finite-element models tend to result in computational algorithms that are readily understood, easily implemented, and exhibit well-known computational advantages. For example, finite-element equation sets tend to be sparse (i.e., to have a relatively small percentage of nonzero terms in the coefficient matrix, with these nonzero entries in a well-defined pattern that can be utilized to achieve efficient computational implementation of required matrix operations) and preserve physical characteristics (such as symmetry and structural stability) that are readily exploited by common matrix algebra schemes.

- Finite-element models are readily shown to possess optimal accuracy characteristics for a substantial portion of problems commonly encountered in engineering practice. For example, for most classes of practical stress analysis, finite-element solution schemes can be shown to yield the most accurate results (measured in terms of minimizing the energy of the error) of any standard numerical approximation scheme, including competing finite-difference and finite-volume methods. In contrast to these latter discretization schemes, finite-element optimal convergence rates are preserved even in the important case of irregular or unstructured meshes. The practical effect of this optimality is that there is often little or no incentive to use any other technique for solution of many computational mechanics problems encountered in the mechanical engineering profession.
- Many finite-element calculations are inherently scalable, in that they can be readily implemented in a manner amenable to parallel computer architectures. The practical result of this scalability property is that current advances in multiprocessor computer architectures are easily leveraged to gain substantial improvements in finite-element program performance with little additional effort required by either users or programmers.
- Finite-element results are defined everywhere in the problem domain, and these results include derived parameters such as stress and strain. Because these solution components are defined everywhere, it is very easy to apply additional computer technology (such as integrated graphical postprocessors) to verify and interpret the results of the finite-element analysis. It is commonly (and incorrectly!) believed that *all* numerical methods provide computational solutions only at a finite number of points in the domain, but this observation is erroneous in the case of finite-element models. The global definition of the finite-element solution is of considerable utility toward the important goal of evaluating the convergence of finite-element models. In addition, the derived secondary solution parameters are also convergent, so that the optimal accuracy characteristics of finite-element models generally apply to both primary (e.g., displacement) and secondary (e.g., stress) solution unknowns. This latter advantage is much harder to prove for many competing numerical methods such as finite-difference and finite-volume schemes.
- It is relatively easy to guarantee the convergence of common finite-element solution schemes, and this characteristic, when combined with the mathematical optimality of the finite-element solution, facilitates trust in the computed finite-element results. The convergence properties of finite-element models are cast in an integral (as opposed to a differential) form, which results in their applicability toward a wide spectrum of engineering problems that contain mathematical pathologies (such as heterogeneous materials, point loads, and complex boundary conditions) that compromise the convergence statements for models (such as finite-difference schemes) that would otherwise be competitive with finite-element models in terms of computational effort.
- For some problems, including those involving substantial convective nonlinearities due to high-speed fluid flow, the optimality characteristics of standard finite-element models can no longer be guaranteed, and an alternative integral convergence formulation must be sought. In the setting of finite-element models for flow, this convergence framework is termed *weak convergence*. It is important to understand that "weak" in this sense is not a pejorative term, but instead refers to the topological structure of the solution space for these flow problems. In fact, weak formulations for problems in science and engineering are useful representations for many practical cases; for example, in the study of statics, the principle of virtual work is a weak formulation of the equations of equilibrium of a rigid body. The weak integral convergence framework for Galerkin finite-element models is an attractive alternative to analogous (but generally more complex) convergence results used in finite-difference modeling for fluid problems.

In simple terms, finite-element models are very attractive candidates for computational approximation efforts. Under very general conditions, they can be demonstrated to be extremely accurate, computationally efficient, and amenable to improvements in computer hardware and software technology. Because

of these computational advantages, finite-element models have become popular in virtually all branches of computational mechanics.

Advantages and Disadvantages of Finite-Element Models

Like all computer approximation schemes, finite-element modeling has characteristic advantages and disadvantages, and the effective use of finite-element technology requires at least an introductory understanding of each. Besides the computational advantages of efficiency and optimality presented above, most finite-element schemes include other practical advantages. Foremost among the advantages are the characteristics of generality of function and flexibility of modeling. The primary disadvantages occur when naive finite-element implementations are applied to mathematically ill-behaved problems.

Finite-element models can be cast into general computer applications that permit a tremendous variety of problems to be solved within the scope of a single program. Some commercial-quality finite-element programs permit the analysis of static and time-dependent problems in solid mechanics, fluid mechanics, and heat conduction within one monolithic computer application. In contrast, alternative computer analysis techniques often place substantial constraints on the range of problems that can be solved, and these limitations require the practicing engineer to be familiar with a variety of different software packages instead of an inclusive general analysis application.

Most finite-element packages are very permissive in the modeling of complexities that regularly occur in engineering practice. These complicating factors include material discontinuities (which are common in composite applications), material anisotropies (implementing more general material response within typical finite-element applications is generally a simple programming task) and complex boundary conditions (such as combinations of traction and displacement on boundaries not readily aligned with the coordinate directions). Because most commercial finite-element applications have very general features, a common source of analysis errors arises from gross mistakes in creating input data (such as specifying a different material model or an incorrect boundary condition). One of the easiest ways to identify gross modeling errors is to make use of solution verification tools (such as integrated graphical pre- and postprocessors) whenever possible. Virtually all high-quality commercial finite-element packages incorporate integrated visualization tools, and utilizing these tools to verify input data and output results is always a good idea.

There are also more complicated problems that can arise in the course of finite-element modeling: the most common pathologies arise from analyses incorporating mixed finite-element models. Guaranteeing the convergence of mixed finite-element models is a much more difficult mathematical problem than the analogous proof of convergence for displacement-based models, and it is possible to get unreliable results from naive use of mixed finite-element models. While most modern commercial finite-element programs will largely guard against the possibility of constructing inherently poor finite-element approximations, there are older or less robust finite-element applications available that do not incorporate completely dependable finite-element technologies. Therefore, it is important to identify potential problem areas for finite-element modeling. The most common sources of difficulty are presented below.

Incompressible Problems. Problems where volume is preserved commonly occur in mechanical engineering practice, but errors arising from inaccurate finite-element modeling techniques are often readily identified and corrected. Incompressible solid elastic materials include rubber and solid propellants, and models for inelastic response of metals also require incompressible or near-incompressible plastic behavior. The most common artifact of poor modeling of incompressibility is known as mesh locking, and this aptly named phenomenon is characterized by grossly underestimated displacements. The easiest way to test for potential incompressibility problems is to solve a suite of simple case studies using small numbers of elements, as shown in Figure 15.3.3. If the finite-element mesh locks in these test examples, then either different elements or a different finite-element application should be considered, as adding more elements to the mesh does not generally improve the results in this case. It is easy to accommodate incompressibility in the setting of standard finite-element libraries. Figure 15.3.4 shows several two-dimensional elements that are accurate when used in an incompressible elasticity setting. Modern

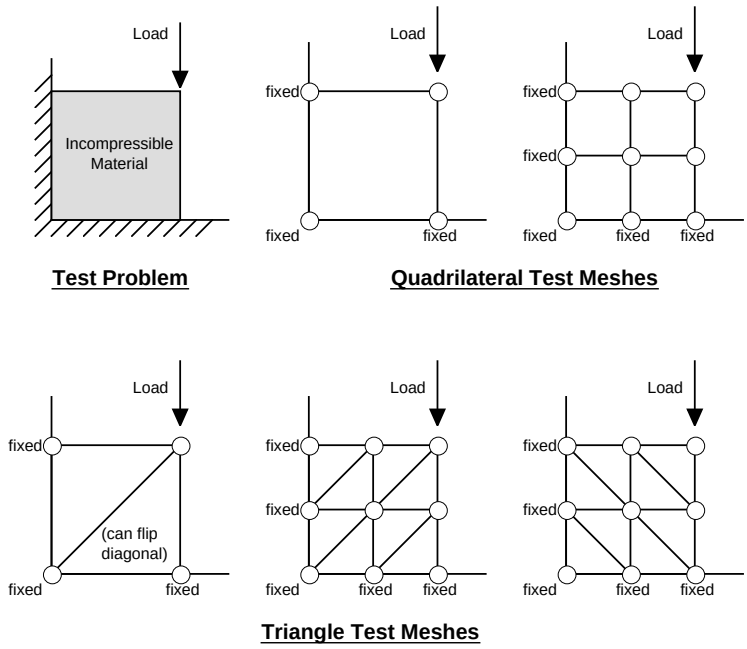


FIGURE 15.3.3 Test problem suite for incompressibility.

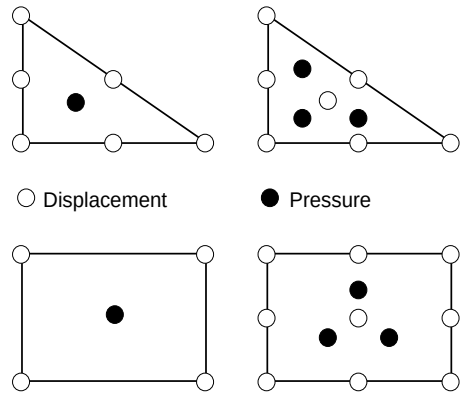


FIGURE 15.3.4 Incompressible elasticity finite elements.

textbooks on finite-element theory (e.g., Hughes, 1987 or Zienkiewicz, 1991) generally provide guidelines on choice of element for incompressible problems, and the user's manuals for many commercial finite-element packages offer similar preventative advice.

Bending Deformation. Problems involving beams, plates, and shells may exhibit solution pathologies similar to those found in incompressible elasticity. In these problems, one-dimensional beam elements or two-dimensional plate elements are loaded in directions normal to the element's reference surface, and naive coding practices for these cases may result in locking of the finite-element mesh, with attendant underprediction of the displacement field. Depending upon the characteristics of the particular element and the stability of the geometric boundary constraints, mesh locking errors may instead be replaced by wildly unstable solution response, with the displacement field oscillating uncontrollably over adjacent elements. In practice, it is very easy to avoid both of these pathologies by incorporating simple additional programming effort, so this problem is rarely encountered when commercial finite-element packages are used. It may be important when the finite-element application is a local or unsupported product, and

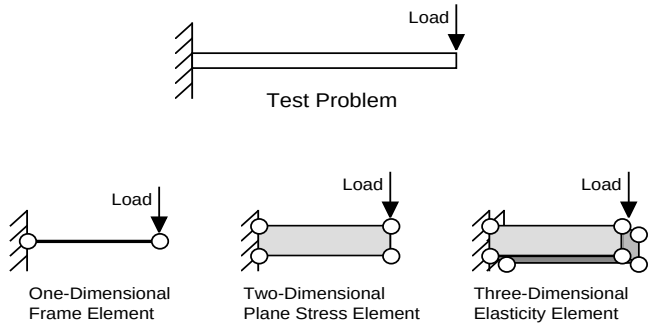


FIGURE 15.3.5 Single-element flexure test problems.

in these cases, appropriate care should be taken to verify the solution. If graphical solution visualization tools are available, they should be used in these cases. If such tools are not available, then the computed results should be compared to simplified problems that can be solved analytically using engineering beam theory analogies or closed-form plate models. One simple test case for naive finite-element coding practices is shown below (see Hughes, 1987); if this simple (but marginally stable) structure is analyzed and gives excessively large or oscillating displacements, then a better finite-element model (or application) should be sought.

Lack of Curvature Deformation in Low-Order Elements. Low-order linear, bilinear, and trilinear elements (such as four-node bilinear quadrilaterals used in plane stress and plane strain applications) are commonly utilized in finite-element packages because they are easy to implement and because many software tools exist for creating two- and three-dimensional meshes for such low-order elements. Unfortunately, there are common circumstances when these low-order elements give inaccurate and unconservative results. One such example is shown in Figure 15.3.5. Many simple finite-element packages will provide accurate answers for the one-dimensional problem, but will substantially underpredict the two- and three-dimensional displacements. Unlike the mesh-locking pathology encountered in incompressible cases, the underprediction of displacements from low-order elasticity elements eventually disappears as more elements are used when the mesh is refined. It is easy to remedy this problem in general by choosing appropriate continuum elements, and many commercial codes provide appropriate low-order elements automatically. Problems due to poorly implemented low-order elements are best avoided by solving simple one-element models first to determine whether the particular element to be used exhibits these errors, and if this is found to be the case, the analyst can either choose a different type of element (or a better finite-element package!) or take pains to insure that the mesh is appropriately refined.

It is important to recognize that many of the problems that arise when using finite-element models for these more difficult cases are not completely due to the particular finite-element model used, but may arise from pathologies in the underlying mathematical problem statement. In these cases, any computational approximation scheme may be susceptible to errors, and so care should be used in evaluating any computational results.

Test Problem Suites for Finite-Element Software

It is easy to avoid most errors in finite-element analysis, by first insuring that no mathematical pathologies are present in the problem or in the elements to be utilized and then by guarding against gross modeling errors in preparation of the input data. The best way to guard against both forms of errors is to use a suite of simpler problems as a “warm-up” to the more complex problems that must be solved. The general idea here is to solve a series of problems of increasing complexity until the final member of the series (namely, the complex problem to be modeled accurately) is obtained. The simpler members of the suite of problems generally involve simplifications in geometry, in material response, and in scale of deformation. For example, if a complicated problem involving large-deformation behavior of an inelastic material is to be analyzed, simpler problems using small-deformation theories, elastic material

models, and simplified problem geometry should be solved and carefully verified. These simpler problems can then be used to verify the response of the complex problem, and this approach is of considerable help in avoiding spectacular engineering blunders.

Another role of a problem suite arises when different competing commercial finite-element packages are to be evaluated. Most commercial finite-element packages are specifically oriented toward different engineering markets, and their respective feature sets depend upon the needs of those particular engineering disciplines. When evaluating commercial finite-element software for potential purchase, a wide selection of representative problems should be analyzed and verified, and two excellent sources of test problems are the pathological examples of the last section (which will provide insight into the overall quality of the commercial finite-element application) and any suite of test cases used to verify more complex problems with a competitive program (which will provide information on how well the proposed application provides support for the particular class of problem encountered in local practice). Problems encountered with either type of test problems should be appropriately investigated before purchase of new finite-element software.

Finite-Difference Approximations in Mechanics

Finite-difference models include many of the first successful examples of large-scale computer simulation for mechanical engineering systems. Finite-difference modeling techniques were originally developed for human calculation efforts, and the history of finite-difference calculations is considerably longer than the history of finite-element modeling. Once automatic computation devices replaced humans for low-level computing tasks, many of the existing finite-difference schemes were automated, and these simple numerical techniques opened up much of engineering analysis to computer simulation methods. Over time, improved finite-difference approximations (including ones that would be prohibitively difficult for human calculation) resulted, and the range of problems solvable by finite-difference models became more diverse.

Throughout the 1970s and 1980s, a natural division developed between finite-difference and finite-element models in computational mechanics. Finite-difference models were utilized primarily for a wide variety of flow problems, and finite-element approximations were used for modeling the mechanical response of solids. While many finite-element approximations were proposed for fluid problems such as the Navier-Stokes equations, in practice these first-generation finite-element models generally performed poorly relative to more advanced finite-difference schemes. The poor performance of naive finite-element models relative to carefully optimized finite-difference approximations was widely (but incorrectly) accepted in the mechanical engineering profession as evidence that fluid problems were best solved by finite-difference methods, even as finite-element technology gained near-exclusive use in such fields as stress analysis.

In fact, the problems associated with early-generation finite-element models in fluids were due primarily to rampant overgeneralization of insight gained in solid mechanics instead of any inherent limitations of finite-element technology applied to fluid mechanics problems. Similar problems developed in the early stages of finite-difference modeling; for example, one of the first finite-difference references published by Richardson used theoretically attractive but practically worthless difference schemes and resulted in a completely unusable algorithm (see Richardson, 1910 or Ames, 1977 for details of this interesting historical development). During the 1980s and 1990s, a better understanding emerged of how to utilize finite-element technology on the relatively more complicated models for fluid behavior, and there has been an explosion of interest in using finite-element approximations to replace finite-difference schemes in a variety of fluid mechanics settings. While finite-difference methods (and especially carefully optimized applications found in certain difficult and narrow specialties in fluid modeling) continue to be widely used in fluid mechanics and electromagnetic simulations, the range of mechanics problems thought to be resistant to finite-element solution approaches is rapidly diminishing.

Finally, a word regarding nomenclature is warranted. Originally, finite-difference models were developed using regular grids of equally spaced points, with the solution approximation computed at the

vertices of this structured lattice. As increasingly complicated problems were solved with finite-difference technology, it became desirable to be able to model more complex problems involving variable material properties, singular data (such as point sources), and irregular geometries. In order to permit these more general problems to be solved, finite-difference modeling was generalized to include various averaged (i.e., integrated) schemes that augmented classical vertex-oriented finite-difference methods to include cell-centered schemes where appropriate integral averaging of the underlying differential equation is utilized to permit modeling discontinuous or singular data. Ultimately, these various improvements led to the development of *finite-volume* schemes, which are a general family of integrated finite-difference approximations that preserve the relative simplicity of finite-difference equations, while casting the underlying theory in a form more amenable to modeling complicated problems. Finite-volume methods are discussed in a later section of this chapter, but the general characteristics given below apply to the entire finite-difference/finite volume family of approximate solution techniques. Because of the widespread application of these integral generalizations of classical vertex-based finite-difference schemes, it has become commonplace to use the term "finite-difference methods" collectively to include all the members of this computational family, including vertex-centered schemes, cell-averaged methods, and finite-volume approximations.

Overview of Finite-Difference Models

Finite-difference methods arise from the combination of two components:

- Differential equations of mathematical physics
- Replacement of differential operators with discrete difference approximations

Differential statements of the equations of mathematical physics are well known, as most of engineering and physics instruction is provided in this form. For example, the familiar relation "$\underline{f} = m\,\underline{a}$" is a vector differential equation relating the components of the applied force to the differential temporal rate of change of momentum. The analogous integral statement of this fundamental relation (which is used to motivate finite-element modeling instead of finite-difference approximation) is probably more useful in computational practice, but it is definitely less well known than its differential brethren. Similar instances occur in every branch of mechanics, as the differential equations for engineering analysis are generally introduced first, with their equivalent integral statement presented later, or not at all. Because of this widespread knowledge of the differential form of standard equations for modeling physical systems, the supply of differential equations that can be converted to analogous finite-difference schemes is essentially inexhaustible.

The computational component of finite-difference modeling arises from the replacement of the continuous differential operators with approximations based on discrete difference relations. The gridwork of points where the discrete difference approximation is calculated is termed the finite-difference *grid*, and the collection of individual grid points used to approximate the differential operator is called (among other terms) the finite-difference *molecule*. Sample finite-difference molecules for common differential operators are diagrammed in Figure 15.3.6. One of the most obvious characteristics of the molecules represented in Figure 15.3.6 is that the discrete set of points where the finite-difference approximation is computed is very structured, with each grid point occurring in a regularly spaced rectangular lattice of solution values.

The fundamental reason for this regular spacing has to do with necessary conditions for optimal convergence rates for the finite-difference solution. This regular spacing of points provides simultaneously the greatest incentive and the biggest constraint toward practical use of finite-difference methods. Performing finite-difference calculations on a regular grid provides desirable fast convergence rates on one hand, but the regularity in the placement of solution sampling points often forces the analyst into a fixed grid structure that is difficult to adapt to nonregular domains without sacrificing optimal rates of convergence. For some problems, it is possible to perform coordinate transformations of the entire physical (i.e., problem) domain onto a regularized computational domain in order to gain simultaneous high convergence rate and geometric modeling flexibility. Unfortunately, these requisite coordinate

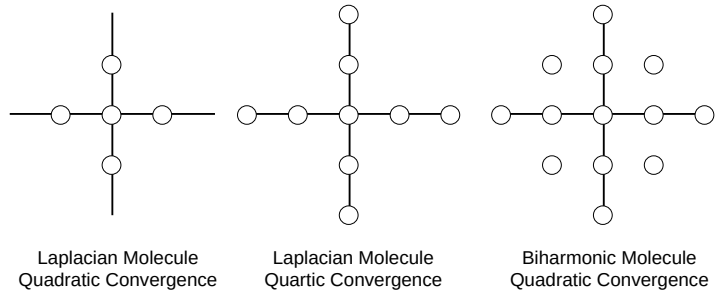


FIGURE 15.3.6 Sample finite-difference molecules.

transformations may be difficult to determine for sufficiently complicated problem geometries. This limitation should be contrasted with the geometric flexibility inherent in analogous finite-element models, where optimal rates of convergence are readily realized even in the presence of irregular or unstructured finite-element meshes.

A geometric motivation for the improved performance of regularly spaced finite-difference molecules can be seen in Figure 15.3.7. Here a representative function has been drawn, along with its tangent derivative and two estimates of the derivative's slope given by standard secant formulas from elementary calculus. In particular, the forward and centered-difference approximations for the slope of the tangent line can be visualized by passing lines through the points $[x, f(x)]$ and $[x + h, f(x + h)]$ for the forward-difference approximation, and $[x - h, f(x - h)]$ and $[x + h, f(x + h)]$ for the centered-difference estimate. (Note that it is also possible to define a backward-difference slope estimate by considering the secant line passing through the function at the points x and $(x - h)$. In problems involving spatial derivatives, this backward-difference estimate has numerical properties that are generally indistinguishable from the forward-difference approximation, but in the setting of time-dependent derivatives, there are considerable differences between forward- and backward-difference schemes.)

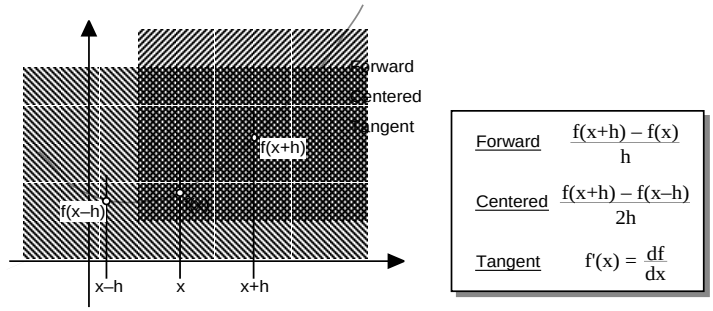


FIGURE 15.3.7 Finite-difference formulas and geometry.

It is apparent from the figure that the centered-difference approximation for the slope is substantially better than that obtained from the forward difference. While it may appear from this particular example that the improved quality of the centered-difference estimate is due to the particular shape of the function used in this case, in fact these results are very general and represent a common observation encountered in finite-difference practice: centered-difference formula are generally optimal in providing the highest rate of convergence relative to the number of points required to evaluate the difference approximation. While this observation can be motivated from a variety of viewpoints (the most common example for purpose of development of finite-difference methods is the use of Taylor Series expansions), it is particularly instructive to perform some computational experimentation regarding this optimality of centered-difference models.

Consider the following simple FORTRAN program:

```
      REAL X, H, FXH, FX, FPRIME, ERR
      INTEGER I
*
*     calculate estimates of the slope of the curve f(x) = exp(x)
*     at the point x = 0
*
      X = 0.0
      FX = EXP(X)
      DO 100 I = 0, 20
         H = (0.1)**I
         FXH = EXP(X + H)
         FPRIME = (FXH - FX)/H
         ERR = 1.0 - FPRIME
         WRITE(*, '(2E20.10)') H, ERR
  100 CONTINUE
*
      STOP
      END
```

This program evaluates forward difference estimates for the derivative of the exponential function at the point $x = 0$. The program is easily modified to provide a centered-difference approximation by adding an evaluation of $exp(x - h)$. The results of running these two computational experiments are diagrammed in Figure 15.3.8, where forward- and centered-difference results are graphed for two different precisions of FORTRAN REAL declarations (32-bit, which represents a lower limit of quality on floating-point numeric representation, and 80-bit, which represents a high-quality floating-point format). Important features presented in these two graphs are

- In each graph there is a region at the right where the error in replacing the differential operator with a difference approximation decreases with a characteristic linear behavior on the log-log plots. In this region, the difference estimates are *converging* to the exact value of tangent slope, as might be expected given the definition of the derivative as the limit of a secant.
- In these linear regions, the centered-difference error decreases with a slope of two, while the forward-difference error decreases with unit slope. This implies that the forward-difference approximation has a *linear* rate of convergence, while the centered-difference estimate is converging *quadratically.*
- The size of these convergent regions depends upon the precision used for floating-point calculations. Extended-precision formats (such as 80-bit reals) provide both a larger region of convergence and improved accuracy (obtained by being able to use smaller values of the step size h).
- To the left (smaller values of the step size h) of these linear regions, the convergent behavior of each estimate breaks down: upon passing through the minimum of each curve (which represents the most accurate estimate that can be obtained for a given precision by using these difference molecules), the effect of decreasing step size is *not* to decrease the error, but instead to ultimately *degrade* the approximation quality.
- Eventually, the denominator of each difference formula underflows (vanishes due to truncation error) as the values of $f(x)$, $f(x + h)$, and $f(x - h)$ become indistinguishable in terms of their respective floating-point representations. This implies that the estimate of the slope is computed as zero, and thus the error eventually tends to unity, which each curve approaches at the left edge [recall that $exp(0) = 1$].

This numerical experiment clearly shows the interplay of convergence rate, numerical precision, and truncation error. The centered-difference formula converges much more rapidly than the forward-differ-

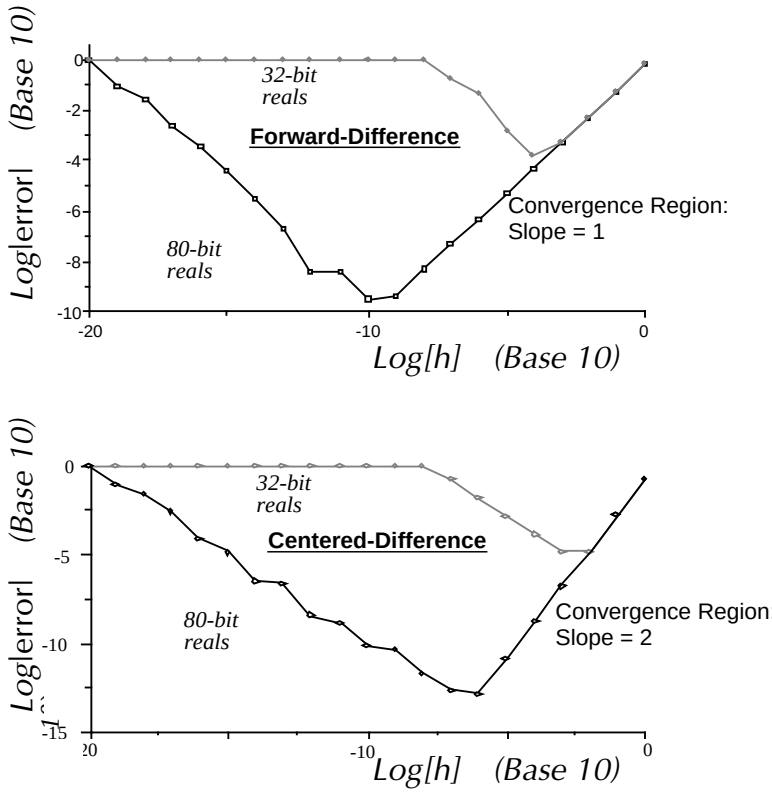


FIGURE 15.3.8 Results of finite-difference numerical experiment.

ence formula, and there exists an optimal step size for the various difference calculations where increasing the step size results in larger errors (in agreement with the rate of convergence of the approximation), and where decreasing the step size results in increased errors (due to the effects of round-off and truncation errors in the calculation of the derivative estimate).

Classification of Finite-Difference Models

As in the case of finite-element models, there are many alternative classifications available for categorizing finite-difference approximations. The first relevant classification is a geometric one, based on whether the finite-difference grid is *structured* (i.e., possesses a regularly spaced lattice structure) or *unstructured*. In order to maintain optimal convergence rates, finite-difference models must use regular spacings of grid points. This regularity is easy to maintain when the problem geometry is simple, but if the natural geometry is complicated, then the regularity of the difference grid will be compromised, with the potential for an attendant loss of convergence rate. In cases of irregular geometry, unstructured meshes are used, and if optimal convergence rates are to be realized, then a coordinate mapping from the actual problem geometry to a regularized computational domain must be constructed.

The problem with using structured grids on irregular or curved domains is diagrammed in Figure 15.3.9. Here a curved boundary is approximated using a low-order linear finite-element mesh (on the left of the figure) and a corresponding finite-difference/finite volume scheme (on the right). In order to capture the effect of boundary curvature, the structured finite-difference/finite-volume representation must approximate the boundary as a stepped pattern of rectangular subdomains. If the curvature is substantial, or if there is a considerable effect of regularizing the boundary on the physical model (a common situation in fluid and electromagnetic applications), this imposition of structure on what is

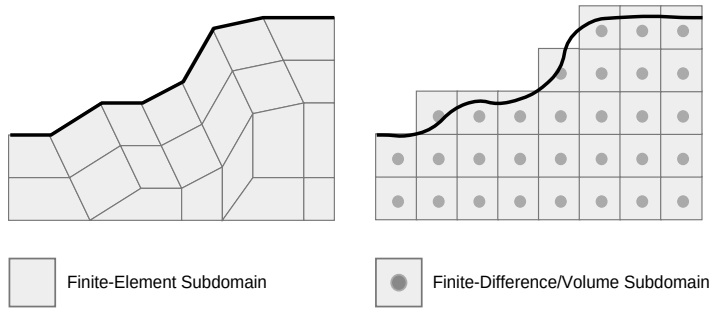


FIGURE 15.3.9 Finite-difference modeling of irregular geometry.

really an unstructured physical problem can seriously compromise the quality of the resulting computational approximation. Finite-element models, in which the presence of unstructured geometries does not substantially affect the solution quality, are thus more natural candidates for problems involving irregular boundaries or poorly structured problem domains. In fact, it is easy using higher-order finite-element approximations (e.g., locally quadratic or cubic interpolants) to approximate the boundary curvature shown in the figure to an essentially exact degree.

Another classification scheme can be based upon whether the difference formulas used are forward, backward, or centered. This classification is especially appropriate for finite-difference time-stepping methods used to solve initial-value problems in engineering and science. In this time-dependent setting, there is a coalescence of finite-element and finite-difference approaches, and the classification presented below permits one unified terminology to be used for both topics. Each class of difference estimate leads to time-marching discretization schemes that have distinctly different characteristics.

Forward difference schemes are commonly used to produce time-stepping methods that are very efficient in terms of computational effort and memory demand. The important class of *explicit* time-stepping algorithms arises largely from using forward-difference approaches: explicit schemes are defined as those requiring no equation solution effort to advance the solution to the next time step (e.g., where the coefficient matrix to be solved is diagonal). The primary disadvantage of forward-difference time-stepping is that these methods are generally only conditionally stable, so that they will converge only if the time step is relatively small. The use of explicit schemes thus involves a trade-off between reduced computational effort at each step, and the attendant requirement of a small time step. Explicit schemes based on forward-difference estimates for temporal derivatives are usually selected when the increased number of steps required is overshadowed by the decrease in the amount of effort required at each step. In some very large problems (such as complex three-dimensional models), explicit schemes are often used exclusively because of their relatively small demand for memory.

Backward difference schemes are commonly used in time-stepping applications where stability of the solution is especially important. In particular, backward-difference schemes generally provide numerical dissipation of high-frequency response, and this characteristic is often desirable in engineering practice. Many transient problems exhibit a wide range of natural time scales; for example, structural dynamics problems in aerospace engineering commonly include long-period vibrations of the entire structure as well as short-period response of individual structural components. One of the characteristics of discretization methods, in general, is that they tend to predict low-frequency behavior very accurately, while high-frequency response is predicted with considerably less accuracy. Since the high-frequency response is therefore often inaccurate (or spurious), there is a strong incentive to attenuate its effects, and choosing backward-difference temporal schemes is a good way to achieve this end. Backward-difference formulations generally require solution of a system of equations to advance the solution across a time step, and hence they are termed *implicit* time-stepping methods.

Centered-difference approximations are generally chosen in order to gain optimal convergence rates for time-stepping methods, and centered schemes are thus one of the most common choices for modeling

transient response. In practice, centered-difference schemes are a mixed blessing. They generally require equation solution to advance across a time step, so they are implicit schemes which require more computational effort and increased demand for memory. They also usually provide a higher convergence rate, so that larger steps can be taken without sacrificing accuracy. In problems where the time-dependent solution is expected to be "well behaved," such as transient heat conduction or similar diffusion problems, centered-difference methods are an excellent choice, especially if there are no spatial pathologies (such as material discontinuities or singular initial conditions) present. In problems where time-dependent response is more problematic (such as nonlinear dynamics of mechanical systems) many centered-difference schemes often give poor results, due to their general inability to attenuate spurious high-frequency discretization effects. There are a number of generalizations of centered-difference time-stepping schemes that have recently been developed to provide optimal accuracy while damping spurious high-frequency response, and these specialized schemes should be considered when complicated dynamic simulations are inevitable (see, for example, the relevant material in Hughes, 1987).

Computational Characteristics of Finite-Difference Models

Finite-difference methods possess many computational similarities to finite-element models, and so their computational characteristics are in large part similar. Both families of approximations generally result in sparse equation sets to be solved, but finite-difference models are less likely to preserve any symmetry of the governing differential equations in the form of matrix symmetry. In addition, many finite-difference molecules (such as those used to represent derivatives of order higher than two) result in equation sets that are not well conditioned and may give poor results when implemented in a low-precision, floating-point environment (such as some inexpensive microcomputers working in single precision).

Convergence rates for many families of finite-difference models can be established using appropriate Taylor series expansions of the solution function. These convergence estimates thus require the conditions under which existence of a Taylor expansion can be guaranteed, and may be compromised when the exact solution of the problem violates these existence conditions. In such cases, the precise rate of convergence of the finite-difference model is more difficult to estimate exactly, though numerical experiments on simple test problems can often be used to obtain estimates of the actual convergence rate. The most common examples of conditions where an appropriate Taylor series expansion may not exist include problems involving interfaces joining dissimilar materials, and singular problem data such as point loads or localized sources. In each of these cases, the exact solution of the underlying differential equation may not possess sufficient continuous derivatives to warrant a Taylor series expansion, and so the rate of convergence may be compromised from that obtained on better-behaved data. For many standard finite-difference schemes, this compromised accuracy is commonly localized around the region possessing the singular or discontinuous data, and outside this disturbed region, better accuracy is obtained. Estimating the extent of contamination of optimal accuracy is a task greatly facilitated by the use of integrated graphics applications for visualizing and verifying solution behavior.

It is worthwhile to contrast this latter potential convergence problem with the analogous finite-element case. Finite-element convergence proofs are generally cast in terms of least-squared error norms (such as the energy norm used for solid mechanics and diffusive flow problems) or weak convergence error measures. Each of these convergence settings is cast in an integral form, so few (if any) constraints upon the existence or continuity of the solution's derivatives are required. The end result of the dependence on integral convergence measures in finite-element modeling is that a substantially more robust mathematical framework exists for guaranteeing optimal convergence of finite-element models.

Advantages and Disadvantages of Finite-Difference Models

Finite-difference schemes possess many important computational advantages. Chief among these is the fact that finite-difference models are very easy to implement for a wide variety of problems, because much of the work of developing a finite-difference application consists merely of converting each derivative encountered in the mathematical statement of the problem to an appropriate difference estimate.

Because there are so many well-understood differential equations available in engineering and science, the scope of finite-difference modeling is appropriately broad.

Finite-difference models are also relatively simple to implement in software. Unlike finite-element models, which possess a bewildering generality that often complicates the task of software development, most finite-difference schemes are sufficiently limited in scope so that the task of computational implementation is less general and therefore considerably simpler. Because these models are easy to motivate and to implement, there is a vast body of existing finite-difference programs that can readily be applied to an equally wide range of problems in mechanical engineering. As long as the problems to be solved fit into the framework of finite-difference modeling, there is a strong incentive to use difference models in mechanical engineering practice.

The primary disadvantages of finite-difference modeling arise from engineering problems that do not fit neatly into the restrictive nature of finite-difference modeling. Real-world phenomena such as material dissimilarities, complex geometries, unusual boundary conditions, and point sources often compromise the utility of finite-difference approximations, and these complicating factors commonly arise in professional practice. Differential formulations of physical problems are inherently restrictive, as the conditions under which derivatives exist are considerably more stringent than those relating to integrability. The end result of relying on differential statements of physical problems is that finite-difference schemes are generally easier to implement on simple problems than are finite-element models, but that for more complex problems, this relative simplicity advantage disappears.

There are other disadvantages of finite-difference models including some important issues in verifying solution results. Since finite-difference models generally only calculate the primary solution parameters at a discrete lattice of sampling points, determining a global representation for the solution over the regions between these lattice points is not necessarily an easy task. If the grid used is sufficiently regular, a finite-element mesh topology can be synthesized and use made of existing finite-element postprocessing applications. If the grid is irregular, then automated schemes to triangulate discrete data points may be required before postprocessing applications can be used.

Another serious disadvantage of restricting the domain of the solution to a lattice of discrete points is that determination of secondary parameters (such as heat flux in thermal analyses or stress and strain in mechanical problems) becomes unnecessarily complicated. In a finite-element model, determination of secondary derived parameters is simplified by two properties: first, the solution is defined globally over the entire domain, which permits solution derivatives to be evaluated easily and accurately; and second, the convergence of the solution is generally cast into a mathematical framework (such as the so-called “energy norm”) that guarantees not only convergence of the primary solution unknowns, but of the derived secondary solution parameters as well. In many finite-difference applications, neither of these important finite-element advantages are easily realized.

Finally, another problem for general application of finite-difference models is in the implementation of boundary conditions. As long as the primary solution unknown is specified on the boundary (and the boundary is easily represented in terms of the finite-difference grid), then implementation of boundary conditions is relatively simple in theory. However, if derivative boundary conditions are required, or if the boundary itself is not readily represented by a finite-difference grid, then various “tricks” may be required to coerce the problem into a form amenable for finite-difference modeling (a particular example of a clever fix is the construction of phantom nodes located outside the problem domain specifically for the purpose of implementing derivative boundary conditions). Competing finite-element models generally permit trivial implementation of a wide spectrum of common boundary conditions.

The implementation of derivative boundary conditions in a simple finite-difference model (in this case, heat conduction, where the temperature function is denoted by u) is diagrammed in Figure 15.3.10. Here it is desired to implement an insulated boundary along a vertical section of the finite-difference grid, where $x = constant$, and hence the normal derivative is easily evaluated. In order to make the normal derivative vanish, which represents the condition of an insulated boundary, it is necessary either to utilize a reduced control volume (shown in gray in the figure) or to implement a fictitious solution

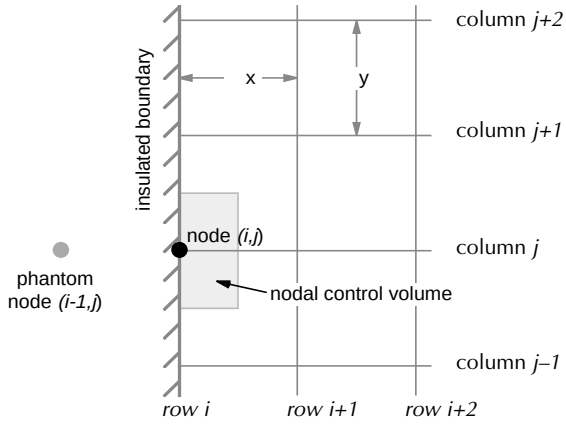


FIGURE 15.3.10 Finite-difference derivative boundary condition implementation.

node outside the problem domain (labeled as a phantom node in the figure). To simplify the resulting equations, it will be assumed that there is no heat generation taking place.

A brief overview of each implementation strategy will be presented below in order to highlight the general methods used in practice. It should be noted that implementing derivative boundary conditions on inclined or curved boundary segments is much more difficult than the case for this simple example. In addition, it is worthwhile to reiterate that derivative boundary conditions are common in engineering practice, as constraints such as insulated boundaries, specified sources, or stress-free surfaces are accurately modeled using derivative boundary conditions. Finally, it should be emphasized that *implementing derivative boundary conditions is a generally trivial task when using finite-element approximations*, a fact which lends considerable practical advantage to finite-element analysis methods.

Implementation of the derivative condition at node (i, j) in the figure is most easily developed using the rectangular control volume shown. If the steady-state conservation of heat energy is considered over this subdomain, then the heat flowing through the top, bottom, and right sides of the control volume must be balanced by heat generation within the volume and by flow through the boundary, and these last two terms are taken to be zero by assumptions of no internal heat generation and of an insulated boundary, respectively. The appropriate mathematical relation for heat conservation over the control volume is given by

$$k\frac{\partial u}{\partial y}\frac{\Delta x}{2}\bigg|_{\substack{\text{bottom}\\\text{face}}} - k\frac{\partial u}{\partial y}\frac{\Delta x}{2}\bigg|_{\substack{\text{top}\\\text{face}}} - k\frac{\partial u}{\partial x}\Delta y\bigg|_{\substack{\text{right}\\\text{edge}}} = 0$$

where k is the thermal conductivity of the material. This partial differential relationship among the heat flows can easily be converted to a discrete set of difference relations by substituting appropriate difference formulas for the partial derivatives, and letting u_{ij} denote the temperature computed at node (i,j). The result of this substitution is given by

$$k\frac{u_{ij} - u_{ij-1}}{\Delta y}\frac{\Delta x}{2} - k\frac{u_{ij+1} - u_{ij}}{\Delta y}\frac{\Delta x}{2} - \frac{u_{i+1,j} - u_{ij}}{\Delta x}\Delta y = 0$$

This single equation can be used to augment the standard difference equations for the nodes not located on the boundary, in order to embed the desired derivative boundary conditions into the complete finite-difference thermal model. For each node located on the insulated boundary (e.g., nodes $[i, j - 1]$, $[i, j]$, $[i, j + 1]$, and $[i, j + 2]$ in the figure), an appropriate difference relationship is synthesized and incorporated

into the set of difference equations governing the complete problem. Details of this procedure for the case of heat conduction can be found in Kreith and Bohn (1993). For more complex problems, appropriate derivations must be carried out to determine the correct form of the boundary relations for the finite-difference/finite-volume model.

In practice, the first two difference terms above are often combined to yield a more accurate centered-difference approximation for the solution. In this case, the third term, which captures the insulated boundary condition specification, can be seen to represent a backward-difference approximation for the boundary flux. Since such two-point backward difference schemes have been shown earlier in this chapter to possess less desirable convergence characteristics than centered-difference formulas, it is often advantageous to convert the insulated boundary derivative condition to a more optimal centered-difference form. One scheme for gaining a centered-difference derivative formula at node (i, j) can be developed by introducing the phantom node shown in the figure as node $(i - 1, j)$. A fictitious (in the sense that the node does not lie within the actual problem domain) thermal solution can be constructed at this node and utilized to gain a centered-difference relationship for the derivative boundary condition that may be more consistent with the relations used elsewhere in the finite-difference approximation. The details of this form of boundary condition implementation can be found in Ames (1977) or any other advanced text on finite-difference numerical analysis.

It is sufficient to note here that the analyst is often confronted with the choice of a feasible but less-than-optimal implementation (such as the control volume approach taken above), or a more accurate but appropriately more computationally complicated scheme for implementing a boundary derivative. Since finite-element models do not exhibit any of these difficulties in boundary-condition implementation, it is often much easier to utilize finite-element software for problems involving derivative boundary conditions.

Test Problem Suite for Finite-Difference Software Evaluation

Before purchasing or using a new finite-difference program, appropriate evaluation measures should be contemplated to insure that the proposed application is capable of solving the particular problems encountered in mechanical engineering practice. As in the case of evaluation of finite-element software, any known solution pathologies can be modeled using the new program in order to determine the quality of the solution scheme. Many of the same mathematical pathologies presented for finite-element evaluation (such as mixed problems arising from incompressible flow constraints) can occur in finite-difference models, so these standard tests may also be appropriate in the finite-difference setting. In addition, any problems commonly encountered in professional practice should be solved with the new code in order to verify that the proposed application is both easy to use and reliable.

Integrated graphical pre- and postprocessors are also important for use with finite-difference programs. Preprocessors that can plot both structured and unstructured finite-element grids are especially valuable, as verification of the coordinate transformations required to map complex physical domains to a regularized coordinate system may necessitate plotting of the finite-element grid in both its irregular physical setting and its regularized transformed computational setting. Postprocessors that automatically determine appropriate secondary solution parameters from the difference solution lattice are especially attractive, as they provide additional solution intepretation functions while minimizing the extra work required on the part of the engineer to construct the desired derived solution function.

If computational simulations of advection-dominated problems are to be generated, it is imperative to insure that the proposed application provides robust algorithms for handling the nonlinear effects of convective acceleration terms. Older computational technology (such as simple upwinding schemes used to more accurately model advective effects in fluid dynamics) often results in poorer solution quality than can be obtained from codes utilizing newer schemes from computational fluid dynamics. Comparing results from various computer programs is a good way to determine which programs implement the best means to capture nonlinear effects. In addition, if the application being evaluated is noncommercial (e.g., public domain codes commonly distributed by academic researchers or special-purpose applications developed within an engineering organization), these comparisons will help to determine whether the

application is capable of robust modeling of the engineering problems encountered in normal professional practice.

Alternative Numerical Schemes for Mechanics Problems

Finite-element and finite-difference approximations are the most common computational schemes used in mechanics, but the practice of computational mechanics is not restricted to these two discretization approaches. There are a variety of alternative methods available for solving many important problems in mechanical engineering. Some of these alternatives are closely related to finite-element and finite-difference models. Others have little in common with these general discretization schemes. A representative sampling of these alternative approaches is presented in the following sections.

Boundary-Element Models

Boundary elements are a novel generalization of finite-element technology to permit better computational modeling of many problems where finite-element methods are unduly complex or expensive (see Brebbia and Dominguez, 1989). In simple terms, boundary elements involve recasting the integral principles that lead to finite-element models into a more advantageous form that only requires discretizations to be constructed on the boundary of the body (hence the term "boundary element"), instead of over the entire solution domain. For many problems involving complicated boundaries, such as infinite domains or multiply connected three-dimensional shapes, boundary element technology is an attractive alternative to finite-element modeling. In addition, because many existing computer-aided drafting applications already idealize shapes in terms of enclosing boundary representations, boundary-element approaches are natural candidates for integration with automated drafting and design technologies.

Boundary-element schemes are not without disadvantages, however. The recasting of the integral formulation from the original domain to the boundary of the problem is not a trivial task, and this conversion substantially complicates the process of developing new boundary-element technology. One of the most important factors limiting the applicability of finite-element models is the formulation of an integral principle for a mechanical problem that is equivalent to a more standard differential problem statement. The conversion of this integral statement into a form amenable to boundary element approximation is yet another laborious task that must be performed, and each of these difficult mathematical problems represents potential errors when implemented in the setting of developing general-purpose analysis software. Many of the potentially attractive capabilities of boundary-element technology, such as the goal of accurate modeling of singular solution behavior (e.g., stress concentrations) that commonly occur on the boundaries of the domain, have not yet been completely realized in existing boundary-element software. However, it is clear that boundary-element models are promising candidates for future computational approximation schemes.

At present, boundary-element codes are most commonly found in research applications, as there is not yet a full range of commercial quality boundary element programs available for widespread use by practicing engineers. While the boundary-element method has a promising future, other advances in finite-element technology (such as robust automated mesh generation schemes that synthesize internal element information from a simple boundary representation) may compromise the demand for the considerable theoretical advantages of boundary element models. It is already feasible to combine features of finite-element and boundary-element technology in the same code, so the future of boundary element techniques may lie in integration with (instead of "replacement of") finite-element applications.

Finite-Volume Schemes

Finite-volume methods are hybrid computational schemes that have been developed to ameliorate some of the problems of finite-difference models. In simple terms, finite-volume models represent integrated finite-difference approximations, where the underlying differential equation is integrated over individual control volumes in order to obtain recurrence relations among discrete solution values. These recurrence formulas are strikingly similar to standard finite-difference approximations, but this integrated approach possesses some clear advantages over related finite-difference technology:

- Because the equations are integrated over control volumes, finite-volume schemes are better able to model material discontinuities, which may cause serious problems in standard finite-difference models.
- The integration process permits rational modeling of point sources, which are commonly implemented in an ad hoc manner in finite-difference schemes.
- The integration process permits more flexibility in modeling unusual geometries, as the individual integration volumes can be adjusted locally to provide such important effects as curved boundaries.

While these are important generalizations for finite-difference theory, in practice they do not realize many of the advantages available with finite-element models. Finite-volume schemes still require substantial regularity in the construction of the solution grid, or the construction of appropriate coordinate transformations to map the problem domain to a simpler computational setting. The integration process provides no relevant aid in the construction of secondary solution parameters such as stresses and strains. Finite-element topology must still be synthesized to enable use of appropriate solution visualization schemes, and this synthesis may be difficult when complex unstructured grids are utilized for the finite-volume model. Therefore, finite-volume schemes are most likely to be found competing with existing finite-difference models in areas where the underlying differential statements of the problem are well established, but where increased modeling flexibility is required.

Finally, the convergence setting for finite-volume methods is similar to that of finite-difference models, and the various integral convergence measures available for proving finite-element convergence are not so readily applied to finite-volume approximations. Thus finite-volume models provide a way to generalize finite-difference approximations to more realistic problem settings, but their limited utilization of integral principles does not provide a unified integral framework of the quality found in finite-element modeling applications.

Semi-Analytical Schemes

Many problems in mechanical engineering still admit solutions based on computational implementation of analytical (i.e., nonnumeric) techniques originally used for hand computation using closed-form solutions. This approach of approximating reality by simplifying the model to the point where it admits a closed-form solution (as opposed to the standard computational approach of putting a complicated model on the computer and using numerical approximations based on finite-element or finite-difference relations) has a rich history in engineering and science, and it includes such classical techniques as Fourier Series expansions for heat conduction problems, representation by orthogonal polynomials, and other eigenfunction expansion schemes.

The advent of high-performance computers has had two effects on semi-analytic schemes such as Fourier Series approximation. First, modern computers are capable of performing the trigonometric evaluations required by Fourier Series representations very rapidly, so that these eigenfunction expansion schemes are now computationally feasible to implement. On the other hand, the application of modern computers to competitive approximation schemes (such as finite-element or finite-difference models for heat conduction applications formerly solved using closed-form Fourier Series techniques) has obviated the need to consider using semi-analytical schemes, as their more general computational approximation competition are now capable of solving more complicated problems with similar amounts of computational effort.

While some applications of semi-analytical techniques can still be found, these computational implementations tend to be localized into special-purpose noncommercial codes. These classical schemes are still useful tools for such niche uses as verification of more complex computational models, but there is little likelihood that these older closed-form schemes will one day become more important relative to modern computational approaches.

Nondeterministic Computational Models

All of the computational models presented so far are completely deterministic: the computer program constructs one solution for every set of input data, and the output is completely determined by the values of the input data set (except, of course, for some differences that may be encountered when using a different computer or changing to a different precision for floating-point computation on the same computer). In some problems, the input data can be characterized with sufficient precision so that these deterministic models are sufficient for analysis. In other problems, however, the input data are not known precisely, but are instead given only in a probabilistic sense. In these latter cases, it is more reasonable not to speak of the precise values that define the solution (such as the nodal displacements of a finite-element model), but instead to consider the problem of determining the probabilistic *distribution* of these solution components.

This nondeterministic approach to finding computational solutions in a probabilistic sense is termed *stochastic* modeling. Stochastic modeling of mechanical engineering systems is still an emerging field, though related engineering disciplines have seen substantial developments in this area. For example, in the modeling of structural systems subjected to earthquake loads (a field of civil/structural engineering very similar to structural dynamics applications found in mechanical engineering practice), nondeterministic models are commonly used to represent the fact that the precise excitation of an earthquake is never known *a priori*. Engineers concerned with the reliability of structures subjected to earthquake loads often resort to stochastic modeling techniques to estimate the probabilistic distribution of forces in a structure in order to design the individual component structural members.

One area of considerable interest is in developing stochastic finite-element models that will permit using much of the existing knowledge base of deterministic finite-element models toward more general nondeterministic computations that can address directly the probabilistic distribution of finite-element response. One of the main constraints to the practical use of such schemes is that large computational models (e.g., ones with substantial numbers of displacement components) generate stochastic approximations that are computationally intractable because of the immense computational expense required to generate and solve the resulting stochastic equation set. At present, only relatively small problems are commonly solved via stochastic models, but this situation is reminiscent of early developments in computational implementation of finite-element and finite-difference schemes. As computer performance continues to improve, so will the practical utility of stochastic models in engineering.

Time-Dependent and Nonlinear Computational Mechanics

Many important problems in mechanical engineering are neither static nor linear-elastic. Mechanical systems undergoing dynamic response or materials deformed past their elastic limit require fundamentally different solution methods than do simpler static or elastic cases. In practice, static and elastic models are often used for initial analysis purposes, and depending upon the analysis results or the solution accuracy required, more complex time-dependent or nonlinear models may be subsequently analyzed. In cases that warrant these more accurate and difficult analysis methods, it is generally a good idea to perform simple static linear analyses first, in order to gain physical intuition into the more difficult problem, and to remove gross modeling errors (such as inconsistent discretizations) before more complex and expensive analyses are performed.

Practical Overview of Time-Dependent and Nonlinear Problems

While it may seem unusual to lump nonlinear and time-dependent problems into a single framework, in practice there are many common features in implementing these two disparate solution approaches. In each case, the more complex problem is solved by looping over a sequence of equivalent pseudostatic or quasilinear computational problems in order to approximate time-dependence or nonlinear behavior. In the case of time-dependent problems, the looping amounts to stepping the solution over a discretization of the time domain; in the nonlinear case, the looping represents an iteration scheme used to trade in a complex nonlinear problem for a series of simpler linear ones. In many problems, the response is both

time-dependent and nonlinear, and then nested looping schemes are generally utilized, requiring internal iterations to solve each nonlinear step composed within a larger time-marching scheme used to advance through the temporal domain of the problem.

Time-dependent problems in computational mechanics are generally solved by discretization of the problem's history into a set of distinct times, and the time-dependent analysis is carried out by stepping the solution between these discrete times. At each step, the temporal variation of the solution is combined with discrete estimates for the temporal derivatives inherent in the problem, and these composed approximations are utilized to transform the advancement across the time step into a computational form that resembles a traditional static problem. In this manner, the sequential nature of the time-dependent problem is transformed into a series of equivalent pseudostatic problems.

Nolinear problems are inherently difficult to solve, as they are commonly ill posed. Nonlinearities are generally accompanied by problems of nonuniqueness (recall that two lines intersect in at most one point, but two curves may intersect at many), and this nonunique character of the solution leads to considerable concerns as to which (if any) of the solutions represents the "real" (feasible) physical state of the problem. In addition, there is little general theory underlying nonlinear solution techniques, and so concrete guarantees that a nonlinear iteration process will converge to a desired solution are generally difficult to obtain.

Test Problems for Time-Dependent Program Evaluation

The most important issues relevant for effective solution of time-dependent problems have already been presented and are reiterated here in a high-level setting. In particular, issues of stability and convergence are especially relevant to practical time-stepping schemes, and insuring that the computed solution history is free from unstable artifacts or inaccurate components is the most important aspect of the practical use of time-stepping algorithms.

The basic trade-off in time-dependent solution processes is between the desire to minimize the number of time steps required for a particular solution history and the need to minimize the amount of computational effort required to construct the solution at each individual time step. Ideally, it would be possible to achieve both goals simultaneously, but this desirable state is seldom reached in practice. In general, increasing the length of the time step involves performing more work at each step, while minimizing the effort at each individual step is accompanied by a substantial increase in the number of steps required to construct the entire solution history. Understanding the interplay of these two desirable goals constitutes much of the practice of solving time-dependent problems. Since many commercial programs that support time-dependent analysis methods provide some form of automated selection of the time step, it is imperative for the analyst to have some idea of the trade-offs involved in the use of these various time-stepping approaches.

Implicit schemes, such as those based upon centered- or backward-difference schemes for approximating temporal derivatives, generally permit taking larger time steps than do explicit schemes, which are often based upon forward-difference approximations. Implicit schemes also require that sets of simultaneous equations be solved to advance the solution across a time step, where explicit schemes require little or no effort of this kind. Therefore, implicit schemes permit fewer steps to be taken with more work per step, where explicit approaches permit rapid advancement of the solution across an attendant larger set of time steps. In general, implicit schemes tend to be more stable than explicit schemes: most practical implicit schemes are unconditionally stable, in that their stability properties do not depend upon the time-step size. Explicit schemes tend to be conditionally stable, with a maximum step size given by a so-called *Courant condition*: step sizes larger than the critical threshold will result in solution oscillations that can easily dwarf the desired solution response.

Centered-difference schemes for time-dependent solution, such as the Crank-Nicolson method used to solve time-dependent heat conduction problems (Ames, 1977) or the Newmark algorithm used to solve structural dynamics problems (Hughes, 1987), are attractive numerical approaches because they possess optimal (quadratic) convergence rates for their time-dependent accuracy behavior. Unfortunately, these optimal error properties are commonly associated with an inability to damp spurious oscillations

that are generally present in discrete computational models. The problem is less important in heat conduction and other dispersive problems, as the spurious oscillations are generally damped over time. In dynamics applications, however, these oscillations can eventually grow and contaminate the solution, and this effect is especially pronounced in the important nonlinear time-dependent case. For this reason, various schemes have been proposed to generalize the more naive centered-difference approach to permit new schemes that damp unwanted high-frequency artifacts but preserve optimal convergence properties (for example, the Hilber-Hughes-Taylor method is a simple extension of Newmark's algorithm, but removes virtually all of the oscillation problems encountered with the latter scheme).

One of the best ways to determine whether the solution accuracy is being eroded by stability or accuracy concerns is to utilize appropriate visualization programs wherever possible. Undesirable or spurious artifacts are readily observed on many graphical displays of solution history, and such solution visualization schemes should be used whenever possible.

Classification of Nonlinear Mechanical Response

Many important problems in computational mechanics are nonlinear. Typical examples include:

- Nonlinear materials, where the material parameters (for example, the material compliances) depend upon the displacements
- Nonlinear geometry, where large deformations require consideration of deformed geometry in the strain-displacement relations, or where Eulerian coordinate formulations are used (as in fluid mechanics problems, where convection introduces strong nonlinearities into otherwise well-posed problems)
- Nonlinear boundary conditions, where the value or type of the boundary condition depends upon the solution of the problem

The different types of nonlinearities often lead to disparate solution qualititative behavior. For example, many material nonlinearities (such as the yielding of metals due to plastic deformation) tend to produce nonlinear problems that are reasonably well posed, in that they are relatively simple to solve and tend not to have serious problems with nonuniqueness of the solution or convergence of the nonlinear solution scheme. Alternatively, many geometrically nonlinear problems (such as high Reynolds' number flow, or arbitrarily large deformations of materials) are considerably more difficult to solve effectively, owing to the presence of alternative (but infeasible) solutions and strong nonlinear effects.

General Overview of Nonlinear Solution Schemes

It is possible to have problems where material and geometric nonlinearities are present simultaneously, and this situation is made commonplace by the fact that inelastic material nonlinearities are often accompanied by plastic flow, which may lead to large deformations that necessitate the use of nonlinear strain-displacement relations. Therefore, it is appropriate to consider solution schemes for nonlinear problems that do not depend upon the particular source of the nonlinearity.

There are many ways to solve nonlinear problems — it often seems like there are *too* many schemes, reflecting the important fact that *there is little general theory for solving nonlinear problems*. Even though there are a wide variety of approaches that can be used to motivate nonlinear solution schemes, one basic fact emerges from all of them: *nonlinear problems are generally solved using an iteration scheme*.

The most common nonlinear iteration schemes are based on *Newton's method*, which provides a relatively simple way to convert an intractable nonlinear equation solution process into a sequence of linear solution processes. The ultimate goal is that this sequence of linear solutions will tend to a fixed solution, and that this solution will solve the underlying nonlinear problem. Thus, Newton's method is based upon a linearization of the nonlinear problem, and it is an especially effective approach in practice because the sequence of linear solutions will often converge quadratically to the nonlinear solution (recall that quadratic rates of convergence are especially attractive, as these schemes cause the error to go to zero rapidly once a good approximation is found).

Finally, a warning on nomenclature ought to be mentioned. Some authors reserve the term *Newton's method* for the iteration scheme resulting from the minimization of a nonquadratic functional such as those that govern many nonlinear problems in solid mechanics. These authors refer to the analogous technique of solving a nonlinear system of equations (which often may not result from a functional extremization problem, as in the case of the Navier-Stokes equations) as *Newton-Raphson iteration.*

Test Problems for Nonlinear Program Evaluation

Just as there is little general theory for solving nonlinear problems, there is little general practice that can be cited to aid the engineer contemplating effective estimation of nonlinear response. Some basic principles based on empirical observations can be given, however, and adherence to these ad hoc guidelines will aid the practicing engineer working in this interesting but difficult field.

The most important observation is that it is unusual to be able to solve a complicated nonlinear problem unless one can also solve a simpler version with the intractable nonlinearities removed. This leads to the concept of "sneaking up on" a nonlinear solution by first performing a sequence of problems of increasing modeling complexity and (hopefully) increasing accuracy. For example, in order to find gross modeling errors (such as incorrect discretizations or geometric inconsistencies), it is sufficient to generate a discretization mesh for the problem, but to use a simpler linear model for approximating material or geometric nonlinearities. Once the gross modeling errors have been removed from these simpler (and computationally less expensive) analyses, more accurate material or geometric models can be constructed from the simpler input data. The results of the simplified analyses may often then be reused to verify the overall mechanical response of the final nonlinear solution.

Another important way to verify nonlinear solution behavior occurs in the setting of time-dependent problems, where the nonlinearities evolve over time from an initial linear state. Most time-dependent algorithms are obtained by trading in a complicated historical problem for a sequence of quasistatic solutions, and at each time step, a standard finite-element or finite-difference quasistatic problem is to be solved (this is especially true when implicit time-stepping schemes are used). In this setting, it is desirable to be able to use large time steps so that the entire solution history can be obtained with as little effort as is possible. However, in many problems (most notably, nonlinear stress analyses), the effect of the nonlinearities on each individual quasistatic problems decreases with the size of the time step. *As the time step gets smaller, the time-dependent component of the solution process becomes more expensive, but each nonlinear iteration scheme generally becomes easier to solve.* Therefore, when solving nonlinear time-dependent processes, it is important to balance the desire to minimize the number of time steps taken with the goal of more rapid convergence of the nonlinear iterations required at each solution step.

Finally, the use of integrated graphics postprocessing to verify solution behavior is incredibly important in the setting of solving nonlinear problems. The problems of nonuniqueness, localization of response, and solution instabilities are commonly encountered in nonlinear analysis settings, and the use of solution visualization techniques (combined with a good sense of engineering judgment and physical intuition) greatly simplifies the task of dealing with these otherwise intractable difficulties.

Standard Packages Criteria

Poor computer acquisition policies can result in spectacular losses for a variety of public and private organizations. Many of these failures arose because empirical concepts developed in the computer industry were not given proper consideration. Guidelines for acquisition of computer hardware and software are outlined below.

Functional Specifications

Regardless of whether an application is to be purchased off the shelf or developed for custom use, the process of acquiring computer software begins with a functional specification of the role of the software. The end result of this process is termed a "functional requirements document", and it is a high-level nontechnical enumeration of the goals ("functions") that the software package is designed to meet. The

language used in a functional specification is normally reasonably devoid of any references to specific technological issues. For instance, if the purchase of an application to aid in document preparation at the "XYZ Engineering" company is contemplated, the functional requirements might include such phrases as "be able to import existing documents for future modification" and "provide support for various printers and other output devices normally used at XYZ Engineering." A good functional specification would not be characterized by "XYZ Engineering should purchase the MicroWizard DocuStuff Version 9.4 software to run on a new WhizBang 9000 Supercomputer:" precise specifications of hardware and software are normally not relevant in a functional specification unless they are used only to indicate the need for back compatibility with existing computer systems.

A functional specification should provide a view of the needs of the organization that are to be satisfied. By stating these needs in broad terms, the functional specification admits the widest possible range of hardware and software solutions. Since computer technology changes rapidly, it is essential to keep technical details from creeping into the functional requirements, as these technical requirements may cause useful alternative measures to be overlooked. When the procurement process is for a government agency (where the lead time between identifying the need and purchasing the software or hardware may be a year or more), it is imperative to insulate the statement of need from technical details that may change rapidly over the course of the acquisition process, and functional specifications are an excellent approach toward this important goal.

Technical Specifications

The next step in the computer materials acquisition process is the translation of the functional requirements to a technical specification. This step is generally performed immediately before bids are sought, so that the most current technical information will be available for review. All of the relevant technical data that was carefully excluded from the functional specification belong in the technical specification, but care should still be taken to avoid unwarranted detail. It is especially important to avoid placing technical information that is inaccurate into the technical specification, or else the resulting procurement process may be compromised.

Evaluation and Testing

Once the technical specification has been finalized, it is either sent to vendors for bidding (in the case of large purchases, where the work of matching specific models and manufacturers of computers can be offloaded to the vendor) or handled internally (for small purchases), where local expertise is used to match the technical requirements with a range of computer models.

One important component of the evaluation process for software procurement is regression testing, where existing data sets are migrated to the new application (or computer), analyzed, and output results compared. This process

- Aids in evaluating whether the application is easy to use
- Provides verification of the software application's capabilities and accuracy
- Gives estimates of performance for the new application

Regression testing of computational mechanics software is greatly aided by the use of integrated visualization tools. Comparing the outputs of complex simulations from two different codes (or even two different revisions of the same code) can be an incredibly tedious process, but the use of graphical display tools provides considerable help toward accurate implementation of the regression testing process.

Overview of Computer Models for Mechanical Engineering Practice

The following outline classifies families of computational models in terms of common mechanical engineering problem characteristics. It emphasizes general "rules of thumb". Although there are instances in which the recommendations below are not optimal, the outline below represents a good starting point for picking the right computational tool for a given task.

- *Simplicity*: Finite-difference models are among the simplest computer models, as they can be derived and implemented directly from the governing diffferential equation whenever the differential form of the boundary-value problem is known. If it is desired to develop, implement, and test a computer model in a hurry, and the problem is well defined in terms of differential equations, finite-difference schemes are a very powerful simulation tool.
- *Generality*: Finite-element models are the most general computational schemes available for mechanical engineers. The mathematical framework of finite-element theory makes it easy to solve coupled problems, as well as problems with discontinuous and singular data. When the underlying problem is intractable in terms of its complexity or material constitution, then finite-element models are generally the first choice for solution.
- *Data definition*: Depending upon how the data are defined, one computer model may be better than others. If all of the data defining the problem domain are given in terms of surfaces (such as would be obtained from some surface-based CAD solid modeling applications), then boundary element technology may be a good choice if the underlying boundary element package can handle the problem encountered. Alternatively, a finite-element package that automatically generates internal element definition from the surface data would be a good choice for such surface-defined modeling. If the data are specified naturally on an unstructured grid, then finite-element schemes generally are easier to attempt than are finite-difference and finite-volume models.
- *Availability*: It is generally a lot easier to use a commercial program to solve a complex mechanical engineering problem than it is to develop the requisite modeling capabilities locally. Therefore, one of the most important starting points for solving complex problems is to become aware of the range of existing applications available. The next section provides an outline guide to a wide spectrum of commercial and public-domain computer-aided engineering tools.

Sample Mechanical Engineering Applications

The following tables (Tables 15.3.2, 15.3.3, 15.3.4, 15.3.5, and 15.3.6) represent a sampling of available software typical mechanical engineering applications. The tables include software for the following tasks:

- Computational solid mechanics and structural analysis
- Computational fluid mechanics
- Computational electromagnetics
- Computational thermal analysis
- Utility software for mechanical optimization and design

These tables only represent a sampling of the state of the software market at the time of publication of this handbook. Since the world of computation and software development proceeds at a rapid pace, it is a good idea to check the current literature (and the addresses listed) to gain a better idea of up-to-date information. Finally, there are several abbreviations given in the tables that are enumerated below:

- FE — finite-element application
- FD — finite-difference application
- FV — finite-volume application
- BE — boundary-element application
- AN — analytical or semi-analytical application
- OPT — design optimization utility (often listed with preferred computational model)

Selection of Software — Benchmarking

Benchmarks are software applications that are used to compare the performance of various computers, or of various operating environments on the same computer. While benchmarks provide a simple means of characterizing computer speed, they are commonly misused and often misunderstood. In particular,

TABLE 15.3.2 Sample Solid Mechanics Applications

Program	Type	Applications	Contact Address
ABAQUS	FE	General solid mechanics, heat transfer, and coupled thermal/stress problems	HKS, Inc. 1080 Main Street Pawtucket, RI 02860
ADINA	FE	General solid mechanics, heat transfer, and coupled thermal/stress problems	ADINA 71 Elton Avenue Watertown, MA 02172
ANSYS	FE	General solid mechanics, heat transfer, and coupled thermal/stress problems	ANSYS, Inc. Johnson Road, P.O. Box 65 Houston, PA 15342-0065
BEASY	BE	Linear-elastic stress analysis and potential (e.g., thermal) problems	Computational Mechanics, Inc. 25 Bridge Street Billerica, MA 01821
HONDO	FE	Static and dynamic stress analysis	Energy Science & Tech. Software Ctr. P.O. Box 1020 Oak Ridge, TN 37831-1020
LS-DYNA3D	FE	3D explicit solid dynamics	Livermore Software Tech. Corp. 2876 Waverley Way Livermore, CA 94550
MARC	FE	General solid mechanics, heat transfer, and coupled thermal/stress problems	MARC Analysis Corp. 260 Sheridan Avenue, Suite 309 Palo Alto, CA 94306
MSC/DYTRAN	FE	Coupled solid-fluid transient interactions	MacNeal-Schwendler Corp. 815 Colorado Boulevard Los Angeles, CA 90041
MSC/NASTRAN	FE	General solid mechanics, heat transfer, and coupled thermal/stress problems	MacNeal-Schwendler Corp. 815 Colorado Boulevard Los Angeles, CA 90041
NIKE 2D/3D	FE	2D/3D implicit solid dynamics	Energy Science & Tech. Software Ctr. P.O. Box 1020 Oak Ridge, TN 37831-1020
NISA II	FE	General solid mechanics, heat transfer, and coupled thermal/stress problems	Engineering Mechanics Research P.O. Box 696, 1607 East Big Beaver Troy, MI 48099
STAAD III	FE	Structural analysis and design	Research Engineers, Inc. 22700 Savi Ranch Yorba Linda, CA 92687
STARDYNE	FE	Static and dynamic structural analysis	The Titan Corporation 9410 Topanga Canyon Boulevard, Suite 103 Chatsworth, CA 91311

many standard benchmarks provide information that is not always relevant to the performance issues encountered in mechanical engineering practice. Furthermore, it is easy to generate benchmark figures that are misleading by citing results out of context, since computer performance is often strongly dependent on a variety of factors not immediately apparent from simple benchmark comparisons. Therefore, in this section some representative schemes for comparing computer speed are presented to aid in the important task of accurately gauging computer performance for mechanical engineering practice.

Choice of Problem Test Suite for Benchmarking

There are many standard benchmarks commonly used in mechanical engineering applications to estimate the relative performance of various aspects of computers. In particular, there are three common types of numeric operations utilized in engineering computation:

TABLE 15.3.3 Sample Fluid Mechanics Applications

Program	Type	Applications	Contact Address
ADINA-F	FE	Incompressible viscous fluid flow	ADINA 71 Elton Avenue Watertown, MA 02172
ANSWER	FV	Laminar or turbulent fluid flow	Analytic and Computational Research 1931 Stradella Road Bel Air, CA 90077
ANSYS	FE	Fluid flow coupled analyses (among many other capabilities)	ANSYS, Inc. Johnson Road, P.O. Box 65 Houston, PA 15342-0065
ARC2D	FD	Navier-Stokes equations in 2D	COSMIC University of Georgia 382 East Broad Street, Athens, GA 30602
FDL3D	FD	Navier-Stokes equations (various explicit/implicit schemes used)	Wright Laboratory WL/FIM Wright Patterson AFB, OH 45433
FIDAP	FE	Fluid flow, heat and mass transfer	Fluid Dynamics International 500 Davis Street, Suite 600 Evanston, IL 60201
FLOW-3D	FV	Fluid dynamics and heat transfer	Flow Science, Inc. P.O. Box 933, 1325 Trinity Drive Los Alamos, NM 87544
FLUENT	FV	Navier-Stokes equations	Fluent, Inc. Centerra Resource Park 10 Cavendish Court, Lebanon, NH 03766
INCA/3D	FV	Navier-Stokes equations	Amtec Engineering, Inc. P.O. Box 3633 Bellevue, WA 98009
MSC/DYTRAN	FE	Coupled solid-fluid transient interactions	MacNeal-Schwendler Corp. 815 Colorado Boulevard Los Angeles, CA 90041
NEKTON	FE	3D incompressible fluid flow, heat transfer	Fluent, Inc. Centerra Resource Park 10 Cavendish Court, Lebanon, NH 03766
NISA/3D-FLUID	FE	General fluid flow and heat transfer	Engineering Mechanics Research P.O. Box 696, 1607 East Big Beaver Troy, MI 48099
RPLUS	FV	Navier-Stokes flows, reacting flows	NYMA 2001 Aerospace Parkway Brook Park, OH 44136
SPEAR/3D	FV	Navier-Stokes equations	Amtec Engineering P.O. Box 3633 Bellevue, WA 98009
TEMPEST	FD	3D fluid dynamics and heat transfer	BATTELLE Pacific NW Laboratories P.O. Box 999 Richland, WA 99352
VSAERO	AN	Subsonic aerodynamics	Analytical Methods, Inc. 2133 152nd Avenue N.E. Redmond, WA 98052

- Integer arithmetic, involving the calculation of sums, differences, products, and quotients of numbers that represent signed integral quantities (i.e., numbers that have no decimal point or exponent in their numeric representation)
- Floating-point arithmetic, which involves the same basic arithmetic calculations as integer arithmetic, but with operands obtained from numbers represented by the composition of sign, mantissa, and exponent (although the computer represents floating-point numbers internally in base 2 format,

TABLE 15.3.4 Sample Electromagnetics Applications

Program	Type	Applications	Contact Address
ANSYS	FE	General electromagnetics (as well as solid, thermal, and fluid analysis)	ANSYS, Inc. Johnson Road, P.O. Box 65 Houston, PA 15342-0065
EMA3D	FD	Maxwell's equations	Electro Magnetic Applications, Inc. 7655 West Mississippi Avenue Suite 300 Lakewood, CO 80226
FLUX2D/3D	FE	Static and dynamic electromagnetic fields	Magsoft Corp. 1223 Peoples Avenue, J Bulding Troy, NY 12180
MAX3D	FV	Time-dependent Maxwell's equations	Wright Laboratory WL/FIM Wright Patterson AFB, OH 45433
MAXWELL3	FE	Time-dependent Maxwell's equations	Energy Science & Tech. Software Ctr. P.O. Box 1020 Oak Ridge, TN 37831-1020
MSC/EMAS	FE	Maxwell's equations for antennas and electrical systems	MacNeal-Schwendler Corp. 815 Colorado Boulevard Los Angeles, CA 90041
NISA/EMAG	FE	General electromagnetic problems	Engineering Mechanics Research P.O. Box 696, 1607 East Big Beaver Troy, MI 48099
TOSCA	FE	3D electrostatics and magnetostatics	Vector Fields, Inc. 1700 North Farnsworth Avenue Aurora, IL 60505

it is generally easier to convert these binary quantities to an equivalent base 10 form, so that they are similar to more standard "scientific notation" for numeric representation used in engineering)

- Transcendental computation, involving floating-point results utilizing more complex mathematical functions, such as trigonometric functions, roots, exponentials, and logarithms (transcendental calculations can be loosely defined as those that cannot be implemented using a finite number of algebraic computational operations)

Each of these classes of numeric computation has particular characteristics: for instance, integer arithmetic is the simplest and fastest type of numeric calculation, as this class of operand does not require manipulation of exponents and mantissas (which often must be renormalized after calculations). Transcendental calculations are the most expensive native numeric operation on computers appropriate for engineering computation, and the speed of this class of operation is often independent of the performance of integer and floating-point arithmetic (owing to the presence and quality of specialized transcendental hardware available to speed up this class of computation). Most algorithms involved in any branch of computing involve substantial integer arithmetic, so nearly all aspects of computing performance increase as integer arithmetic improves. Many common forms of computer use (such as word processing or some drawing packages) require little or no floating-point calculation, so these applications have performance characteristics largely independent of the speed of a computer's floating-point arithmetic hardware.

Most mechanical engineering applications (and especially computational mechanics software) are highly dependent on floating-point performance issues, and so measuring the floating-point speed of computers is generally important in mechanical engineering settings. While many floating-point applications in mechanical engineering involve negligible transcendental calculation (for example, most finite-element or finite-difference schemes involve little or no transcendental results), there are some particular software packages that require substantial amounts of transcendental computation (for example, Fourier Series approximations, which require large-scale evaluation of trigonometric functions). Depending upon

TABLE 15.3.5 Sample Thermal Applications

Program	Type	Applications	Contact Address
ABAQUS	FE	Heat transfer, as well as solid mechanics and coupled thermal/stress problems	HKS, Inc. 1080 Main Street Pawtucket, RI 02860
ADINA	FE	Heat transfer, as well as solid mechanics and coupled thermal/stress problems	ADINA 71 Elton Avenue Watertown, MA 02172
ANSYS	FE	Heat transfer, as well as solid mechanics and coupled thermal/stress problems	ANSYS, Inc. Johnson Road, P.O. Box 65 Houston, PA 15342-0065
BEASY	BE	Potential simulations (including thermal) and linear-elastic stress analysis	Computational Mechanics, Inc. 25 Bridge Street Billerica, MA 01821
MARC	FE	Heat transfer, as well as solid mechanics and coupled thermal/stress problems	MARC Analysis Corp. 260 Sheridan Avenue, Suite 309 Palo Alto, CA 94306
MSC/NASTRAN	FE	Heat transfer, as well as solid mechanics and coupled thermal/stress problems	MacNeal-Schwendler Corp. 815 Colorado Boulevard Los Angeles, CA 90041
NISA II	FE	Heat transfer, as well as solid mechanics and coupled thermal/stress problems	Engineering Mechanics Research P.O. Box 696, 1607 East Big Beaver Troy, MI 48099
TACO3D	FE	Heat transfer	Energy Science & Tech. Software Ctr. P.O. Box 1020 Oak Ridge, TN 37831-1020
TAU	FE	Heat transfer	AEA Technology Risley Warrington Cheshire, England WA3 6AT
TMG	FD	Heat transfer	MAYA Heat Transfer Tech. Ltd. 43 Thornhill Montreal, Quebec Canada H3Y 2E3

the types of calculations expected, various benchmarks for engineering computation have been commonly used for decades, including:

- LINPACK benchmarks, which measure a mix of integer and floating-point instructions that are commonly associated with manipulation of matrices such as those resulting from many engineering and scientific applications. LINPACK benchmarks are generally reported in terms of the number of floating-point operations performed per second (generally abbreviated as FLOPS, or MFLOPS and GFLOPS, as typical modern computers execute millions or billions of these operations per second).
- Whetstone benchmarks, which (supposedly) measure an instruction mix representative of engineering and scientific computation, but with a considerably simpler implementation than that available for the complicated LINPACK estimate. (Whetstone benchmarks are reported in terms of the number of Whetstones performed per second).
- Savage benchmarks, which measure both transcendental performance and round-off error in this component of floating-point computation. The Savage benchmark is especially simple, and because of its simplicity, its performance estimates are heavily dependent upon compiler technology quality.
- Pure-integer benchmarks, such as the Sieve of Eratosthenes, which require no floating-point computation at all, measuring only the speed of integer arithmetic. Integer benchmarks are generally the most commonly quoted measures of computer performance, but these benchmarks are often only of marginal utility in “real-world” mechanical engineering applications, as the

TABLE 15.3.6 Sample Optimization Utilities

Program	Type	Applications	Contact Address
ASTROS	FE OPT	Preliminary structural design or structural modification	USAF WL/FIBRA Wright Laboratory Wright Patterson AFB, OH 45433
CSAR/OPTIM2	FE OPT	Automated optimal structural design for NASTRAN	Computerized Structural Analysis Corp. 28035 Dorothy Drive Agoura Hills, CA 91301
DOT	OPT	General-purpose numerical optimizer suitable for use with other programs	Vanderplaats, Miura & Associates 1767 S. 8th Street , Suite M200 Colorado Springs, CO 80906
GENESIS	FE OPT	General-purpose structural optimization	Vanderplaats, Miura & Associates 1767 S. 8th Street , Suite M200 Colorado Springs, CO 80906
OPTDES	OPT	Design and optimization toolbox	MGA Software 200 Baker Avenue Concord, MA 01742
OPTISTRUCT	FE OPT	Optimum structural design with finite elements	Altair Engineering 1757 Maplelawn Drive Troy, MI 48084
STARSTRUC	FE OPT	Structural optimization application	Zentech Incorporated 8582 Katy Freeway, Suite 205 Houston, TX 77024

instruction mix of engineering calculations usually involves a considerable proportion of floating-point computation.

While these and other benchmarks are commonly quoted and compared, there are many practical problems with their use. Many benchmarks may exhibit wildly different values depending upon minor modifications of the source code, and hence may lead to inaccurate results. For instance, in the C language implementation of many matrix benchmarks, various ways of declaring integer subscripts to be used as matrix indices can result in substantial changes in the performance measured. Since the C language specification includes a "register" declaration (which advises the compiler that the integer variable will be used often, and is hence best placed in a fast on-chip storage location such as a register, instead of being stored in slower main random-access-memory), the use of "register" declarations for some benchmarks will result in substantial speed-ups on many CPUs that contain many general-purpose registers. In some cases, published comparisons are based on a "lowest-common-denominator" approach, so that if one CPU doesn't support general-purpose register allocation (for example, many earlier Intel CPU designs), this feature is disallowed on all CPUs compared, which can give rise to misleading performance results.

In order to remove some of these cross-platform performance problems, better benchmarking schemes have evolved. One improved scheme involves standardized integer and floating-point calculations measured in "SpecMarks", after the computer-industry standards organization (i.e., the *Standard Performance Evaluation Corporation*) that proposes them. These benchmarks are termed the SpecInt and SpecFp, and measure integer and floating-point performance, respectively. The standards are updated every few years, as CPU manufacturers often implement design features specifically oriented to improve these important benchmarks (it should be noted that the problem of hardware manufacturers tuning designs to improve standard benchmark performance is a widespread one and is not only encountered among CPU vendors). The particular benchmark is generally suffixed with the year of its standardization (e.g., SpecInt95 for the 1995 version of the integer benchmark suite). SpecMark benchmarks are widely quoted figures available for various processors, and the SpecMark family has been evolving to model more accurately the increasingly complicated schemes available with modern CPU architectures to speed up overall computational performance. Unfortunately, many published SpecMark benchmarks are given

only for integer performance, so the important issue of floating-point speed may not be available without more detailed research efforts.

For practicing mechanical engineers, the best benchmarks are those that most accurately reflect the computational characteristics of the problems encountered. One good way to measure the relative performance of various computers is to perform large-scale simulations of typical practical problems on the individual computer platforms. The best comparisons are obtained using the same software package on each computer, but if this is not possible, different packages can be run of different machines and the overall results compared (this latter case results in a higher uncertainty, as differences between the individual software implementations may cloud issues of hardware performance). The wider the range of problems run, the better estimate of relative performance that will be obtained. These larger-scale benchmarks can be supplemented with simpler benchmarks such as those discussed above to obtain an estimate of relative computer performance characteristics.

Accurate Measurement of Program Performance

Benchmark estimates of computer performance are a useful means for evaluating computer quality, but there are additional considerations to estimate their reliability. In particular, many features available on modern computer conspire to complicate the analysis and interpretation of standard benchmark measures. Some of these issues are addressed below.

First, one must consider the hardware and software environment utilized on the computer compared. Different computers provide different support for various types of numeric computation, and these disparate levels of support may substantially alter benchmark results. For instance, many older computers store integers as 16-bit quantities, where more modern computers use 32- and 64-bit storage locations to represent integers. Other things being equal, it is a lot easier to manipulate 16-bit integers than longer representations, and so benchmarks that do not require any integers longer than 16 bits (e.g., the Sieve of Erasthosenes utilized to find all of the primes smaller than $65{,}536 = 2^{16}$) may give potentially inaccurate results, especially if problems requiring manipulation of integers larger than 65,536 will be required (this latter case occurs nearly always in engineering practice). In order to compare benchmarks accurately, it is important to know some background information concerning the range of numeric representations used in the published results, and how this range corresponds to expected computational practice and to the ranges readily implemented on a particular CPU.

In addition, different types of operating systems can also influence performance of benchmarks. Modern preemptive multitasking operating systems slightly compromise program execution speed by adding the overhead effort of a scheduler subsystem (these and other operating systems principles are detailed later in this chapter), and this additional operating system software effort simultaneously improves the overall quality of the operating environment of the computer, while potentially degrading its numeric performance. If a benchmark requires substantial amounts of random access memory (such as the LINPACK benchmarks, where large quantities of matrix storage must be allocated), then operating systems that implement protected paged virtual-memory subsystems may exhibit substantial slowdowns, especially if the computer's physical memory is smaller than the virtual memory demanded by the benchmark application. In these cases, the benchmark actually measures a complex combination of CPU, operating system, and disk subsystem performance, as the memory is paged from memory to disk and back by the virtual memory resource subsystem.

Finally, the computer language utilized to implement the benchmark may alter the results as well. This phenomenon may be due to obvious causes (such as the overall quality of available compilers, which may penalize newer but faster CPUs that do not have carefully optimized native compilers available relative to slower but more mature platforms), or more subtle effects (such as the original C language specification, which required all floating-point arithmetic to be done in more accurate but less efficient double-precision storage — this feature of the C language often results in slower performance on some benchmarks than that obtained by using single-precision arithmetic, which is the usual default in the FORTRAN language).

Cross-Platform Comparison of Benchmark Metrics

In addition to specific computer software and hardware issues outlined above, there are other important factors that govern comparisons of benchmarks run on different types of computers. When considering performance analyses of different computer platforms, it is important to account for such factors as:

- Presence or absence of specialized hardware, such as floating-point processors, vector processors, or other numeric accelerators
- Relative quality of optimizing compilers available for the individual platforms
- Differences among operating system overhead for the various computers to be compared
- Specialized hardware for auxiliary tasks, such as input/output accelerators for problems requiring substantial access to input/output subsystems (for example, mainframes and supercomputers often have only marginally better raw floating-point performance than many workstations and high-end PCs, but for more realistic problems involving substantial transfer of data, these more expensive computers are considerably faster than simple benchmarks generally indicate)

Many of these specific hardware issues are outlined in the next section of this chapter, but all of them should be investigated to insure that accurate benchmarks comparisons are measured. In all cases, the basic principle is to be able to normalize the benchmark results to account for variations in the equipment tested (vs. what might be purchased), as well as trade-offs in performance due to desirable features (such as preemptive multitasking resources in an operating system). A benchmark suite containing a range of applications that exercise the aspects of computer performance that are of practical importance in the particular mechanical engineering enterprise can be a useful tool for evaluating relative performance of different computers, different operating environments, and different software applications.

15.4 Computer Intelligence

Computer intelligence encompasses a wide variety of thought and practice, ranging from philosophical considerations of whether computers are capable of thought to the implementations of complex programs that can play chess as well as virtually any human being. On a purely pragmatic level, computer intelligence can be considered as including nearly every application of computation found in this chapter, as all of the speedy calculations inherent in large-scale computer models can be viewed as the leveraging of the natural (but slow) computational paradigms used by humans to solve problems for centuries. A more narrow and useful view of computer intelligence can be developed by considering the computer as an approximation to a thinking machine, in that computer intelligence can be limited to considerations of simulations of higher-level human activities such as judgment and decision making, instead of the rote issues of multiplication and addition. The narrowext viewpoint of computer intelligence (often connected with the concept of "strong artificial intelligence") is the belief that computers are actually capable of true thought, just as humans are.

In this section, the intermediate philosophy of so-called "weak artificial intelligence" is considered, by which computers are utilized as useful adjuncts for modeling high-level human activities such as decision making. This practical approach to machine intelligence relegates purely computational issues (such as the computational mechanics applications of the last section) to "nonintelligent" status, yet sidesteps the philosophical quagmire of whether the computer is actually thinking. In addition, there are many other topics in the field of artificial intelligence, such as autonomous robots, machine vision, and intelligent control systems, that are addressed elsewhere in this handbook. Finally, the field of artificial intelligence is one of the richest and most diverse areas of research in computer science, with applications that span the full range of human endeavor; the material presented in this section is necessarily only a brief introduction to a few specific areas of interest to practicing mechanical engineers. Broader and deeper treatments of this important field can be found in references such as Shapiro (1992) and Kurzweil (1990).

Artificial Intelligence

Considerations of artificial computer intelligence often begin with the famous "Turing test," in which a computer is judged to be capable of thought if and only if expert human judges (who are carefully insulated from being able to sense whether they are dealing with another human or with a computer) pose a series of questions and receive corresponding answers: if they cannot determine whether the questions were answered by a human or a computer, then the computer is judged to be capable of human thought. Although no computer has passed the Turing test (and it is implicitly assumed that the question "are you a human or a computer?" is disallowed), computers seem to be gaining ground.

A simpler motivation is to consider the task of making machines "intelligent", in that they behave in intuitive and predictable ways, much like people behave (or in some cases, how we *wish* people would behave). In this broader sense, artificial intelligence ranges in mechanical engineering application all the way from robotics, where machines behave as highly skilled machinists and industrial technicians, to computer interfaces, where such desirable characteristics as consistency and "user friendliness" can be seen as emulating desirable human traits in software.

Classification of Artificial Intelligence Methods

There are many facets of artificial intelligence, and only a few will be outlined below. The various threads that constitute this important field include (but are not necessarily limited to):

- Expert systems, which attempt to emulate the decision-making capacities of human experts
- Neural networks, which mimic the perceived operation of human brains in the hope that computers can become self-learning

- Fuzzy logic, which utilizes generalizations of classical mathematics to develop models capable of handling imprecise or incomplete data
- Data mining, which involves seeking implicitly defined new patterns in existing large data sets
- Symbolic mathematics, which utilizes computers to perform many of the most difficult aspects of mathematical derivations, such as integration and algebraic manipulation
- Machine vision (or computer vision), where visual pattern-seeking algorithms are used to approximate human visual processing

Comparison of Artificial Intelligence and Computational Science

Artificial intelligence and computational science are two contrasting approaches to the same problem in engineering: leveraging the productivity of human engineers by efficient utilization of computer technology. How this augmenting of human efforts is realized is vastly different between the two approaches, however. In artificial intelligence applications, the typical rationale for the computer application is to eliminate wholesale human interaction from some component of the engineering analysis or design process. As a concrete example, expert systems used in design applications have the general goal of codifying the rules and inferences made by a human engineering "expert" into a computer implementation that no longer requires that the engineer make those design decisions. On the other hand, computational science (which represents the integration of high-level engineering knowledge and judgment, relevant computer science, and mathematical numerical analysis) is motivated by the desire to immerse the engineer in the problem-solving process by off-loading the difficult or tedious aspects of the engineering workflow to the computer, while letting the human engineer make the high-level decisions supported by computer tools. This process of putting the engineer into the design/analysis loop (instead of the artificial intelligence motivation of *removing* engineers from the same process) leads to the class of computer-assisted engineering applications collectively termed "engineer in the loop" solutions. Several particular examples of engineer-in-the-loop applications to computer-aided drafting and design are presented in a later section of this chapter.

Expert Systems in Mechanical Engineering Design

Expert systems are one of the oldest and most successful applications of artificial intelligence. Expert systems have been used in such diverse fields as engineering, medicine, and finance, and their range of applicability is enormous. There is an incredible variety of software tools available for doing expert systems development, and this rich library of development tools makes this field especially productive for generating artificial intelligence applications. A substantial component of artificial intelligence practice arises from expert systems, including efforts to develop these applications as well as research in competing technologies (such as neural networks) that have developed at least in part as a reaction to the limitations of expert systems approaches to computer intelligence solutions.

Overview of Expert Systems Methods

A simple schematic of an expert system is shown in Figure 15.4.1. There are three fundamental components of the system: a knowledge base representing the concrete expression of rules known to a human expert (and previously communicated to the computer), an inference engine that encapsulates the "how" of the problem-solving process that the expert system is designed to implement, and a data base representing the data relevant to the problem at hand (such as material parameters for a mechanical design expert system, or answers to questions from a sick human patient for a medical diagnostic application). Other important components (such as input/output systems or the computer itself) are not shown in this simplified figure.

The expert system uses reasoning based on the rules present in the knowledge base, along with the data at hand, to arrive at appropriate output in the form of decisions, designs, or checks of specification. Once implemented, a well-designed expert system often works very well, but the process of programming such a system is often very difficult. In addition to the obvious aspects of the problem, including

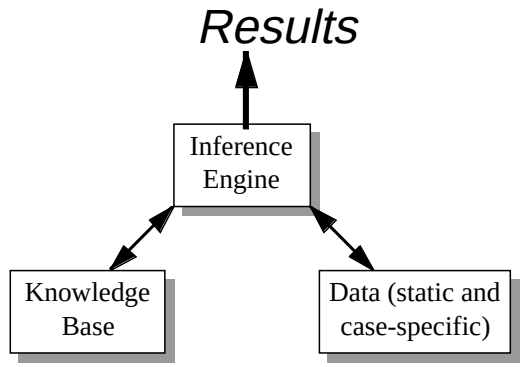


FIGURE 15.4.1 Schematic representation of an expert system.

implementing the data base, there are other nontrivial issues that must be solved, such as coaxing information out of experts who are not always sure what they are doing when they make informed judgments. While programming expert systems is a complicated task, any incomplete progress on elucidating rules for the knowledge base or implementation of an appropriate inference engine may ultimately compromise the overall utility of the resulting expert system.

Advantages and Disadvantages of Expert Systems in Engineering

The main advantage of expert systems is that they permit an incredible leveraging of human expertise. Because they are software systems, they are readily cloned (copied) and they don't forget or misrepresent what they've been taught. This amplification of the efforts of the human experts that were used to create the expert system's knowledge base is frozen within the expert system, and can be readily generalized to include new information or modified to accommodate new ways of dealing with old situations. Human beings have a certain amount of psychological inertia that often prevents them from learning "new tricks", but this problem does not affect expert systems. In addition, other human foibles, such as emotion or indigestion, can foil the opinions of human experts, but expert systems are completely consistent in their exercise of computer decision-making skills. Finally, expert systems are readily scalable to take advantage of faster computers or parallel computational resources, which permits them to solve larger or more complicated problems than human experts could possibly handle.

The biggest disadvantage of expert systems is that they represent only the knowledge base that has been programmed into them, so that such peculiarly human features as engineering judgment or creative problem-solving (which are often important components of engineering decision-making practice) are notoriously absent in this class of artificial intelligence. Another serious disadvantage of expert systems is that they require prodigious amounts of training, which makes their development and implementation costs large for appropriately complex problems. The latter trait of not being capable of autonomous learning is such an important limitation of expert systems approaches that the field of neural networks has developed in large part to remedy this important shortcoming of expert systems technology.

Sample Applications of Expert Systems in Engineering

Expert systems are among the most successful implementations of artificial intelligence, and along with this success has come a wide range of problems that have been shown to be amenable to solution using expert systems technology. Areas in which expert systems have been utilized include engineering design (usually in settings of limited scope, where the rules of practice for design can be readily articulated for programming an expert system), management (where expert systems-based decision-making tools have enjoyed considerable success), biomedicine (where expert systems for diagnostic applications have worked well in practice), and finance (where expert systems can be interfaced with computer models for financial modeling to elicit information that would be too complex for most humans to discern). In addition, research oriented toward using expert systems in conjunction with fuzzy logic has shown

considerable promise for generalizing expert systems approaches to handle noisy, poorly specified, or incomplete data.

Neural Network Simulations

Neural networks represent a promising approach to developing artificial intelligence applications in engineering, and neural nets have many characteristics that are very different from traditional expert system-based solutions. The basic principles of neural networks are derived from inquiries into the nature of biological neural networks such as the human brain. These biological systems are not well understood, but are believed to function efficiently because of the web-like system of interconnected nerve cells ("neurons"), which differs markedly from the regular topology of most conventional computers. Artificial computer-based neural networks represent an attempt to capture some of the beneficial characteristics of human thinking, such as autonomous programming and simple pattern-based learning, by mimicking what is known of the human brain. There are many good introductory references to the basic principles of neural network modeling, including books by Fausett (1994) and Wasserman (1989).

Overview of Neural Network Modeling

A computational neural network can be idealized as a three-layer system, consisting of individual processors, termed units, and a web of communications circuitry, called connections. The topology of the network is diagrammed in Figure 15.4.2. The individual processors operate only on the data that they store locally or that they send and receive via the communications web. This distributed memory architecture makes neural network computing readily parallelizable. In practice, the hidden layer that connects the input and output layers may be composed of a variety of individual component layers, but this aspect of the hidden layer's internal structure can be ignored in this elementary overview of neural networks.

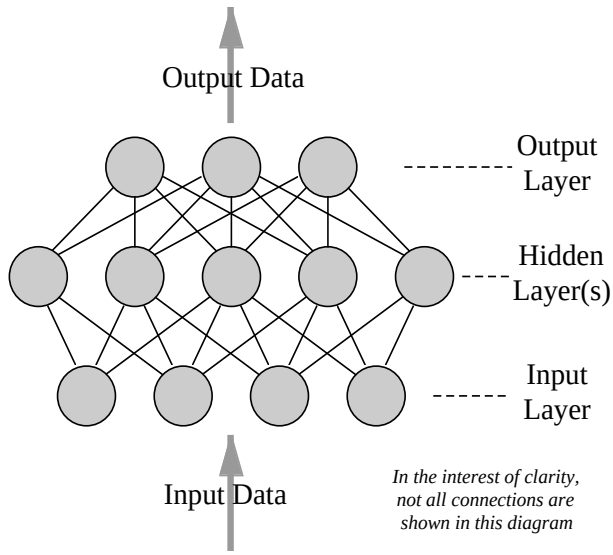


FIGURE 15.4.2 Schematic representation of a neural network.

While neural networks are theoretically capable of performing standard deterministic computation (such as the numeric algorithms used in computational mechanics), in practice neural schemes cannot presently compete with traditional computer architectures for these relatively straightforward tasks. It is in artificial intelligence applications that neural networks are used primarily, because they have an important advantage over virtually every other computational model: neural networks *learn* in a very natural and relatively autonomous fashion. In many regards, neural networks are self-programming, and

this incredible characteristic makes them very useful in practical computer intelligence applications. In simple terms, they learn by example and develop the capacity to generalize what they have learned to new situations.

The mechanism by which the most common forms of neural networks learn to solve a problem is relatively straightforward and similar to human learning: a battery of training examples are fed into the inputs of the neural network, while corresponding outputs are compared to known "correct" values. An alternative mathematical interpretation of this learning is that a weighted sum of the inputs is computed and filtered in a nonlinear manner through the hidden layers to the output layer, and this output response is compared to known output patterns in order to compute an error measure. This learning process is repeated many times utilizing some form of error minimization process to perform a nonlinear regression on the problem data, until the response of the network is sufficiently close to the desired response on all of the sample learning cases. At this point, the system has been trained to solve some set of problems, in that it can generate outputs for any set of input data. The technique described above is termed "backpropagation of error" and is commonly used to develop and train neural network models.

In practice, neural networks are not quite this simple to program, as there are many practical issues that must be contended with, and solving these difficult problems is an area of active research in the neural network intelligence field. The whole issue of how the weights are assigned, propagated through the model, and how they evolve to yield an accurate regression of output response upon input data has been glossed over, as has the more vexing problem of how best to train the network to learn autonomously instead of simply memorizing the peculiarities of the training ensemble; this last undesirable trait is termed overfitting and is an important stumbling block in autonomous neural network development. Regardless of these practical considerations, the fact that neural networks are easy to construct and straightforward to program has made them one of the fastest-growing and most important areas of application of computer intelligence.

Advantages and Disadvantages of Neural Networks in Engineering

The most obvious advantage of neural networks is that they are very simple to implement and to program. Since they are readily parallelizable and extremely simple to train, they can be easily applied to a wide variety of useful applications requiring pattern matching and machine recognition skills. Neural networks are also inherently scalable, in that their distributed architecture makes them easy to adapt to large problems.

One of the most important disadvantages of neural-network systems is that they are presently used only for certain classes of artificial intelligence problems: this limitation arises from their simple architecture and the difficulty of a completely rigorous understanding of how neural networks work. Other important disadvantages include difficult considerations of topology, number and type of connections, training methods, and avoidance of pathological behaviors such as overfitting. As neural networks become more commonplace and well understood, some of these disadvantages should become less important.

Sample Applications of Neural Networks in Engineering

In addition to the obvious examples of pattern matching in production and quality control, neural networks have found wide acceptance in a variety of important engineering roles. Other pattern-matching applications commonly occur in imaging problems, and neural networks have performed well in these cases. In addition, neural networks are natural candidates for use in computer-aided decision-making applications, including management and business decision-making cases. In these settings, other computer programs (such as computer analysis programs) may be utilized to generate case studies to be used to train the neural network, and the resulting artificial intelligence application can readily determine reasonable results from generalization of the training examples. When decisions are to be made based on running very expensive computer programs (such as nonlinear transient engineering analysis programs discussed in the computational mechanics section of this chapter), it is often more efficient to run a suite

of training problems, and then let a large neural network handle the prediction of results from other values of input data without rerunning the original analysis application.

Fuzzy Logic and Other Statistical Methods

Fuzzy logic represents an extension of the principles of classical mathematics and statistics to handle "real-world" systems, which are often imprecisely defined and poorly delineated. In simple terms, fuzzy logic achieves this generalization of traditional logic by replacing Boolean values of "either true or false" with a continuous spectrum of "partially true and partially false." A common analogy is to think of traditional statistical principles as based on black and white issues (each premise is either true or false), while fuzzy logic allows premises to take on different shades of gray (ranging continuously from black to white) and permits more realistic responses to be captured within its conceptual framework. For example, such imprecise and yet common qualifiers as "slightly" and "extremely" are readily handled within the framework of fuzzy methods, yet cause considerable difficulties for many traditional alternative modeling schemes.

Engineering systems based on fuzzy logic represent an intermediate state between traditional mathematical representations, such as standard deterministic design and analysis techniques, and purely rule-based representations, such as expert systems. In principle, fuzzy logic methods are superficially similar to expert systems in that they are both dependent upon artificial intelligence and both remove the engineer from many aspects of the design process. In practice, however, fuzzy systems are more closely related to neural networks than expert systems, as fuzzy models permit considerable automation of the process of converting descriptions of system behavior into a corresponding mathematical model. There are many good references on fuzzy logic fundamentals (e.g., Klir and Folger, 1988).

Overview of Fuzzy Set Theory and Fuzzy Logic Principles

In traditional set theory, objects are categorized as either members of a set or of its complement (i.e., everything in the universe except the set in question). In real-life applications, it is much more difficult to categorize whether anything is a member of a set. In most practical applications, engineers are interested in determining quickly whether objects are members of a set, and such determinations are naturally "fuzzy," both in the fact that they may be imprecise in their underlying definitions, and in the recognition that the test for set membership may be carried out only on small samples taken from a larger product population. In order to permit modeling mathematical constructs based on such incomplete, poorly defined, or noisy data, fuzzy set theory and fuzzy logic were developed (see Section 19.15).

In a similar manner, traditional mathematical logic represents premises as either true or false. Fuzzy logic recasts these extremes by considering the degree of truth of individual premises as a spectrum of numbers ranging from zero to one. The theory of fuzzy logic is based on extending more complex principles of traditional logic, such as logical inference and Boolean operations, to a wider realm that preserves the structure of the classical approach, yet permits an astonishing amount of realistic behavior by permitting considerable flexibility to model poorly defined systems. There are many connections between fuzzy set theory and fuzzy logic, including the obvious one of logical test of membership in fuzzy sets. Many principles of one field (such as statistical clustering applications from fuzzy set theory) are readily utilized in the other.

Advantages and Disadvantages of Fuzzy Principles in Engineering

One of the most important advantages of fuzzy applications in mechanical engineering is their flexibility. Many industrial models represent engineering systems that are either not perfectly characterized or are continuously evolving to better suit changing needs. In these settings, fuzzy systems are often excellent candidates for control and decision making, because the resulting fuzzy model is simpler than traditional approaches, owing to its intrinsic ability to adapt to poorly defined, incomplete, or changing data. Fuzzy inference systems, which are similar to expert systems in that they are utilized for rule-based decision making, are similarly flexible to specify and to implement, especially for engineers experienced in applications of fuzzy principles to engineering control systems.

Another important advantage of fuzzy models is that they are relatively simple to develop and implement. This advantage is primarily due to the fact that they can more accurately model realistic engineering systems which are difficult to specify either completely or accurately. In general, simpler models are less expensive to develop, and the simplicity of fuzzy models for engineering applications is often directly correlated to improved economies in implementation and maintenance.

Probably the most serious disadvantage of fuzzy methods in practice is that they are a relatively new approach to artificial intelligence applications. Because of their novelty, there is not much professional experience with these methods, and a critical mass of software tools for specifying and developing fuzzy models has only recently become available. Until these methods are commonly utilized and understood, it is difficult to categorize their practical advantages and disadvantages.

Sample Applications of Fuzzy Methods in Mechanical Engineering

One of the first successful commercial applications of fuzzy logic in an engineering system was developed for autofocusing features in video camcorders. Early autofocus systems for consumer-grade video recorders failed to work in a wide variety of settings, as it is difficult to design a control system for such a complex and dynamic system. In addition to the technical optical issues of rangefinding and focus control, there are other pertinent (and naturally fuzzy) questions that must be solved to enable such automatic devices to work robustly in a motion-recording setting:

- What do you focus on when there are many objects in the field of view?
- What if some or all of these objects are moving through the field?
- What should be done when some objects are approaching and others receding from the viewer?

Each of these issues is readily modeled using fuzzy logical principles, and recasting such device control problems into a fuzzy statistical setting often transforms the underlying engineering system into a much simpler formulation. Since simpler models are generally easier to implement and to modify, the cost of developing and extending fuzzy-based engineering models is often substantially lower than the cost of competing traditional control systems.

The range of problems thought to be amenable to fuzzy implementations is enormous. Examples range from automotive applications, where sensors generally reflect continuous response instead of Boolean extremes, to tasks requiring categorization of objects as representatives of sets, such as pattern-matching applications or quality-control applications. Many household appliances now incorporate fuzzy methods for control, and the growth of this component of artificial-intelligence engineering applications is expected to be rapid. In general, the more imprecise the data available from an engineering system, the more likely a fuzzy approach will work well in comparison to more traditional artificial intelligence approaches.

15.5 Computer-Aided Design (CAD)

Joseph Mello

Introduction

The computer is a powerful tool in the design of mechanical components and entire mechanical engineering systems. There are currently many diverse computer applications that were created to augment the mechanical design process. The development of a mechanical component or system can generally be divided into three tasks: design, analysis, and manufacturing.

Historically, the design task was restricted to describing the part geometrically. This task included conceptual sketches, layouts, and detailed drawings, all of which were created on a drawing board. The designer or draftsman generally created drawings using orthographic or axonometric projections of the three-dimensional object being designed or developed. These drawings served to illustrate the mechanical part, and in the case of a detailed dimensional drawing, served as a data base for transferring the design information to the manufacturer or to the shop floor.

The drawing performed in the design process has been automated and is now performed using the computer and dedicated software. This automation process is termed computer-aided design, or CAD. CAD generally refers to the geometric description of the part using the computer, but as the computer and its corresponding software have evolved, other tools have been integrated within contemporary CAD systems which are historically categorized as analysis or manufacturing tasks. This form of design, analysis, and manufacturing integration will be discussed later in this section.

CAD systems presently fall into two general categories: those based on entities and those based on solids or features. This classification refers to the basic tools used to describe a mechanical part. The entity-based system can be thought of as a three-dimensional electronic drawing board. The designer in this case uses lines, arcs, and curves to describe the part in a manner similar to the way objects were defined on the drawing board. Solid modeler CAD systems take a different approach: the designer in this latter case must think and work more like a manufacturer. The part is created or described using more complicated solids or volumes, for example. Both of these systems are discussed in the following sections.

Entity-Based CAD Systems

The entity-based CAD system, as mentioned earlier, can be thought of as a powerful electronic implementation of the drawing board. These systems have been designed for nearly all computer platforms and operating systems. Currently, their low cost and portability allow them to run on an Intel-based personal computer or on an Apple Macintosh, and these advantages result in their availability at a majority of small or medium-sized design engineering or manufacturing firms. The workhouse and a typical example system is the AutoCAD application running on an Intel-based microcomputer using a Pentium microprocessor. Some of these CAD systems have optional digitizer pads that are used in the design process, while most use standard keyboard and mouse hardware for input.

Overview of Modeling and Drafting

There are a variety of ways to describe a mechanical component using an entity-based CAD program. For a simple well-defined part, the designer or draftsman may elect to make a two-dimensional (2D) engineering drawing directly. This work is analogous to using a drawing board, in that the designer or draftsman uses basic commands to draft and dimension the part. For a more complex component, the designer may choose to create a three-dimensional (3D) wireframe model of the component, which can be meshed with surfaces. The motivation to do this may be for visualization purposes or for the export of the geometry to an analysis package or a computer-aided manufacturing application. The following material briefly outlines the use of an entity-based CAD system.

Two-Dimensional Modeling

User Interface. The user interface varies among different systems, but some common user-interface elements include pull-down menus, iconic menus, and dialog boxes for input of data. Menu systems are manipulated using the mouse and keyboard. Some systems may use a digitizer tablet with icons to select different drawing operations. The software may also provide the experienced user with a command line input prompt as an alternative option to pulling down menus or dialog boxes.

Coordinate Systems. The key element of any entity-based CAD application is the Cartesian coordinate system. All such applications have a default or global system and also allow the user to specify any number of user-defined coordinate systems relative to the global system. Systems usually also offer the user polar coordinate systems. All geometric entity data are referred to a particular coordinate system.

Basic Drawing Entities. The designer or draftsman at the drawing board uses a drafting machine to draw straight lines, a compass or circle template to draw circles and arcs, and a French curve to draft smooth curves. The entity-based CAD system has analogous basic entities which are the line, circle or arc, and spline.

Lines are drawn by a variety of methods. The line is defined by its end points in the system coordinates. Basic line construction is often accomplished by pulling down a line construction menu, selecting the option to define the line's global coordinate end points, and then typing in the Cartesian coordinate pair. Another method for line construction is to use the control points of existing geometry: control points are a line's or arc's end or midpoints. All systems allow the user to "snap to" or select control points. Given two lines defined by coordinate end points, one can close in a box with two more lines by selecting the end points of the existing entities.

Circles and arcs are constructed in a manner similar to line definition. The most common circle construction is to specify the center point in terms of an active coordinate system and either the radius or diameter. Some other circle construction methods are to define three points on the circumference or two diameter points. There are generally various ways to construct entities in addition to these methods. Arcs are constructed in the same fashion as circles. Some examples of arc construction methods include: start, contour, and end point definition, start and end point radius, or start and end point and angle of arc. This list is not inclusive and of course is system dependent. One important class of arc is the fillet. Most systems offer the user the ability to round a corner by selecting two lines with a common end point and then simply defining the fillet radius.

The spline is the mathematical equivalent of using the French curve to fit a smooth curve between points. The CAD user defines the curve's points and the system fits a cubic spline to these points. There are generally options for specifying a spline's end point tangents or other end conditions.

Design Aids. The utility of a CAD system can be realized when the user appeals to the myriad functions designed to aid in drawing on the computer. Some typical tools are discussed below.

- Layering or layer control gives the designer significant advantages when compared to the drawing board. Layers can be thought of as electronic consecutive sheets of tracing paper. Entities can be placed on different named or numbered layers. Thus the designer can put a whole system design in one file with each component on its own layer. These layers can be turned on or off, and therefore are visible and active or are visible for reference only.
- Entity color, thickness, and line type can be set and later changed or modified. Thickness refers to the displayed and plotted line, arc, or spline width. Type is reserved for hidden, centerline, or phantom line types which are generally a variety of dashed lines.
- Selection control determines how entities are picked and created. Systems have orthogonal grids analogous to graph paper which can be used to sketch parts. Entity end points can be made to snap to grid points. Grid scales can be modified to suit. Angles can be set to force all geometry to be drawn at an angle to an active coordinate system. When building geometry by selecting end

points of existing geometry one can filter selections. An example would be to set the appropriate filters so that only entities red in color may be picked or selected.

- View control refers to windowing, viewing, zooming, and panning. Generally all parts are created at full scale in a CAD system. One then uses view control to look at the whole part or to zoom in on a small portion of it. These view controls are essential, as even a large computer monitor is deficient in view area compared to an "E" sheet on the drawing board.
- Design information menus or calculations are built into CAD systems. These features can be as simple as showing what layer an entity is on or measuring its length or location in Cartesian space. Even the most simple systems can determine the area between groups of entities. Some systems similarly calculate centroids and inertias.

Drafting

Drafting is defined herein as operations such as dimensioning, tolerancing, hatching cross-section views, and writing notes and title blocks in the creation of engineering drawings. These tasks are usually performed after the model geometry has been completed.

Basic Dimensioning. Using a CAD system does not change where dimensions are placed on a drawing. The drafter still applies the traditional rules, but the process of creating dimensions changes completely. Since the component's geometry has been drawn to scale very accurately, a drafter needs to only tell the CAD system where to start, stop, and place the dimension. The system automatically determines the dimension value, draws extension/dimension lines and arrows, and places the dimension text.

There are generally five different types of dimensioning commands. Linear dimensions can be aligned with horizontal or vertical geometry. Angular dimensions measure the distance between two nonparallel lines. Radius and diameter commands dimension arcs and circles. Ordinate dimensioning commands call out Cartesian coordinates of entity features such as spline points. These features, combined with the ability to attach notes, allow the drafter to define the part formally.

Dimensioning Variables and Standards. Dimensioning variables are system standards that control dimension format. Examples are arrowhead size, text location, and text size. These are generally controlled by a set of dimensioning subcommands. Most systems' default settings conform to ANSI Y14.5 standards. CAD systems also are designed to automate geometric dimensioning and tolerancing. Menus and commands create and label reference data. There is usually an extensive ANSI Y14.5 symbol library so that a drafter may add features such as control references, symbols, and finish marks.

Drafting Aids. CAD systems are usually equipped with tools or commands to create drafting entities. These entities include section and center lines. Cross-hatching is generally automated to some degree and a large library of ANSI hatch styles are usually available. If the part drawings are created from a 3D wire frame model as some systems allow, hidden line detection or removal may be automated. This is a very powerful perceptual cue for aiding in the visualization of 3D entities.

Libraries and Data Bases

CAD systems are designed to build up data bases and libraries; for example, special symbol libraries can be created. Any component drawn or modeled can be grouped or blocked together, and the resulting part can be saved and imported into future components or drawings as a single entity. It may then be ungrouped or exploded into distinct entities if the designer wishes to modify it to create a different version of the component.

Three-Dimensional Wireframe and Mesh Modeling

Entity-based CAD systems generally allow the designer to model in three dimensions. Historically, three-dimensional (3D) drafting augmented standard engineering drawings with an isometric or oblique view of the component being described. CAD systems have in a sense reversed this process, as complex parts are now modeled in 3D and then rendered onto engineering drawing views from the 3D data base.

3D modeling is a simple extension 2D drawing and modeling using CAD systems. Any of the basic entities of lines, arcs, circles, and splines can be drawn using X, Y, and Z Cartesian coordinates for control points. The basic entities can thus be used to create a "wireframe model". The wireframe model essentially has every edge of the component or part defined by basic entities. This extension to 3D is simple in theory but is somewhat complex in practice because the interface and display capabilities are two dimensional.

Key tools in productive 3D CAD efforts are coordinate system and view angle manipulation. Some systems also allow definition of construction planes which defines a 2D plane in space where all current entities are placed. In addition, some systems allow computer hardware-assisted rotation of viewing which makes visualization of the part in 3D extremely fast.

Surfaces. Points, lines, arcs, and splines are used to create wireframes, but surfaces are used to represent an object accurately. Surfaces can be shaded to render objects in order to make them look realistic or they can be exported to analysis or manufacturing systems. Surfaces on a entity-based system can be idealized as complex entities, as they can be created from wireframe entities such as lines, arcs, and splines or specified directly by input of the particular defining surface parameters.

Meshes are arrays of faces. Through the use of meshes, engineers can define multifaceted surfaces as single entities. Some common surface types are

- Ruled surface: a surface between two defined curves such as lines, arcs, or splines
- Tabulated surface: a surface created by projecting a drafting curve some distance in space
- Surface of revolution: a surface created by revolving a generator curve about an axis
- Edge-defined coons surface patch: surface created by interpolating with a bicubic function from four adjoining edges defined by lines, arcs, or splines

Solid or Feature-Based CAD Systems

CAD systems that are based on solid modeling or feature modeling are significantly different in appearance and use from entity-based systems. There is generally no means to draft a simple part in 2D using a feature-based system, so descriptions of feature-based CAD begins with solid modeling.

Overview of Solid Modeling

The design process starts with creating a solid model using solids, primitive shapes, or features. Examples of these include solid extrusions, revolved solids, or solids created from general surfaces. There are also solid removal features such as holes, cuts, and chamfers. One way to build a solid model is to extrude a block and then to cut, make holes, and round the edges in a manner analogous to actual manufacturing, where a block of material may be sawed, drilled, and milled into the desired shape. Some details of creating a solid model follow in a simple case study. It should be noted that some solid modelers are parametric. This allows the part to be reshaped by changing key dimensions. The parametric CAD system automatically reshapes the model or part when the dimensions or parameters are modified.

After creation of the solid model the designer can combine solid models or parts to create assemblies. The assembly process with a solid modeler also has analogies to the actual manufacturing assembly process. For example, part surfaces can be mated and aligned to create an assembly.

The solid modeler usually has a separate drafting or drawing module that is used to create engineering drawings. Views are generated nearly automatically from the solid model. Reference dimensions can be added along with symbols and other drafting details. In the most powerful systems the drawing, assemblies, and parts are associative; the designer can make changes in parts at the drawing or assembly level, and these changes are automatically reflected in the model. The part, assembly, and drawing features are separate entities but are associated to insure a consistent data base.

An important advantage of solid modelers is that one essentially "builds" the parts and assemblies. Shading and viewing from all sides are generally accomplished using simple keyboard commands and standard mouse movements. Complete design information is available from a solid model, including

such parameters as volume, weight, centroids, areas, and inertias. Clearance and interference-checking algorithms prove or test assembly function prior to manufacturing. If detailed computational analyses are to be done or the parts to be made use computer-controlled machinery, complete part or neutral (i.e., intermediate transfer) files can be used to transfer information to analysis or manufacturing applications. The current trend is to integrate CAE and CAM functions with the solid modeler, thus eliminating the need for neutral files.

Hardware and Operating System Requirements

The CAD solid modeler generally requires significant hardware and software resources. The typical large CAD system runs optimally on UNIX-based workstations with significant graphics hardware. The recent increased power of personal computers and the availability of Windows NT operating system with its integrated graphic libraries have resulted in the availability of solid modeling on lower-cost RISC workstations and Pentium-based personal computers. This is currently giving small- and medium-sized companies design tools once reserved for aerospace and automaking industries. The software is expensive, but many companies realize payoff if accounting is made for shorter, more efficient design cycles and fewer errors that result from the use of such a system.

Part Solid Modeling Details

A case study of the design of a connecting rod will be used in the following section as a concrete demonstration of how solid modeling systems are utilized in practice. The system used in this case study was Parametric Technology Corporation's Pro/ENGINEER, which is currently considered to be one of the best CAD systems. The hardware used was a Silicon Graphic Indy workstation equipped with a MIPS R4400 processor and 64 megabytes of random access memory.

Datums, Sections, and Base Features. Pro/ENGINEER (Pro/E) uses what is termed "feature-based modeling" to construct a part. This refers to the fact that the part is designed by creating all the features that define the part. One generally begins the model of a part by creating default orthogonal datum planes. Next, a base feature or solid is typically constructed: this is a solid that is approximately the size and shape of the component to be created. Other features are then added to shape the base solid into the part that is desired.

The connecting rod's base feature in this case will be an extrusion that is defined by a section normal to the axis defined by the connecting rod holes. The section for the extrusion is sketched on a default data plane. Mouse and menu picks allow the engineer to sketch the shape using the "sketcher" mode. This part of the process is somewhat similar to entity-based CAD drafting, except that the sketch initially need not be exact. The section is then dimensioned: once again this task is similar to dimensioning on an entity-based system. The part is then "regenerated" or solved mathematically, assuming the given dimensions and alignment to datums are adequate to define the part. The dimensions or "parameters" are then modified to shape and size the part: see Figure 15.5.1 for the connecting rod's dimensioned section. The resulting extrusion is shown in Figure 15.5.2 as a wireframe plot. The designer has the option of viewing wireframes or shaded solids during the model creation process.

Rounds, Holes, and Cuts. The connecting rod can be made lighter in weight with a cut, in this case, that involves an extrusion that removed material leaving the rod cross section in the form of an "H". The dimensioned cut section and resulting solid are shown in Figure 15.5.3. Given the cut base feature, rounds can be added in this case to smooth out stress concentrations. These "edge rounds" are constructed by selecting an edge and defining a radius. The solid modeler adds material in the form of a round. Figure 15.5.4 shows the cut and rounded extrusion produced using this approach.

Hole features are now used to create the bores for the piston wrist pin and the crankshaft. These hole features are created by defining the hole location, setting the parameters for the diameter, and describing a bore depth or, in this case, by instructing the modeler to drill "through all."

The connecting rod is finished with a collection of cuts, rounds, and chamfers. A shaded image of the finished connecting rod is shown in Figure 15.5.5. Two different versions of the rod are shown. The

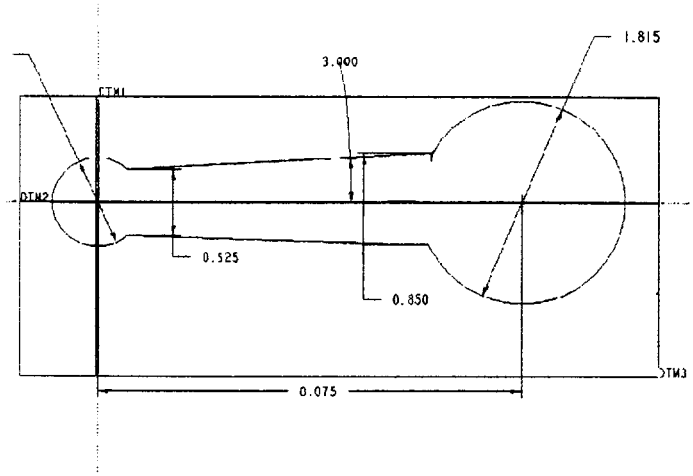


FIGURE 15.5.1 Connecting rod dimensioned section.

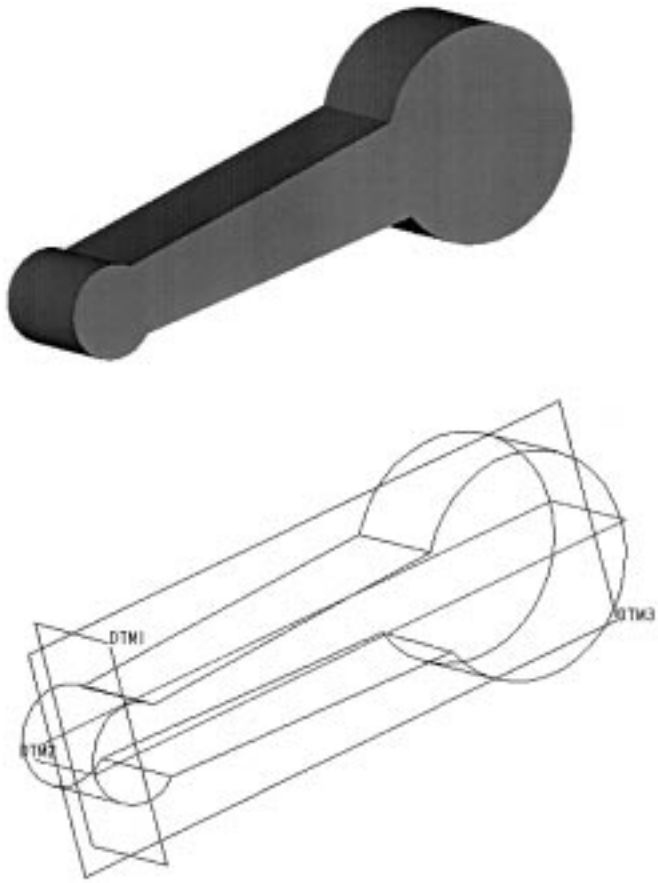


FIGURE 15.5.2 Wireframe plot of connecting rod.

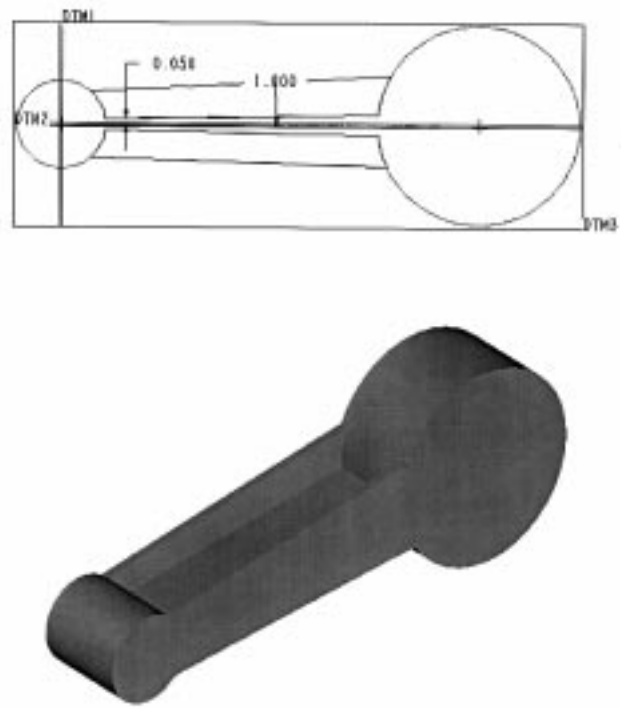


FIGURE 15.5.3 Dimensioned cut rod, solid view.

FIGURE 15.5.4 Edge rounds added to extrusion.

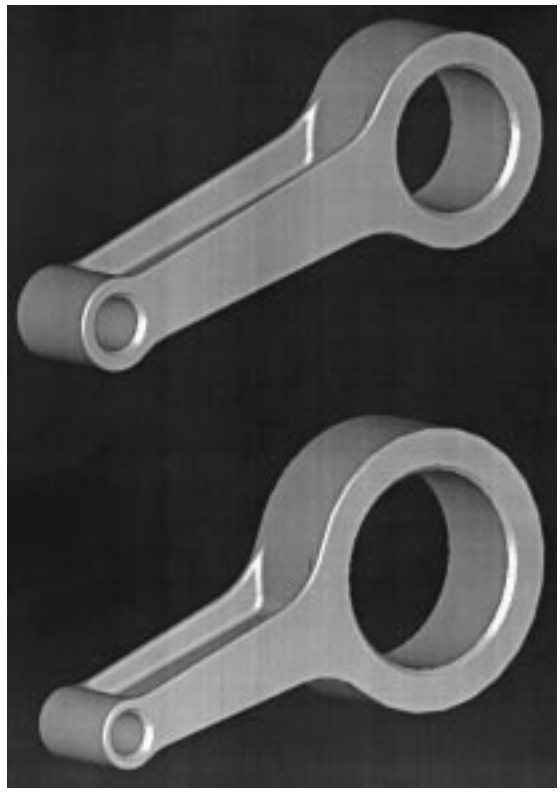


FIGURE 15.5.5 Finished connection rod.

second rendering is a shorter rod with larger main bore and a smaller wrist pin bore. This part was created simply by selecting and changing a few of the defining dimensions and telling the modeler to regenerate the part. This example illustrates the parametric nature of the modeler, which permits design changes to be easily visualized and accommodated.

Other Features. The features used to create this simple example are obviously not an inclusive enumeration of the solid features available to the designer. A brief list of other key features and corresponding descriptions follows to aid in summarizing the solid modeler's capabilities.

- Surface: extrude, revolve, sweep, or blend sections to create a surface
- Shaft: a revolved addition to the part using a specified section
- Slot: create a slot from a sketching plane
- Protrusion: boss created from the sketching plane by revolving, extruding, sweeping, or blending
- Neck: create a groove around a revolved part using a specified axial section
- Flange: create a ring around the surface of a revolved part using a specified axial section
- Rib: create a rib from a specified sketching plane
- Shell: remove the inside of a solid to make it a thin shell
- Pipe: create a 3D tube/pipe/wire
- Tweak: create part draft, dome, offset, or other surface deformation features

Part Modeling Tools. The solid modeler contains a large assortment of tools that can be used to manipulate the features or solids that comprise the component being designed. A description of a few of these functions follows.

Solids modelers generally provide a layer control similar to that discussed for entity-based CAD systems. This function is particularly useful for controlling the large number of datums and other features created by a solid modeler. Features may be redefined, reordered, and suppressed or blanked.

There are also capabilities built in to allow features to be patterned, copied, or mirrored. Patterns can be used to create splines in a shaft or hole arrays in a flange, for example. Solid protrusions can also be patterned to create spokes on a vehicle wheel.

Cross sections can be created by building or selecting a datum plane that passes through the part. The system automatically creates the cross section at this plane. This information can be later used in the drawing mode to create section views at key locations.

Pro/ENGINEER has a very powerful utility termed "relations." A feature's geometric dimensions can be made functions of other features or geometry. This is similar to having a spread sheet controlling the geometry. Relations can be expressed in terms of algebraic or trigonometric functions as needed. Logic can be built into these relations, as well as utilizing the solution of systems of simultaneous equations.

Assemblies

Parts or solid models that have been created can then be put together to form "assemblies." This process is analogous to assembling actual components in a manufacturing environment. The process starts by placing a base part or component in the assembly. This base part could be the connecting rod example discussed previously. After placement of the base part one can, for example, add or assemble components to the base part such as a wrist pin and piston in this case study. Components are added by defining appropriate constraints. Some typical assembly constraints are

- Mate: Planar surfaces are constrained to be coplanar and facing each other.
- Align: Planar surfaces are constrained tube coplanar and in the same direction.
- Insert: The axis of a feature male part is constrained to be coaxial with a female feature on a parts axis.
- Coordinate system: Two data coordinate systems on different components can be aligned.

The wrist pin is added to the connecting rod by mating the bore of the rod to the cylindrical surface of the pin and aligning the datum plane at the middle of each component. The piston is added to the assembly in the same way. Figure 15.5.6 shows a shaded view of this example subassembly. It should be noted that changes to any of the components can be made at the assembly level and the associated part is automatically updated even though they are separate entities.

There are some advanced tools provided at the assembly level. Material may be added or subtracted from one set of parts to the other set of parts in the assembly. For example, the connecting rod could be added to a block of material such as a mold base. Then the blank could be cut using the connecting rod to form a mold with which to forge or cast the rod. Parts can also be merged at the assembly level, thus becoming essentially one part.

Design Information Tools

Pro/Engineer provides utilities that calculate engineering information for both parts and assemblies. There is an option to assign material properties directly or they can be picked from a data base. Given the material data the modeler can generate the mass properties for a part: Figure 15.5.7 shows the mass property results for the connecting rod example. Basic measurement and surface analysis functions are also available, and cross-section properties can also be generated. Assembly mass properties can be calculated in the assembly mode. Powerful clearance and interference calculations can be performed for any combination of subassembly parts, surfaces, or entities — these serve to check the function of hardware before committing to actual manufacturing. In general, the added effort of building solid models is more than offset by the information and checking that are provided.

FIGURE 15.5.6 View of example subassembly.

Drawing

Documenting designs by creating engineering drawings with a solid modeling system such as Pro/E is done using a drawing module. Since the part is completely described by a solid model, most of the work normally associated with drafting or drawing is already completed. The process of creating a drawing from a part or solid model is outlined briefly below.

The first step in drawing creation is to name the drawing, select the drawing size or sheet, and retrieve the appropriate format. Formats are essentially borders and title blocks. A format mode editor exists to create standard or custom formats with any drafting geometry, notes, and call outs desired. Views of the desired model are then placed within the format. The user first places what is called a general view and orients it with view control menu commands. Now the user can place projected views off the first view. The system automatically creates the projected views from the model selected.

View types can be projections, auxiliary, or detailed portions of the geometry. View options include full views, broken views, x-sections, or scale settings to scale a particular view independently. Views can subsequently be moved, resized, and modified in a variety of ways. The most important is the drawing scale which the system calculates automatically depending on model size and sheet selected. A new scale can be defined from a modify menu and all drawing views are rescaled accordingly.

Dimensions are now placed within the drawing. Since the solid model is dimensioned in the design process, this step simplifies to that of instructing the drawing module to “show all” detailed dimensions. The system then automatically places dimensions on the views defined. The real effort of dimensioning becomes that of cleaning up the dimensions in order to have them to meet traditional drafting standards. Sometimes the dimensions used to model the part are inadequate in a detailed drafting sense, so a complete drafting utility exists to create what are essentially reference dimensions.

```
                    MASS PROPERTIES OF THE PART CRC_ROD

                         VOLUME =  2.3581262e+00 INCH^3
                   SURFACE AREA =  2.6669756e+01 INCH^2
                        DENSITY =  1.0000000e-01 POUND / INCH^3
                           MASS =  2.3581262e-01 POUND

          CENTER OF GRAVITY with respect to _CRC_ROD coordinate frame:
X   Y   Z          -2.7834183e+00 -2.4942842e-03  0.0000000e+00  INCH

      INERTIA with respect to _CRC_ROD coordinate frame:  (POUND * INCH^2)

INERTIA TENSOR:
Ixx Ixy Ixz         6.7806120e-02 -1.2601220e-03  0.0000000e+00
Iyx Iyy Iyz        -1.2601220e-03  2.4115880e+00  0.0000000e+00
Izx Izy Izz         0.0000000e+00  0.0000000e+00  2.4404084e+00

     INERTIA at CENTER OF GRAVITY with respect to _CRC_ROD coordinate frame:
                                                          (POUND * INCH^2)
INERTIA TENSOR:
Ixx Ixy Ixz         6.7804653e-02  3.7703925e-04  0.0000000e+00
Iyx Iyy Iyz         3.7703925e-04  5.8464929e-01  0.0000000e+00
Izx Izy Izz         0.0000000e+00  0.0000000e+00  6.1346817e-01

                PRINCIPAL MOMENTS OF INERTIA:  (POUND * INCH^2)
I1  I2  I3          6.7804378e-02  5.8464957e-01  6.1346817e-01

          ROTATION MATRIX from _CRC_ROD orientation to PRINCIPAL AXES:
                        1.00000        0.00073        0.00000
                       -0.00073        1.00000        0.00000
                        0.00000        0.00000        1.00000

     ROTATION ANGLES from _CRC_ROD orientation to PRINCIPAL AXES (degrees):
angles about x  y  z    0.000          0.000          0.000

               RADII OF GYRATION with respect to PRINCIPAL AXES:
R1  R2  R3          5.3622289e-01  1.5745784e+00  1.6129188e+00  INCH
```

FIGURE 15.5.7 Mass properties calculated for example cross section.

As discussed previously for entity-based CAD systems, the solid modeler drawing or drafting utility has detailed functions designed to create basic dimensions, specify geometric tolerances, and to set data as reference. Large ANSI symbol libraries are also available to aid in the detailing.

The part and the drawing exhibit "bi-directional associativity": if the engineer completes the drawing and then goes back to the part to make modifications, the drawing geometry associated with these changes is automatically updated. In a similar manner, the part definition is changed if a defining dimension is modified on the drawing.

CAD Interface Overview

All contemporary CAD systems have utilities to import and export design data. These utilities have translators which support many types of file formats. The most common interfaces and files will be discussed in the following paragraphs. Since the CAD system usually defines the part the focus will be on, export of data and Pro/Engineer will be used as an example.

Printing and Plotting. A very important interface is with printing and plotting hardware to get hard copies of shaded images, wireframe plots, and drawings for use in design reviews and for archiving. Shaded images are commonly written in PostScript page-description format for printing on a color printer with such capabilities. Plot files of objects or parts, drawings, or assemblies can be done with the standard

HPGL (Hewlett-Packard Graphics Language) or PostScript formats. Pro/ENGINEER, for example, has specific configuration files for a variety of Hewlett-Packard, Calcomp, Gerber, Tektronix (for shaded images), and Versatec plotters.

Computer-Aided Engineering (CAE). It is often desirable to export CAD geometry to other CAD systems or analysis packages. Example destinations are

- Thermal/Structural Finite Element Analysis
- Computational Fluid Dynamics Analysis
- Moldflow Analysis Packages
- Kinematics and Dynamics Programs
- Photorealistic Graphics Rendering Programs

The most common format to transfer graphic and textual data is the Initial Graphics Exchange Specification (IGES) file format. It is important to determine what type of IGES entities the receiving system desires, then adjust the IGES configuration prior to creation of the file. As an example, a variety of Finite Element Analysis (FEA) packages use an IGES trimmed surface file as opposed to wireframe entities where part edges are output. A finite element model of the case study connecting rod was formulated using an exported IGES trimmed surface file of the Pro/E solid model part file. A stress contour plot of a simple static analysis of this rod model is shown in Figure 15.5.8. The model was automeshed with hexahedral elements directly from the surface files with little user intervention. The meshing procedure took 30 to 40 min to initiate and the automesh time was about 10 hr on a Silicon Graphics Workstation. Most production-quality finite element codes either already have or will eventually possess such capabilities.

FIGURE 15.5.8 Stress contours and geometry of optimized rod.

Data exchange format (DXF) files are also very common and are used to export model and drawing data to products such as AutoCAD or to products that were designed to interface with AutoCAD. Pro/E also has its own neutral file which is a formatted text file containing more design information than an IGES file. Part geometry is formatted so other software packages can use the complete design data.

Pro/E also directly writes other software package neutral files such as PATRAN geometry files (PATRAN is a generalized finite element pre- and postprocessor).

This is only a partial list of export options available. There exists a crucial problem with export, however: associativity is lost. If the part is, for example, exported and analyzed, any subsequent part modification renders the analysis model inaccurate. This is a problem as the current trend is to use computational analysis tools early in the design process instead of using them only as a final check. Presently, there are software package alliances and cooperatives forming that are addressing this problem.

This direct integration of analysis packages has been accomplished by Structural Dynamics Research Corporation I-DEAS and Pro/Engineer Mechanica, among others. Once again the Pro/Mechanica integration will be used as an example to review the current capabilities of such systems. The Mechanica packages are thermal, motion simulation, and structural analysis packages that are completely integrated with Pro/E's solid modeler. Analysis is thus another large set of pull-down menus and dialog boxes within the CAD package. The beauty of the system is that it is a complete bi-directional link, in that design changes made in Pro/E automatically update finite element models or vice versa. This feature greatly simplifies the task of structural design optimization. Mechanica has routines which will optimize and change model geometry based on static stress and structural dynamic considerations. There are practical limitations as Mechanica is currently limited to linear elastic structural analysis. The Mechanica analysis package uses an adaptive p-element, where the order of the tetrahedral finite element is adjusted until convergence is achieved. The adaptive scheme uses up to a ninth-order polynomial in the element formulation. If the model does not converge, the mesh can be refined and the analysis tried again.

Computer-Aided Manufacturing (CAM)

Parts that are modeled in 3D are generally fabricated with CAM. This fabrication task can involve rapid prototyping, directly machining the part, or mold making to support casting, forging, or injection molding.

Rapid Prototyping

Rapid prototyping is a relatively new manufacturing technology which is complementary to CAD solid modeling. A CAD design file of an object is converted into a physical model using special computer-controlled sintering, layering, or deposition processes. SLA stereolithography machines by companies such as 3D Systems convert 3D CAD data of objects into vertical stacks of slices. A computer-controlled low-power ultraviolet laser then traces across photocurable resin solidifying the part one layer or slice at a time. Selective laser sintering and fused deposition processes operate in a similar manner. In all cases, an accurate nonstructural model is created which can be used in design reviews, mock-ups, and prototype assemblies. A physical model such as this is a very powerful tool in early product design reviews and marketing efforts.

The interface to rapid prototyping is an SLA or (*.stl) file. SLA files represent the solid model as groups of small polygons that make up the surfaces. The polygons are written to an ASCII text or binary file. Pro/E has an SLA selection in its interface menu. It is important to note that the designer partially controls the quality of the prototype through two variables, the chord height of the tessellated surface and an angle control variable. The connecting rod example is again used to demonstrate the effects of chord height setting in Figure 15.5.9, which shows the polygon surfaces for SLA files with two different chord heights.

Computer Numerical Control Machining

Numerical control machining refers to created parts using computer-controlled machines such as mills, lathes, or wire electric discharge machines. Parts or molds are made by machining away material. In the case of complex parts it is essential to transfer CAD geometric data to a machining software package and finally to the computer-controlled machine.

A typical machining center has a numerical control machine wired directly to the serial port of a microcomputer. This configuration is termed computer numerical control or CNC, as it allows the machinist to program the machine tool off-line using specialized machining programs. Most CAD

FIGURE 15.5.9 Polygonal representation for stereolithography.

systems have optional integrated manufacturing modules similar in architecture and use to their integrated analysis packages. There are also a variety of stand-alone machining software packages which accept a spectrum of CAD input in the form of neutral, IGES, and DXF files. These packages or modules can import and create geometry, as well as determine and simulate cutter paths. As with many CAD programs, these machining applications now have 3D interactive graphic environments. As the programmer generates steps for the machine tool path, 3D models show the part being cut by the tool. Line-by-line programming in a text mode is presently being replaced by graphically selecting geometry, tools, and tool paths. These programs have postprocessors to write the "G and M code" specific to the machine tool being used ("G and M" code is a standard language for programming machine tools). These ASCII text files based on numerical codes using G and M as prefixes form an instruction list for a wide variety of machine tool operations.

Further Information

There are a wide variety of excellent references available that summarize many of the principles found in this chapter. For practical discussions of computer technology, the Internet-based UseNet newsgroups form the most detailed and up-to-date sources of information on programming languages, computer architecture, and software engineering. Most UseNet groups provide a list of Frequently Asked Questions (FAQs), which often address introductory issues in considerably more detail than elementary textbooks, and consulting such UseNet FAQs is a good way to begin more detailed study of most of the computational topics outlined in this chapter. Sample UseNet newsgroups include comp.ai (for general information on artificial intelligence research and application), comp.object (for general information on object-oriented programming), and comp.sys.intel (for information regarding Intel hardware). These general UseNet topics are generally further specialized toward more detailed discussions, such as the philosophical information regarding computer intelligence that can be found in comp.ai.philosophy. Since UseNet groups are created (and sometimes disappear) at near-random intervals, it is best to use standard Internet browsing tools to determine the best newsgroups for researching a particular topic.

More traditional printed works are available for all the topics found in this chapter. The history and philosophy of software engineering are presented in engaging detail in Yourdon's *Writings of the Revolution: Selected Readings on Software Engineering* and Brooks's classic *The Mythical Man-Month*. The classic modern reference on object-oriented programming is Booch's *Object-Oriented Analysis and Design*, which also contains good descriptions of procedural programming models. An excellent overview of artificial intelligence can be found in Kurzweil's *The Age of Intelligent Machines*.

One of the most detailed treatments of finite-element modeling available is Hughes' *The Finite Element Method: Linear Static and Dynamic Finite Element Analysis*, which presents a comprehensive and detailed view of finite-element modeling as oriented toward mechanical engineering applications, and does an excellent job of presenting time-dependent solution techniques amenable to any numerical discretization approach (including finite-difference methods). There are many standard references on finite-difference methods, including the well-known book by Ames, *Numerical Methods for Partial Differential Equations*.

Terms and Associated Acronyms

Artificial Intelligence (AI): a spectrum of ideas on the general topic of computer modeling of human intelligence, ranging from "Strong AI" (defined as the premise that computers are capable of emulating human intelligence) to "Weak AI" (defined as the premise that computer can be utilized as useful adjuncts for modeling high-level human activities such as decision-making).

Benchmarks: application programs specifically designed to quantify the performance of a particular computer system, or of its component subsystems. Benchmarks provide a convenient means to compare measures of system performance obtained from disparate computer platforms.

Boundary-Element Method (BEM): a numerical approximation technique based upon integral equations and related to the finite-element method. Boundary elements possess the advantage of requiring only surface-based representations of the problem domain, in contrast to the volume-oriented data required by finite-element methods.

Central Processing Unit (CPU): the main processor (i.e., the "brain") of a computer, consisting of the hardware necessary to perform the low-level arithmetic and logical operations that form the building blocks of computer software.

Compressor/Decompressor (CODEC): a pair of computational algorithms designed to compress the bandwidth required to process large streams of data. The most common applications of CODECs are in computer video/animation applications, where it is physically impractical to process the uncompressed raw data needed to maintain the illusion of continuous motion within a computer animation.

Computational Mechanics: the field oriented toward integrating principles of mechanics with tools from computer science and mathematics toward the goal of constructing accurate numerical models of mechanical systems.

Computer-Aided Design (CAD): a general term outlining areas where computer technology is used to speed up or eliminate efforts by engineers within the design/analysis cycle.

Computer-Aided Engineering (CAE): application of the computer toward general tasks encountered in mechanical engineering, including analysis, design, and production.

Computer-Aided Manufacture (CAM): the integration of the computer into the manufacturing process, including such tasks as controlling real-time production devices on the factory floor.

Computer-Assisted Software Engineering (CASE): software development tools oriented toward utilizing the computer to perform many rote programming development tasks.

Computer Numerical Control (CNC): the creation of mechanical parts using computer-controlled machinery such as lathes, mills, or wire electric discharge machines.

Convergence: in the setting of computer-aided engineering, convergence is taken as a measure of how rapidly the computer approximation tends toward the "exact" solution, measured as a function of increased computational effort. Ideally, as more computer effort is expended, the computational approximation rapidly becomes appropriately more accurate.

Data base Management: computational techniques specifically oriented toward allocation, manipulation, storage, and management of organized data structures.

Discretization: the fundamental tenet of most computational models for physical systems, in which a complex continuous real system is approximated by a simpler discrete replacement. In theory, as the level of refinement of the discretization is increased (i.e., as the discrete components get simultaneously smaller and more numerous), the discrete system more closely approximates the real one.

Distributed Multiprocessing: a parallel computer architecture characterized by independent processors working cooperatively on a distributed set of resources (for example, a network of computers, each equipped with its own local memory and input/output devices). In a distributed multiprocessing system, the communication bandwidth of the network connecting the individual CPUs is often the limiting factor in overall performance.

Engineer-in-the-Loop: a class of computer-aided design applications oriented toward immersing the engineer into the computational simulation with the goal of separating the rote calculation components of the task from the decision-making processes. The latter efforts are handled by the engineer, while the former are dealt with by the computer.

Eulerian Reference Frame: a reference frame attached to a region of space, to be used for constructing computational mechanics approximations. Eulerian reference frames are generally used for fluids, though they may also be profitably utilized in modeling the large deformations of solids.

Expert Systems: a class of artificial intelligence applications based on emulating the decision-making capacities of human experts by using specialized computer software for inference and knowledge representation.

Expressiveness: the desirable characteristic of a program language's facility for controlling program execution and abstracting programming functions and data.

Finite-Difference Method (FDM): a family of numerical approximation methods based on replacing the differential equation-based statement of a physical problem by a set of discrete difference equations that directly approximate the underlying differential equations.

Finite-Element Method (FEM): a family of numerical methods based on approximation of integral statements of the problems of mathematical physics. This emphasis on integral formulations (in contrast to the differential formulations used by finite-difference models) provides considerable numerical advantages, such as better convergence properties and the ability to model discontinuous problem data.

Finite-Volume Method (FVM): a family of numerical methods based on integrating the differential statements of physical problems over small subdomains. Finite-volume modeling is essentially a means to maintain the simplicity of formulation of finite-difference models, while permitting realization of some of the advantages of integral finite-element schemes.

Fuzzy Methods: computational methods for artificial intelligence that generalize the simple Boolean logical theories that traditionally form the basis of deterministic computer programs. These methods are based on more qualitative (i.e., "fuzzy") representations of data and instructions than commonly found in conventional computer programs.

Graphical User Interface (GUI): a model for software design that emphasizes graphical human/computer interaction through the desired characteristics of responsiveness, intuitiveness, and consistency.

Inheritance: a feature of object-oriented programming languages that permits derivation of new objects using a child–parent hierarchy. Inheritance makes is possible to express relationships among objects in a flexible manner that facilitates programmer productivity and the reuse of code.

Lagrangian Reference Frame: a reference frame attached to a physical body, to be used for constructing computational mechanics approximations. Lagrangian reference frames are primarily used for solids (where it is relatively easy to track the deformation of the underlying physical domain), and generally result in a simpler formulation of the conservation laws used in engineering mechanicsmultitasking

Multiprocessing: the process of distributing threads of execution over different processors on the same computer, or over different computers connected over a network. The various flavors of multiprocessing form the architectural divisions among different types of parallel processing.

Multitasking: managing the execution of separate computer programs concurrently by allocating processing time to each application in sequence. This sequential processing of different applications makes the computer appear to be executing more than one software application at a time.

Multithreading: managing the execution of separate concurrent threads of program execution by allocating processing time to each thread. This low-level software resource management model forms the basis of many multitasking operating systems and permits various forms of parallel computation by distributing distinct threads over different processors.

Neural Networks: a class of largely self-programming artificial intelligence applications based on webs of interconnected computational processes. The aggregate goal of the interconnected network of computation is to emulate human thought patterns, and especially the autonomous programming routinely exhibited by humans and notably lacking in computers.

Newton-Raphson Iteration: the basic algorithm for solving nonlinear problems, in which a nonlinear equation set is traded for a sequence of linear equation sets, with the ultimate goal of rapid convergence of the linear equation sequence to the solution of the underlying nonlinear problem.

Object-Oriented Programming: a standard for software development that emphasizes modeling the data and behavior of the underlying objects that define the program. Object-oriented programs are comprised of a collection of cooperating software objects, with these objects derived from an underlying hierarchical organization of generic classes. Contrast this concept with procedural programming, which uses algorithms as the fundamental component of program architecture.

Operating System: the software layers that provide low-level resources for execution of computer programs, including file services for input/output, graphical display routines, scheduling, and memory management.

Polymorphism: a feature of object-oriented programming languages where a single name (such as the name of a variable) can be recognized by many different objects that are related by class inheritance. Polymorphism is one of the most important features of object-oriented programming, as it permits many different objects to respond to a common event, a process that is considerably more difficult to implement using procedural programming languages.

Portability: the highly desirable program characteristic of ease of implementation on different types of computers. Portable programs are generally written in higher-level computer languages, such as C.

Procedural Programming: computer software development oriented toward abstraction of programming function by subdividing the overall task into successively smaller subtasks, which are generically referred to as procedures. Procedures represent the software implementation of algorithms, and hence procedural programming is oriented toward modeling the algorithms of the underlying system.

Random Access Memory (RAM): the collection of semiconductor hardware used for memory storage on modern computers. The main memory of a computer is usually composed of Dynamic Random Access Memory (DRAM), which is volatile and does not persist when the power is turned off.

Rapid Application Development (RAD): programming tools oriented specifically designed to aid in development of prototype software applications for demonstration and initial application design.

Relational Data Base: a standard for data base management emphasizing modeling data by using flexible relations among the various data structures of the program.

Schema: the structure of a data base (the term schema generally is taken to mean the structure and organization of the tables in a tabular data base).

Software Engineering: the profession concerned with the effective specification, design, implementation, and maintenance of computer programs.

Solvability: mathematical conditions relating to the existence and uniqueness of solutions to problems in mathematical physics. Problems that do not possess unique solutions often lead to pathological behavior when implemented on the computer, so determining the solvability characteristics of a problem is an important first step in computational engineering applications.

Spiral Model: a standard for the incremental design and implementation of computer programs, characterized by applying regular updates to a flexible design during prototyping and delivery phases.

Stability: the numerical characteristic by which errors can be guaranteed not to increase in magnitude as they propagate throughout the mathematical domain of the problem. Unstable methods permit errors to grow until the computed solution may become hopelessly corrupted, while stable methods provide some insurance against this unfortunate outcome.

Stress Analysis: a family of computational schemes oriented toward determining the state of stress and deformation of a physical medium. Stress analysis is normally associated with solid mechanics, though its methods can be successfully applied to fluid systems as well.

Structured Grid: used to characterize finite-difference and finite-volume discretizations, where considerable regularity (such as an evenly gridded rectangular geometry) is imposed on the topological structure of the underlying grid. Structured grids are generally associated with optimal convergence characteristics, but may not accurately represent the geometry of the underlying physical problem.

Structured Query Language (SQL): SQL is a standard and portable language for creating and modifying data bases, retrieving information from data bases, and adding information to data bases.

Symmetric Multiprocessing (SMP): a parallel computer architecture generally characterized by independent processors working cooperatively on a shared set of resources, such as memory or input/output devices. In a symmetric multiprocessing system, the issue of contention for resources is often the limiting factor in overall performance.

Unstructured Grid: used to characterize finite-difference and finite-volume discretizations, where little regularity is imposed on the geometric and topological structure of the underlying grid. Unstructured grids are more easily adapted to realistic problem formulations, but require more effort and often result in poorer convergence characteristics.

Vectorization: a process by which certain types of numeric computation (most notably, those involving well-defined operations on vectors) can be internally pipelined on a computer's processing units, leading to considerable efficiencies in performing the associated vector operations.

Virtual Memory: an operating system characteristic where some contents of main memory are temporarily cached on a persistent storage device in order to permit allocation and management of memory exceeding the physically available supply.

Virtual Reality: a class of computer graphics applications specifically designed to maintain the illusion of an artificial world in which the user is immersed.

Visualization: the graphical representation of data in order to facilitate understanding by human observers. The art and science of visualization exist independently of computers, but their principles are widely used in computer graphics to display the data sets that are commonly encountered in large-scale computation.

Waterfall Model: a standard for the sequential design and implementation of computer programs, characterized by long lead times to delivery and limited scope for substantial modifications to the program's function or architecture.

References

Ames, W.F. 1977. *Numerical Methods for Partial Differential Equations*. Academic Press, New York.

Apple Computer, Inc. 1985. *Inside Macintosh*, Vol. 1. Addison-Wesley, Reading, MA.

Barton, J.J. and Nackman, L.R. 1994. *Scientific and Engineering C++: An Introduction with Advanced Techniques and Examples*, Addison-Wesley, Reading, MA.

Booch, G. 1994. *Object-Oriented Analysis and Design with Applications*, 2nd ed. Benjamin/Cummings, Redwood City, CA.

Brooks, F.P. 1975. *The Mythical Man-Month*, Anniversary ed. Addison-Wesley, Reading, MA (original edition published in 1975).

Brebbia, C.A. and Dominguez, J. 1989. *Boundary Elements: An Introductory Course*. McGraw-Hill, New York.

DOD Trusted Computer System Evaluation Criteria. December 1985. DOD 5200.28-STD.

Fausett, L.V. 1994. *Fundamentals of Neural Networks: Architectures, Algorithms, and Applications*. Prentice-Hall, Englewood Cliffs, NJ.

Foley, J.D. and VanDam, A. 1982. *Fundamentals of Interactive Computer Graphics*. Addison-Wesley, Reading, MA.

Hughes, T.J.R. 1987. *The Finite Element Method: Linear Static and Dynamic Finite Element Analysis*. Prentice-Hall, Englewood Cliffs, NJ.

Humphrey, W.S. 1989. *Managing the Software Process*. Addison-Wesley, Reading, MA.

Kernighan, B.W. and Plauger, P.J. 1976. *Software Tools*. Addison-Wesley, Reading, MA.

Kernighan, B.W. and Plauger, P.J. 1978. *The Elements of Programming Style*. McGraw-Hill, New York.

Keyes, J. (ed.). 1994. *The McGraw-Hill Multimedia Handbook.* McGraw-Hill, New York.

Klir, G.J. and Folger, T.A. 1988. *Fuzzy Sets, Uncertainty, and Information.* Prentice-Hall, Englewood Cliffs, NJ.

Kreith, F. and Bohn, M.S. 1993. *Principles of Heat Transfer,* 5th ed. West Publishing, St. Paul, MN.

Kurzweil, R. 1990. *The Age of Intelligent Machines.* MIT Press, Cambridge, MA.

Metzger, P. 1983. *Managing a Programming Project.* Prentice-Hall, Englewood Cliffs, NJ.

Richardson, L.F. 1910. *Philos. Trans. R. Soc.* A210: 307.

Shapiro, S.C. (ed.). 1992. *Encyclopedia of Artificial Intelligence*, 2nd ed. John Wiley & Sons, New York (1st ed., 1987).

Shugar, T. 1994. Visualization of skull pressure response: a historical perspective. In Proc. of the Chico NSF Workshop on Visualization Applications in Earthquake Engineering, Chico, CA.

Wasserman, P.D. 1989. *Neural Computing: Theory and Practice*. Van Nostrand Reinhold, New York.

Yourdon, E. 1982. *Writings of the Revolution: Selected Readings on Software Engineering.* Yourdon Press, New York.

Yourdon, E. 1993, *Decline and Fall of the American Programmer.* Yourdon Press, New York.

Zienkiewicz, O.C. and Taylor, R.L. 1991. *The Finite Element Method*, 4th ed. McGraw-Hill, Berkshire, England.

Kreider, J.F.; et. al. “Environmental Engineering”
Mechanical Engineering Handbook
Ed. Frank Kreith
Boca Raton: CRC Press LLC, 1999

Environmental Engineering

Jan F. Kreider
University of Colorado and JFK & Associates, Inc.

Ronald R. Hewitt Cohen
Colorado School of Mines

Nevis E. Cook, Jr.
Colorado School of Mines

Peter S. Curtiss
University of Colorado and Architectural Energy Corp.

Tissa Illangasekare
University of Colorado

Frank Kreith
University of Colorado

Ari Rabl
Ecole des Mines de Paris

Paolo Zannetti
Failure Analysis Associates

16.1 Introduction

Environmental Engineering and the Role of Mechanical Engineers

Ari Rabl and Jan F. Kreider

The subject of environmental science and management is vast and interdisciplinary, ranging from highly technical matters such as the design of emission control equipment to socioeconomic matters such as the valuation of the impacts of pollution. The goal is to prevent or reduce undesirable impacts of human activities on the environment. Within this endeavor there are several different areas where mechanical engineers can make important contributions. One type of contribution concerns the design of equipment, in particular for the control of emissions; an example is an electrostatic precipitator to reduce emissions of particulates from the flue gas of a boiler or furnace. Another type of contribution concerns the modeling of the dispersion of pollutants in the environment. This chapter covers air pollution, surface water pollution, and groundwater pollution. Since space is limited and since mechanical engineers are most likely to be involved in air pollution analysis and abatement projects, the emphasis is on air pollution problems.

Environmental Burdens and Impacts

Ari Rabl

As general guidance to the field of environmental engineering, it may be helpful to present the most important environmental burdens and impacts in the form of a matrix, as in Figure 16.1.1. Burdens, e.g., the emission of a pollutant, are listed in the column on the left; impact categories are listed as a row at the top. Each element in this matrix corresponds to a specific impact of a specific burden. An X indicates that the impact from the corresponding burden is likely to be significant. Particulate air pollution, for instance, has been shown to cause a significant increase in mortality.

As an added feature we have indicated the spatial and temporal extent to the burdens. The classic air pollutants (particulates, NO_x, and SO_x) are dispersed over distances on the order of a thousand kilometers, and they affect essentially only the present generation; thus, the second and third columns show the letters R (for regional) and P (for present generation). Global warming from greenhouse gases, on the other hand, affects the entire globe and will persist over decades or centuries, hence the letters G (for global) and P, F (for present and future generations).

The classification in Figure 16.1.1 is not without ambiguities or problems. For example, we have indicated the impact of greenhouse gases as "climate change" only, even though this category includes such effects as deaths from flooding. The relative importance of impacts may change with improved scientific understanding and with the evolution of societal preferences. One should also note that the assignment of effects to causes is in many cases quite uncertain; for instance, the observed mortality from air pollution could be due to particulates or due to SO_2.

Some impacts, especially thermal pollution and noise, can be highly site dependent. The cooling water from a power plant, for instance, could damage the ecosystem of a river or it could be used to improve the output of a fishery.

Each of the categories in Figure 16.1.1 could be broken down into subcategories:

- Health
  - Mortality
  - Injury
  - Cancer
  - Respiratory illness
- Natural environment
  - Recreational value of land (including forests)
  - Recreational value of water (including fishing)
  - Biodiversity
- Agricultural environment
  - Crops
  - Cattle (milk, meat, fur, …)
  - Wood production by forests
  - Commercial fishing
- Man-made environment
  - Functional buildings
  - Historical buildings
  - Other objects (bridges, cars, …)
  - Noise

	Impacts						
	Extent				Environment		
Burdens	**Space**	**Time**	**Climate**	**Health**	**Natural**	**Agricultural**	**Man-Made**
Primary air-pollutants							
CO_2	G	P, F	X				
CH_4	G	P, F	X				
Other greenhouse gases	G	P, F	X				
Particulates	R	P		X			X
SO_2	R	P		X	X	X	X
NO_x	R	P		X	X	X	X
CO	R	P		X			
Heavy metals (Pb, Hg, Cd, …)	R	P, F		X	X		
Toxic organic compounds (e.g., dioxins)	R	P, F		X	X		
VOC (volatile organic compounds, etc.)	R	P		X			
Secondary air pollutants							
O_3 (from NO + VOC)	R	P	X	X	X		
Acid rain (from NO_x, SO_x)	R	P		X	X	X	X
Aerosols (from NO_x, SO_x, etc.)	R	P	X	X	X	X	X
Liquid residues							
Heavy metals (Pb, Hg, Cd, …)	L, R	P, F		X	X		
Toxic organic compounds (e.g., dioxins)	L, R	P, F		X	X		
COD (chemical oxygen demand)	L, R	P, F		X	X	X	
BOD (biological oxygen demand)	L, R	P, F		X	X	X	
Solid residues	L	P, F		X	X	X	
Other							
Thermal	L	P			X		
Noise, odor	L	P					

Impacts: X = potentially important; blank = usually not important. **Extent:** L = local (up to tens of kilometers); P = present generation; R = regional (hundreds to thousands of kilometers); G = global; F = future generations.

FIGURE 16.1.1 Overview of environmental burdens and major impact categories, with approximate indication of typical geographic extent and typical importance of impact.

16.2 Benchmarks and Reference Conditions

Ari Rabl, Nevis Cook, Ronald R. Hewitt Cohen, and Tissa Illangasekare

Natural Environment

Air Basins

Unpolluted air is an idealization, but its composition has been defined as indicated in Table 16.2.1. Unfortunately, measurements of truly unpolluted air were not and can never be made because measurement techniques and even the interest in measurements did not exist when air was unpolluted. Now even the most remote sites have mildly polluted air.

TABLE 16.2.1 Gaseous Composition of Unpolluted Air (dry basis)

	ppm (vol)	μg/m^3
Nitrogen	780,000	8.95×10^8
Oxygen	209,400	2.74×10^8
Water	—	—
Argon	9,300	1.52×10^7
Carbon dioxide	315	5.67×10^5
Neon	18	1.49×10^4
Helium	5.2	8.50×10^2
Methane	1.0–1.2	$6.56–7.87 \times 10^2$
Krypton	1.0	3.43×10^3
Nitrous oxide	0.5	9.00×10^2
Hydrogen	0.5	4.13×10^1
Xenon	0.08	4.29×10^2
Organic vapors	~0.02	—

Although measurements of the undisturbed atmosphere were not made, we can gain some insight into trends of air pollutant burden growth by examining emissions over the last century. Figure 16.2.1 shows the growth of emissions of the classical air pollutant species since 1900. The latter two panels of the figure show that emissions from regulated sectors in the United States have abated in the past 20 years. Those regulations are described later. Note that Table 16.2.1 uses two sets of units for gaseous pollutants, one volumetric, the other mass based. To convert from one to the other the ideal gas law is used with the result (at 1 atm and 25°C):

$$1 \text{ ppm} = \text{MW} * 40.9 \ \mu\text{g/m}^3$$

where MW is the molecular weight.

Surface Water Basins

Human activity has also dramatically altered the distribution, quantity, and quality of Earth's water, especially since the industrial revolution. Accurate measurement of many water impurities, particularly trace impurities, has only become possible in the latter part of the 20th century. Given the quantities and wide distribution of human-generated wastes delivered directly or indirectly (via atmospheric deposition or surface runoff and erosion) to water bodies, recent water quality surveys might not be representative of truly "natural" conditions. As an example, a "pristine," undeveloped alpine lake may show traces of plutonium that are residuals of 1950s through 1960s atmospheric testing of nuclear weapons. Lead from automobile emissions can be detected in the bottom sediments of the Atlantic Ocean, more than 1500 km from the nearest landmass. A tabulation of the averages and ranges of concentrations of many naturally occurring substances detected in minimally impacted waters can serve

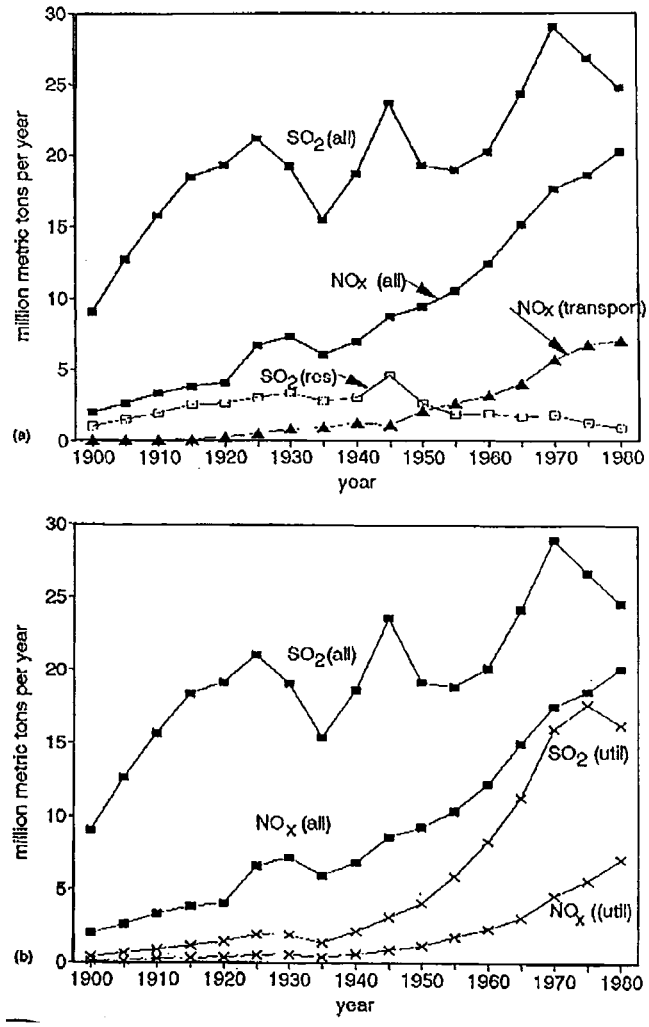


FIGURE 16.2.1 Trends in U.S. national emmissions of air pollutants. (a) SO_2 and NO_X ,1900–1980, all sources, residential SO_2 and transportation NO_X. (b) SO_2 and NO_X, 1900–1980, all sources and electric utility sources. (c) SO_2 and NO_X, 1970–1990, all sources, transportation and electric utility sources. (d) SO_2 and NO_X, and particulates, 1970–1990. (Data from EPA annual reports; Gschwandtner et al., 1985.)

as a benchmark, admittedly imperfect, against which to compare "polluted" waters, water quality criteria, and regulatory standards.

Choice of Reference Conditions. The authors propose as reference conditions, the concentrations of measured impurities in the oceans, "average" rainwaters, and "average" river waters. This choice of reference waters was based on two distinct criteria:

1. Either the volume of water was so large (such as the open oceans) that human activity has had little detectable effect on average water quality conditions, or
2. The waste input is small and the water body is rapidly and continuously renewed by unpolluted sources (such as a tributary to a major river).

The major and minor constituents of water and their quantities are easily presented in tables. The use of the word "major" indicates materials present or required in large quantities. Table 16.2.2 summarizes data on the constituents of selected natural waters. Note that the concentration data for major constituents

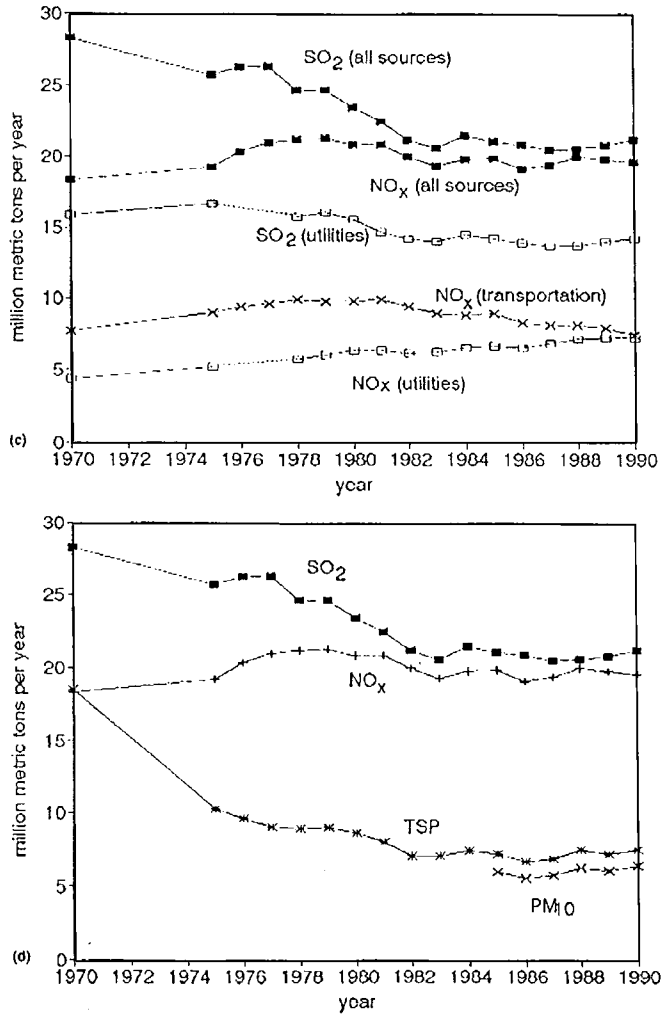


FIGURE 16.2.1 continued

(Table 16.2.2) is given in milligrams per liter and the concentration data for the minor constituents (Table 16.2.3) is given in micrograms per liter. Inclusion of boron and fluoride as major constituents is somewhat arbitrary and was based on their occurrence at greater than 1 mg/L in seawater. In these tables, individual entries are *average values* reported by U.S. regional or national surveys. *Ranges* are derived from the majority of the data from similar surveys, excluding data obtained from waters apparently contaminated by pollution. Some surveys presented results by stating that "very few" samples exceeded certain concentration values or simply reported measurements as "less than" due to analytical detection limits. These results are preceded by the symbol < in the tables. Although the data given are believed to be representative of water found in natural settings, it should be appreciated that, especially for industrially important trace constituents, the upper limits of the concentration ranges may include some small level of **anthropogenic** inputs.

There are additional categories of natural and contaminant components of water. There are complex, difficult to characterize, organic humic materials in natural water bodies that represent the end point of decay of formerly living organic matter. There are synthetic chemicals by the thousands, produced currently and in the past, by the chemical industry. Some of these compounds and polymers are completely new to the natural environment. Many are considered toxic to aquatic organisms and/or

TABLE 16.2.2 Major Constituents and Characteristics of Natural Waters (constituent concentrations in mg/L, unless otherwise noted)

Constituent	Oceans	Rivers	Rain
Cl^-	19,000	5.8, 7.8	0.2–0.6
Na^+	10,500	5.3, 6.3	0.2–0.6
SO_4^{2-}	2,700	8.3, 11	1.1, 2.2
Mg^{++}	1,350	3.4, 4.1	0.3–1.5
Ca^{++}	410	13.4, 15	0.05–1.5
K^+	390	1.3, 2.3	0.07–0.11
HCO_3^-	142	52, 58	1–10
Br^-	0.67	<20	<0.15
Sr^{++}	8	0.06–0.11	—
SiO_2	6.4	10.4, 13	0.1
B^{3-}	4.5	0.3	—
F^-	1.3	<1	—
pH (units)	8.2	6, 7.2, 7.5	5.7
Hardness (total)	—	10–200	—
Ammonia, as N	—	0.05–0.5	—
Nitrate, as N	—	0.1–2.0	—
BOD	—	2–4	—

Note: BOD is biochemical oxygen demand, mg/L oxygen; nitrate is NO_3^-, ammonia is NH_3.

TABLE 16.2.3 Minor Constituents of Natural Waters (constituent concentrations in μg/L)

Constituent	Oceans	Rivers	Rain
N	670	0–1000	0–620
C (organic)	100	3, 6, 19	—
P	90	10–30	—
Ba	20	43, 45	—
Zn	10	5–45, 10, 20	3.6
Ni	7	0.3, 10	—
As	3	0.15–0.45	0.45
Cu	3	10	2.5
Fe	3	10	—
Mn	2	—	—
Sb	0.3	0.54	—
Ag	0.3	0.3	0.001–0.1
Hg	0.2	<0.3	0.2
Cd	0.11	1, <10	—
Se	0.09	0.1, 0.2	—
Cr	0.05	0.43, 1.4, 5.8, <10	0.1–.02
Pb	0.03	<1	<1

humans. For lists of contaminants considered toxic and their effluent limitations, there are many documents available from the U.S. Environmental Protection Agency (USEPA) and other sources.

Soils and Water Basin Definitions

The unconsolidated sediment that covers a comparatively thin mantle of the land surface in general is referred to as soil. Soils are complex mixtures of solid, liquid, and gases. The solid phase consists of a mineral inorganic fraction that is produced by weathering of rocks or transported material. The predominant inorganic elements are silicon, aluminum, and iron. The organic fraction consists of partially or fully decomposed products of flora and fauna. The liquid phase is the water that occupies the pore spaces between and within grains of the solid material. In its natural form this water, which is referred to as

soil water, contains dissolved substances introduced from the solids or transported from the ground surface. The pore spaces that are not occupied by water will be filled by water vapor, gases, and air. The complexity of soil systems derives from the fact that the mixture of the solid, liquid, and gases is very heterogeneous and that the composition of the individual phases and the mixture changes in space and with time.

The soils in their natural environment can be subjected to drastic changes not only as a result of the interaction among the solid, liquid, and gas phases but also by external factors that are controlled by pressure, temperature, and light. The physical and chemical characteristics of the changing solid phase have a significant influence on the thermal behavior, water retention and flow, adsorption and entrapment of chemicals and wastes, and transport of dissolved substances. All of the above processes are of importance in the study, understanding, and solution of problems in environmental sciences and engineering associated with soil contamination.

The physical characteristics of soils as a porous medium are affected by the shape, size and size distribution, and the arrangement of the solid phase or the soil grains. The *shapes* and *sizes* of the soil grains vary widely from small colloids to large sand and gravel. Particles that are less than the arbitrary size of 2 μm are the clay fraction that is formed as a secondary product of weathering of rocks or derived from the transported deposits. These particles are platelike or disk shaped. The non-clay fraction formed from inert minerals and fragments of rock consists of silt, sand, and gravel. In a particle size classification used by agricultural scientists that was developed by the U.S. Department of Agriculture (USDA), non-clay particles that are in the size range 2 to 50 μm are classified as silts, in the range 50 to 2000 μm as sands, and above 2000 μm as gravel. In this classification system the sands are further divided into subgroups of very fine, fine, medium, coarse, and very coarse sands. A second system, by the International Society of Soil Science (ISSS) that is also used by agricultural scientists, classifies silts to be in the range 2 to 20 μm, fine sands 20 to 200 μm, and coarse sands 200 to 2000 μm. The American Society of Testing and Materials (ASTM) classifies colloids as particles that are less than 0.1 μm in size.

The surface area of the solid particles contained in a known volume of the soil has a significant influence on the physical and chemical processes that occur on the surfaces of the soil grains. The ratio of the internal solid surface area to the total volume is referred to as the *specific surface*. This parameter is also sometimes expressed as the ratio of the surface area to the mass of the soil grains. This parameter of the soil is affected by the size and shape of the individual soil grains. Because of the platelike or disk-shaped nature of small clay particles, clays in general have very large specific surface compared with the non-clay particles (silts, sand, and gravel). For example, the three common clay minerals kaolinite, illite, and montmorillonite have specific surface of 45, 175, and 800 m^2/gm, respectively (Corey, 1994). Because of these enormously large surface areas, the clay fraction in soils has a significant influence on the chemical reactive processes that are controlled by the surface area of the soil grains. The silt and sand fraction will not have a significant influence on the chemical processes, and also the smaller surface areas result in small water retention capacities as compared with clay.

The spaces between the grains that are referred to as intergranular pore spaces control the flow behavior and capacity to hold water by a soil. A physical parameter that is known as the *average porosity* or *porosity* characterizes the secondary pore space enclosed between the aggregates. The porosity of a soil sample is defined as the ratio of the volume of interconnected pores to the total volume of the sample. Soil porosity depends on many factors that include its structure, shape of soil grains, size distribution of the grains, the degree of mixing of the various-sized particles, and the way the soil grains are packed. Under conditions of normal packing the porosities of unconsolidated sand vary in the range 0.39 to 0.41 and soil with structure in the range 0.45 to 0.55. The organic matter in soil binds the inorganic fraction to form larger aggregates. The primary pore spaces within the soil aggregate play a significant role in retardation and attenuation of dissolved chemicals that are contained in the aqueous phase.

The subsurface soil–water environment that exists below the ground surface is divided into two zones, namely, the *unsaturated* or *vadose zone* and the *saturated zone*. In the unsaturated zone pore spaces contain both water and air. The capillary forces created by the surface tension in the water–air interfaces produce water pressures that are less than atmospheric (suction). The water flowing through

the unsaturated zone is subjected to capillary driving forces in addition to the gravitational forces. In the saturated zone, the water pores are fully saturated with water and the water pressures are greater than atmospheric. The water in the saturated zone is driven by gravitational forces. The surface that separates the unsaturated and saturated zones is referred to as the *water table*. By definition, the water pressure at the water table is atmospheric. The location of the water table below the ground surface at waste and spill sites becomes critical in determination of the fate and transport of pollutants in the subsurface soil–water environment as described later. Figure 16.2.2 shows the essentials of the situation schematically.

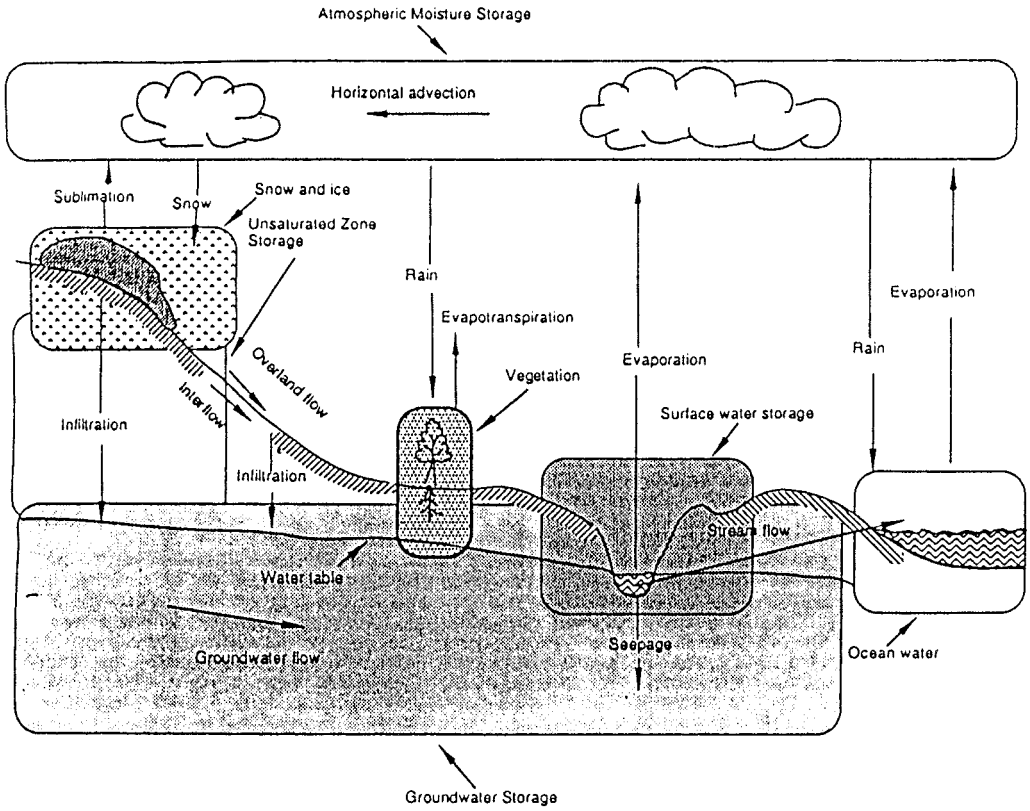


FIGURE 16.2.2 A column of Lagrangian boxes simulating atmospheric dynamics.

Acceptable Levels of Pollutants

Water

Criteria vs. Standards. The *objective* of water quality control programs, for example, is to protect current or potential *beneficial uses* of water. Criteria define the specific characteristics necessary to protect such uses. Criteria are not absolute, but are based on a combination of experience, current scientific knowledge, and the judgment of experts. Toxicological data for many of the water components often are limited. There may be results of tests for acute, or immediate, impact to organisms (including humans), but little information on impact on reproduction, long-term health or tolerance at low concentrations, organism adaptability, acclimatization, or interaction with other substances. General criteria are intended to protect sensitive species subjected to long-term exposure and may be overly conservative in specific situations. The aquatic life column of Table 16.2.4 presents typical, continuous, maximum concentration criteria for freshwaters intended to support a variety of aquatic life. Surface water quality criteria and their

TABLE 16.2.4 Water Quality Criteria and Standards (inorganic contaminants in μg/L)

Constituent	Beneficial Use or Protection Category		
	Aquatic Life	Irrigation	Drinking Water
B	—	750, 500–3,000	—
F	—	1,000	4,000[a]
Ammonia-N	20[b]	—	—
Nitrite-N	—	—	1,000
Nitrate-N	—	5,000–30,000	10,000
Ba	—	—	2,000
Zn	110	2,000	5,000[c]
Ni	160	200	—
As	190	100	50
Cu	12	200, 100–1,000	1,300[c]
Fe	1,000	5,000	300[c]
Mn	—	200	50[c]
Sb	—	—	6
Ag	3[d]	—	—
Hg	0.012, 0.05[e]	—	2
Cd	1.1	10	5
Se	5	20	50
Cr	11	100	100
Pb	3.2	—	15
pH, units	No vertebrates below 4–4.5	—	6.5–8.5

[a] Considered beneficial in drinking water at somewhat lower concentrations.
[b] Temperature and pH dependent, only unionized form (NH_3) is toxic to aquatic life.
[c] Secondary standards based on use impacts (tastes, staining), not health related.
[d] Controversial, based on toxicity of Ag^+, rarely present in natural waters.
[e] Extremely low levels partly due to food chain bioaccumulation potential.

relationship to permitted discharges are further discussed later in this chapter. The irrigation criteria given in Table 16.2.4 are intended to protect the yields of agricultural crops.

Due to resource limitations (time, money, data, or a shortage of knowledgeable experts), conservative criteria are often adopted as *standards*. Standards development should, however, include considerations of actual local needs and conditions, economics, technical feasibility, and the defined objectives of environmental policy. In contrast to criteria, standards are usually incorporated into laws and regulations and are often absolute. Either standards are violated or they are not.

Summarized as simple, USEPA (1979) definitions follow:

1. *Water Quality Criterion:* That concentration of a water quality measure that will meet a specific water use.
2. *Water Quality Standard:* The translation of a water quality criterion into a legally enforceable mass discharge of effluent limitation.

The drinking water column of Table 16.2.4 presents a subset of the current U.S. National Standards for drinking water supplied to the public. The health-based standards are enforceable. It should be noted that the drinking water standards list in Table 16.2.2 is not comprehensive. Maximum concentration levels (MCLs) for some inorganics, microbial contaminants, and the many regulated toxic organics have been omitted. Many drinking water MCLs are controversial because they rely on dose–response models which cannot be directly verified at the low levels of exposure that are typical.

Air

Tables 16.2.5 to 16.2.9 summarize the current U.S. ambient air quality standards for carbon monoxide, sulfur dioxide, nitrogen dioxide, ozone, and particulate matter. These criteria represent myriad effects depending on the receptor, species of pollutant, and duration and severity of impact. These criteria are

TABLE 16.2.5 U.S. Ambient Air Quality Criteria for Carbon Monoxide

CoHb in Blood, %	Associated Human Symptoms
80	Death
60	Loss of consciousness; death if exposure is continued
40	Collapse on exercise; confusion
30	Headache, fatigue; judgment disturbed
20	Cardiovascular damage; electrocardiographic abnormalities
5	Decline (linear with increasing CoHb level) in maximal oxygen uptake of healthy young men undergoing strenuous exercise; decrements in visual perception, manual dexterity, and performance of complex sensorimotor tasks
4	Decrements in vigilance (i.e., ability to detect small changes in one's environment that occur at unpredictable times); decreased exercise performance in both healthy persons and those with chronic obstructive pulmonary disease
3–6	Aggravation of cardiovascular disease (i.e., decreased exercise capacity in patients with angina pectoris, intermittent claudication, or peripheral arteriosclerosis)

CoHb = carboxy hemoglobin.

Source: Henderson, Y. and Haggard, H.W., *Noxious Gases,* Chemical Catalog Co., New York, 1927; and USEPA, Air Quality Criteria for Carbon Monoxide, EPA/600/8-90/045F, Research Triangle Park, NC, December 1991.

TABLE 16.2.6 U.S. Ambient Air Quality Criteria for Sulfur Dioxide

Concentration of SO_2 in Air (ppm)	Exposure Time	Human Symptoms and Effects on Vegetation
400	—	Lung edema; bronchial inflammation
20	—	Eye irritation; coughing in healthy adults
15	1 hr	Decreased mucociliary activity
10	10 min	Bronchospasm
10	2 hr	Visible foliar injury to vegetation in arid regions
8	—	Throat irritation in healthy adults
5	10 min	Increased airway resistance in healthy adults at rest
1	10 min	Increased airway resistance in people with asthma at rest and in healthy adults at exercise
1	5 min	Visible injury to sensitive vegetation in humid regions
0.5	10 min	Increased airway resistance in people with asthma at exercise
0.5	—	Odor threshold
0.5	1 hr	Visible injury to sensitive vegetation in humid regions
0.5	3 hr	U.S. National Secondary Ambient Air Quality Standard promulgated in 1973
0.2	3 hr	Visible injury to sensitive vegetation in humid regions
0.19	24 hr[a]	Aggravation of chronic respiratory disease in adults
0.14	24 hr	U.S. National Primary Ambient Air Quality Standard promulgated in 1971[b]
0.07	Annual[a]	Aggravation of chronic respiratory disease in children
0.03	Annual	U.S. National Primary Ambient Air Quality Standard promulgated in 1971[b]

[a] In the presence of high concentrations of particulate matter.

[b] *Sources:* Air Quality Criteria for Particulate Matter and Sulfur Oxides, final draft, USEPA, Research Triangle Park, NC, December 1981; Review of the National Ambient Air Quality Standards for Sulfur Oxides: Assessment of Scientific and Technical Information, Draft OAQPS Staff Paper, USEPA, Research Triangle Park, NC, April 1982.

descriptive and are used for emission and air quality standards that are summarized briefly. As better data become available, these tables may change.

TABLE 16.2.7 U.S. Air Quality Criteria for Nitrogen Dioxide

Concentration of NO_2 in Air (ppm)	Exposure Time	Human Symptoms and Effects on Vegetation, Maerials, and Visibility
300	—	Rapid death
150	—	Death after 2 or 3 weeks by bronchiolitis fibrosa obliterans
50	—	Reversible, nonfatal bronchiolitis
10	—	Impairment of ability to detect odor of NO_2
5	15 min	Impairment of normal transport of gases between the blood and lungs in healthy adults
2.5	2 hr	Increased airway resistance in healthy adults
2	4 hr	Foliar injury to vegetation
1.0	15 min	Increased airway resistance in individuals with bronchitis
1.0	48 hr	Slight leaf spotting of pinto bean, endive, and cotton
0.3	—	Brownish color of target 1 km distant
0.25	Growing season	Decrease of growth and yield of tomatoes and oranges
0.2	8 hr	Yellowing of white fabrics
0.12	—	Odor perception threshold of NO_2
0.1	12 weeks	Fading of dyes on nylon
0.1	20 weeks	Reduction in growth of Kentucky bluegrass
0.05	12 weeks	Fading of dyes on cotton and rayon
0.03	—	Brownish color of target 10 km distant
0.003	—	Brownish color of target 100 km distant

TABLE 16.2.8 U.S. Ambient Air Quality Criteria for Ozone

Concentration of O_3 in Air (ppm)[a]	Human Symptoms and Vegetation Injury Threshold
10.0	Severe pulmonary edema; possible acute bronchiolitis; decreased blood pressure; rapid weak pulse
1.0	Coughing; extreme fatigue; lack of coordination; increased airway resistance; decreased forced expiratory volume
0.5	Chest constriction; impaired carbon monoxide diffusion capacity; decrease in lung function without exercise
0.3	Headache; chest discomfort sufficient to prevent completion of exercise; decrease in lung function in exercising subjects
0.25	Increase in incidence and severity of asthma attacks; moderate eye irritation
0.15	For sensitive individuals, reduction in pulmonary lung function; chest discomfort; irritation of the respiratory tract, coughing, and wheezing. Threshold for injury to vegetation
0.12	U.S. National Primary and Secondary Ambient Air Quality Standard, attained when the expected number of days per calendar year with maximum hourly average concentrations above 0.12 ppm is equal to or less than 1, as determined in a specified manner

TABLE 16.2.9 U.S. Ambient Air Quality Criteria for Particulate Matter

Concentration of Particulate Matter in Air ($\mu g\ m^{-3}$)				
TSP < 25 μm	**TP < 10 μm**	**FP < 2.5 μm**	**Exposure Time**	**Human Symptoms and Effects on Visibility**
2000	—	—	2 hr	Personal discomfort
1000	—	—	10 min	Direct respiratory mechanical changes
—	350	—		Aggravation of bronchitis
	150	—	24 hr	U.S. Primary National Ambient Air Quality Standard as of September 1987
180	90	—		Increased respiratory disease symptoms
	150	—	24 hr	U.S. Primary National Ambient Air Quality Standard as of September 1987
110	55	—	24 hr	Increased respiratory disease risk
	50	—	Annual geometric mean	U.S. Primary National Air Quality Standard as of September 1987
	50	—	Annual geometric mean	U.S. Secondary National Ambient Air Quality Standard as of September 1987
—	—	22	13 weeks	Usual summer visibility in eastern U.S., nonurban sites

TSP = total suspended particulates; TP = thoracic particulates; FP = fine particulates.

16.3 Sources of Pollution and Regulations

Sources

Jan F. Kreider, Nevis Cook, Tissa Illangasekare, and Ronald R. Hewitt Cohen

Air

Air pollutants are found in the form of gases (e.g., SO_2) and particulate matter (e.g., fine dust). They are emitted into the atmosphere from natural sources (e.g., volcanoes) and *anthropogenic* sources (e.g., industrial activities). These pollutants are called "primary" because they are directly emitted from local sources. Primary pollutants may undergo chemical reactions that result in the subsequent formation of other species called "secondary" pollutants (e.g., O_3).

Air pollution is found at all spatial scales, ranging from a few meters (e.g., indoor pollution) to local, urban, regional, and global scales (Milford and Russell, 1993). *Indoor pollution* is of great concern because a large fraction of our time is spent indoors. Indoor sources of pollutants include combustion processes, aging materials (e.g., formaldehyde emitted from particleboard and plywood), and radon — a natural indoor pollutant which migrates through the soil and may penetrate and accumulate inside buildings. The *local scale* ranges from 100 m to a few kilometers. At this scale, pollution dynamics are dominated by atmospheric diffusion and the role of primary pollutants. The local scale is the one where we experience major exposure to toxic substances and flammable compounds during catastrophic and emergency releases.

The *urban scale* ranges from 10 to 100 km and is characterized by both primary and secondary pollutants. In fact, characteristic residence times are on the order of 1 or a few days, thus, allowing enough time for chemical transformation to play a role. Two types of urban smog are well known: the "London" smog and the "Los Angeles" smog. The former is characterized by stagnant, foggy meteorological conditions in winter which allow a buildup, over a few days, of SO_2 and particulate matter. The latter, which has become the most common type of atmospheric pollution throughout the cities of the world, is a **photochemical** smog associated with clear, sunny days. Photochemical smog is a mixture of many gaseous compounds and particulate matter, among which the two most important constituents are ozone (a colorless secondary pollutant) and fine secondary particulate matter (such as sulfates, nitrates, and organic particles), which is responsible for most of the visual haze.

The *regional scale* ranges from hundreds to thousands of kilometers (the upper regional scale is also called *continental scale*). Characteristic residence times are on the order of 1 week. At this scale, pollution dynamics are dominated by chemical transformation and ground deposition phenomena. Acidic deposition, often referred to as *acid rain,* is a phenomenon in which acid substances, such as sulfuric and nitric acid, are brought to Earth by dry and wet deposition. Some lakes are very sensitive to acidic deposition because of their limited buffering capacity. *Global air pollution* is characterized by relatively unreactive compounds, such as CO_2, methane (CH_4), and chlorofluorocarbons (CFCs). The long lifetime of these pollutants allows their global dispersion and accumulation.

Air pollution at any scale creates several adverse effects. Air pollution can just be a nuisance (e.g., odors) or be aesthetically offensive (e.g., visibility impairment). Some adverse effects, however, are of fundamental importance to the welfare of the population and public health. For example, many pollutants cause respiratory effects; some pollutants have mutagenic effects; others have shown carcinogenic effects; some pollutants also have synergistic effects (e.g., the damage of SO_2 to the human respiratory system can be greatly enhanced by the presence of fine particles).

In the rest of this chapter, an overview is presented of analytical and numerical techniques for simulating air pollution phenomena. Air quality modeling is an essential tool for most air pollution studies today. Models can be divided into **deterministic** models, based on fundamental mathematical descriptions of atmospheric processes, and statistical models, based upon semiempirical relationships extracted from data and measurements. Deterministic models, in particular, have become a major tool

in providing objective assessment of air pollution scenarios and evaluating the effectiveness of different air pollution control measures. In fact, only deterministic models provide an unambiguous assessment of the fraction of responsibility of each pollution source in each receptor area, thus, allowing the definition and implementation of the most cost-effective cleanup strategy. Statistical models, instead, are useful in situations where deterministic models do not perform well, e.g., for real-time forecasting of air pollution episodes.

Waterborne

Table 16.3.1 summarizes important sources of surface water pollution species by economic or technical sectors of the U.S. economy. An X indicates a key source and no entry indicates a second-order effect at most. Table 16.3.2 lists groundwater pollution sources by cause, extent and waste type.

Soil and Groundwater

Various waste products and chemicals are generated from industrial, agricultural, commercial, and domestics activities. Unmonitored and uncontrolled long-term application on the land, accidental spillage, and improper storage and disposal result in these chemicals and wastes acting as potential sources of soil and groundwater contamination. La Grega et al. (1994) use an engineering classification system in which hazardous wastes are grouped under six categories, namely, inorganic aqueous waste, organic aqueous waste, organic liquids, oils, inorganic sludges/solids, and organic sludges/solids. In addition to these, pathogens and nuclear wastes act as sources of contamination. As most of these materials are fully or partially soluble in water, they ultimately will contaminate the water phase contained within the soil pores and the water passing through the soil both in the unsaturated and saturated zones of the subsurface. Figure 16.3.1 shows in schematic form how groundwater pollution occurs.

Inorganic aqueous wastes generated from manufacturing activities involving galvanizing, metal finishing, electroplating, etc. mostly contain acids, alkalis, or concentrated solutions of inorganic wastes such as heavy metals. Organic aqueous wastes are liquids that primarily contain mixtures of dilute concentrations of organic substances such as pesticides. Organic liquid wastes are complex mixtures or concentrated solutions of **organic compounds**. A common example of organic liquid wastes is halogenated solvents that are used in metal degreasing. Most of the oily wastes are derived from petroleum. Oils are used as fuels, lubricating oils in engines, and cutting oils in manufacturing. Inorganic sludges/solids wastes are produced from wastewater treatment, smelters, coking, and metal fabrication. Tars and sludges that are produced from manufacturing activities are some of the examples of wastes that are in the form of organic sludges.

Waste materials and chemicals are released to the soil and groundwater from various sources. The most common sources are leaking underground chemical storage tanks, septic tanks, municipal landfills, industrial landfills, surface impoundments, and abandoned hazardous waste sites. In addition, injection wells, regulated hazardous waste sites, land application, road salting, saltwater intrusion, and oil and gas brine pits contribute to soil and groundwater contamination.

The most frequent use of underground tanks is for storage of gasoline in service stations. Hazardous wastes and chemicals and oils are also stored in buried tanks at industrial sites, farms, and homes. The potential for leakage exists when these tanks corrode internally or externally. In addition, chemicals can leak from pipe joints or holes in the pipes that are connected to the tanks.

Older landfills that were not properly designed for containment of liquids or leachates were used to dispose of garbage, sludges from wastewater treatment plants, construction wastes and debris, incinerator ash, waste from foundries, and many other industrial and domestic hazardous and nonhazardous waste products. The chemicals in the landfill become a source of contamination when rain or surface water infiltrating through the landfill produces leachate. These leachates containing the dissolved constituents of the waste, eventually contaminate the underlying soil and groundwater. Modern well-designed landfills have synthetic liners and leachate-collection systems.

Surface impoundments in the form of open pits, lagoons, or ponds are designed to accept liquid wastes or mixed solids and liquids. The chemical wastes in these temporary storage sites are treated and

TABLE 16.3.1 Sources of Water Pollution vs. Contaminant Emission Categories

	General Emissions Type or Category									
Pollution Source	**BOD**	**Pathogens**	**TSS**	**Turbidity**	**TDS**	**Nutrients N, P**	**Synthetic Organics**	**Volatile Organics**	**Metals**	**pH**
Sewage	X	X	X	—	—	X	—	—	—	—
Mining operation	—	—	X	X	X	—	—	—	X	X
Timber operation	—	—	X	X	—	—	—	—	—	—
Agriculture	—	X	—	X	X	—	—	—	—	—
Food processing	X	—	X	—	X	—	—	—	—	X
Chemical manufacturing	X	—	—	—	—	—	X	X	X	X
Textiles manufacturing	X	—	X	X	X	—	—	—	—	X
Primary metals	—	—	—	—	—	—	—	—	X	X
Pulp and paper	—	—	X	X	X	—	—	—	—	X
Petroleum refining	X	—	X	—	—	X	—	X	X	X
Rubber/plastics	X	—	X	—	X	—	—	—	—	X
Septic systems	—	X	—	—	—	X	—	—	—	—
Injection wells	—	—	—	—	X	—	X	X	X	X
Sewage sludge	—	X	—	—	—	X	—	—	X	—
Urban runoff	—	X	—	X	X	—	X	X	X	—
Landfill leachate	X	—	—	—	X	X	—	—	—	—

Notation: BOD = biochemical oxygen demand; TSS = total suspended solids; TDS = total dissolved solids; synthetic organics = synthetic organic compounds; volatile organics = volatile organic compounds; metals = industrial metals; pH = high (alkaline) or low (acid) pH.

TABLE 16.3.2 Groundwater Pollution Sources

Source	Cause	Extent	Chemical/Wastes
Underground storage tanks	Hole due to internal and external corrosion leaks chemical into soil and groundwater	• 2.5 million • 47 states • 35% of 800,000 fuel tanks leaked	Gasoline Oil Hazardous chemicals Waste products
Landfills	Rainwater leaching chemicals to groundwater	• 2,395 open dumps • 24,000–36,000 closed or abandoned landfills • 75,000 on-site industrial landfills • 12,000–18,000 municipal landfills contain hazardous wastes	Garbage and trash Sludge Incinerator ash Foundry waste Hazardous substances
Surface impoundments	Direct infiltration to the saturated zone of aquifer	• 180,000 waste impoundments (1982) • 37,000 municipal • 19,400 agricultural • 27,912 industrial • 25,000 mining • 65,688 brine pits for oil and gas • Industrial sites evaluated 95% within 1 mile of drinking wells, 70% unlined, 50% on top of aquifers	Settling ponds from municipal wastewater and sewage treatment Animal feed lots and farms Oil and gas industries Mining Paper Chemical operation
Waste-disposal injection wells	Direct dumping through wells to aquifers	• Groundwater contamination in 20 states • Millions of tons through thousands of wells	Toxic Hazardous Radioactive Metals Wood Preservatives Petroleum
Septic systems	Surfacing and flooding due to failure; leaching into aquifers	• 22 million in U.S. • 0.5 million installed every year • 30% of population served	Variety of organic and inorganic compounds Fecal coliform Nitrates and nitrites Trichloroethylene, benzene, and methylene chloride
Agricultural waste	Leaching through unsaturated zone to groundwater	• 50,000 pesticides with 600 active ingredients • 10 million t of nitrogen	Nitrates Salts Dissolved solids
Land application	Leachates reaching groundwater	• Major threat to groundwater in 20 states • 40% of hazardous wastes in California treated by land farming	Heavy metals Toxic chemical Nitrogen Pathogens
Radioactive contaminants	Potential migration to groundwater	• Massive production of radioactive isotopes after World War II	Uranium Plutonium

After Bedient, P.B., Rifai, H.S., and Newell, C.J. 1994. *Ground Water Contamination: Transport and Remediation*, Prentice Hall, Englewood Cliffs, NJ, 541.

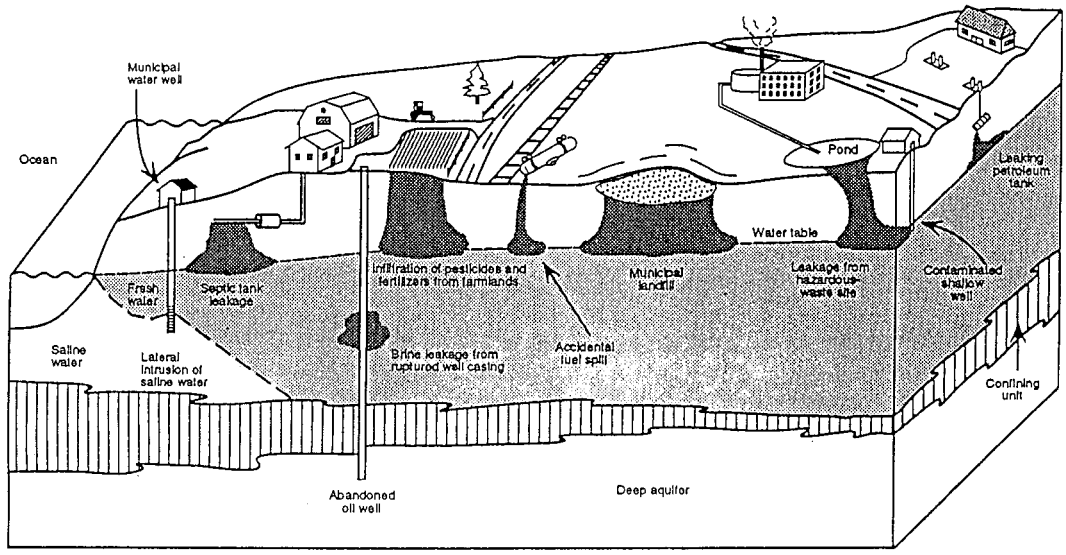


FIGURE 16.3.1 Mechanisms of groundwater contamination (From Fetter, 1993).

discharged or allowed to evaporate or infiltrate into the ground. In addition, to store and dispose of wastewater and products from sewage treatment, surface impoundments are used in paper, mining, and oil and gas industries. They are also used in farms and large feed lots. In unlined impoundments, the liquids leak, and when undetected, can result in soil and groundwater contamination in large zones of the subsurface.

Because of the large number of septic tanks that are in operation in homes, small businesses, service stations, laundries, and industry, septic tanks have become a major source of soil and groundwater contamination. When these systems fail, the sludges or septage floods the drainage field and leachates that are generated act as a source of contamination. The discharges from septic tanks contain many organic and inorganic chemical products, suspended solids, fecal coliform, pathogenic bacteria and viruses, nitrates, and nitrites. Industrial septic systems used in commercial operations discharge hazardous waste chemicals that include synthetic organics (e.g., trichloroethylene, tetrachloroethylene, toluene, etc.) and heavy metals (e.g., lead, copper, and zinc).

Injection wells that are drilled into deep aquifers or aquifers that are not used for water supply are used to discharge liquid wastes into the subsurface. Large volumes of toxic, hazardous, and radioactive wastes that are generated in chemical, petroleum, metals, mining, and wood-treatment industries are disposed of using this method. The contamination of aquifers that are used for drinking water can occur when the injection wells are not designed properly and are placed in formations where the hydrogeolgic conditions are not well known or understood.

Many types of agriculture-related activities use chemicals and produce wastes that contaminate soil and groundwater. Among these products are pesticides, fertilizer, and feed lot waste. Various types of pesticides are used in farming, golf courses, gardens, and parks to control insects and weeds. Many of the modern-day pesticides are biodegradable, but some of the stable ones are carried by rain and irrigation water through the soil to the groundwater. Fertilizers primarily contain nitrogen, potassium, and phosphorus. Because of comparatively high mobility in soil, nitrogen leachates are the primary contaminant of concern that is associated with application of fertilizers. Waste generated at large feed lots can introduce nitrogen, bacteria, and salts to the underlying aquifers. In addition to the above sources, the salts that get accumulated in soils from long-term application of irrigation water act as a source of soil and groundwater degradation.

In a treatment and disposal method known as land treatment, the wastewater and sludges that are generated from treatment plants and industrial operations are directly applied on the ground surface. Contamination of the soil and groundwater occurs when heavy metals, toxic chemicals, nitrogen, and pathogens leach through the unsaturated zone.

Another source of soil contamination is the radioactive wastes that are produced in the weapons and nuclear industries. These wastes are primarily associated with the elements uranium and plutonium. The ionizing radiation in the form of alpha and beta particles and gamma rays are damaging to human and animal life. As these contaminants follow a first-order exponential decay law, they remain hazardous for very long time periods on the order of hundreds to thousands of years. Table 16.3.2 summarizes several of the key sources of groundwater pollution, their extent and the major associated pollutants.

Pollutant Monitoring

Jan F. Kreider and Tissa Illangasekare

Groundwater

Contaminants that are released to the aqueous phase from the soil or external sources move with the flowing groundwater creating a solute plume. Monitoring the groundwater flow velocity and the propagation of the solute plume is necessary to design schemes and strategies to protect the quality of groundwater that is used for drinking.

Monitoring wells are installed in aquifers to measure water pressure and to sample groundwater water quality. Wells in general consist of a standpipe with a screened interval. The screened interval allows for the aquifer water to flow into the well pipe. The water pressures are used to determine the head gradients, which in turn can be used to estimate the magnitude and direction of groundwater flow. Water samples that are collected at monitoring wells can be used to obtain information on the quality of groundwater.

Piezometric head is the sum of pressure and the elevation heads. The pressure head at any point in the aquifer intercepted by the screen is the height of the water surface in the well bore (or standpipe) above the point on the screen. The elevation head at a point is the elevation of the point above or below a datum. The gradient of the piezometric head determines the magnitude and direction of groundwater flow. By measuring the piezometric heads at many monitoring wells it is possible to draw the contour lines of equal piezometric head or potential. A flow net is obtained by constructing a set of flow lines that are orthogonal to the equipotential lines. The groundwater flow direction at a point is determined by drawing a tangent to the flow line passing through the point. By constructing the flow net so that the intersection of equipotential lines and the flow lines form curvilinear squares, it is possible to estimate the groundwater flow velocities and discharge in an aquifer. The monitoring well data can also be used to calibrate a groundwater model that then can be used to estimate and monitor groundwater flow.

Groundwater samples collected from monitoring wells can be analyzed in the laboratory to determine the chemical contents. Probes can be inserted into the monitoring wells for the *in situ* measurement of conductivity and pH.

Process Emissions Monitoring. Methods and instrumentation for the monitoring of industrial waste discharge streams are discussed in detail in the USEPA *Handbook for Sampling and Sample Preservation of Water and Wastewater.* ASTM publishes a frequently updated *Annual Book of ASTM Standards* that extensively details recommended sampling techniques.

The topic of sampling is complex and includes sampling timing and frequency; sample preservation (until analyses can be done); sample preprocessing (i.e., filtering to separate particulate and dissolved components); questions of whether waste streams are continuously monitored for particular environmental variables or whether discrete samples are taken; whether a single "grab" sample is taken or a series of samples in time or space are retrieved and composted into a single "representative" sample; whether there is a single, worker-sampled, grab sample of an automated, continuous sampling apparatus; whether there is retention of volatile compounds for later analysis. Additionally, there are issues of

laboratory analytical techniques, quality control, and quality assurance from the moment of sampling through the reporting of the data, acceptable detection limits, sample contamination, materials used to sample particular chemicals. The only way to account for all these variables is to know thoroughly the data-reporting requirements of the industrial discharge permits or agreements. Also, the objectives of the sampling and the compounds of concern will dictate the entire structure of the sampling regime. The authors will discuss, briefly, some of the above topics.

The ideal characteristics of a process waste discharge stream for monitoring are steady, uniform flow that is chemically homogeneous throughout the cross section. Under such conditions, several grab samples over the day may yield representative results. In many industries, the waste streams vary due to process switching, change of shifts of workers, and cleaning and preparation of process equipment and machinery. Several specified times a day a local beer brewery flushes process wastes from the floor using high-pressure hoses. Occasional grab sampling might miss completely this pulse of waste load to the treatment system. There are continuous sampling devices that take a grab sample at specified time periods, then rotate to another sample bottle in preparation for the next sample. There are more-sophisticated samplers that draw in a sample at time intervals proportional to stream discharge. Often, the type and frequency of discharge sampling is specified in the discharge permit. For best results, the sampling planner or coordinator should know the plant operation cycles in order to be able to characterize the time variation of discharges.

The material of which a sampling device and storage vessel is constructed must be matched to the materials sampled. Some organics require the use of tetrafluoroethylene (Teflon) tubing and glass storage vials that have been washed with an organic-free cleanser. Using the wrong tubing may result in the leaching of plasticizers into the sample. It is recommended that metal sampling devices *not* be used for collection of metal-laden discharge samples that require later analysis for metal ions. Samplers must be washed between samplings to avoid sample cross contamination.

Environmental Quality Monitoring. Most of the mechanics and issues of sampling discussed just above for monitoring discharges hold true for environmental sampling, i.e., after the discharge has entered a stream or lake. Detailed sampling plans must be prepared. Existing data is gathered (from company files; land use maps; USEPA; U.S. Geological Survey gauging and water quality station results; state departments of environment, resources, and health; other federal and local agencies). It helps to have site characteristics prior to the inception of the waste discharge.

The data collection for environmental monitoring often is governed by legal considerations concerning the validity and admissibility of the data. To this end, a quality assurance/quality control (QA/QC) plan must be prepared according to USEPA or state guidelines. These plans will specify the frequency of taking duplicate samples (taken at approximately the same time and place), split samples (one sample is split into two, separate containers), spiked samples (an additional injection of chemicals anticipated to be in the environmental sample are added to a duplicate sample, and the recovery of that known amount of spike is reported). A good QA/QC operation will ensure that the best effort is being made to

1. Obtain representative samples,
2. Use appropriate sampling methods,
3. Use appropriate analytical methods,
4. Ensure adequate records of chain of custody of samples, and
5. Develop a sound and acceptable database.

The USEPA can be contacted to send *Interim Guidelines and Specifications for Preparing Quality Assurance Project Plans* (or QAPPS — pronounced “kwaps”).

Other documents that may be required include

1. Sampling plans;
2. Site background information;
3. Planned sampling locations;

4. Planned sampling methodology and sample preservation (some chemicals decay on exposure to light, microbes may induce decomposition of organic compounds at ambient temperatures, chemically reduced compounds may oxidize on exposure to a head space of air, volatile compounds may escape from an insufficiently sealed vial, solid materials may dissolve, or dissolved materials may precipitate);
5. Health and safety plans;
6. Plans for sample-handling procedures and chain of custody of samples;
7. Request forms for sample analyses;
8. Frequency and format of data reports.

Air Quality Monitoring

Air quality monitoring has many of the same goals and general techniques as water quality monitoring. A stationary monitoring network should yield the following information:

1. Background concentration
2. Highest concentration levels
3. Representative levels in populated areas
4. Impact of local sources
5. Impact of remote sources
6. Relative impact of natural and anthropogenic sources

Spatial scales range from micro- (up to 100 m) to regional scales (tens to hundreds of kilometers). Site selection is a key part of network design because microclimates can affect readings and cause them to be nonrepresentative of the region.

Mobile monitoring is useful when sites not monitored with the fixed network need assessment. The key drawback of such systems is the sparsity of equipment suitable for reliable and durable mobile monitoring. Remote sensing offers a second alternative to stationary networks.

Quality assurance is a continuous concern of monitoring systems. Good instrumentation installation and maintenance practice with careful record keeping is essential.

Space precludes presentation of further details on air quality monitoring, but Boubel et al. (1994) has a thorough overview.

16.4 Regulations and Emission Standards

Water

Nevis Cook and Ronald R. Hewitt Cohen

The Clean Water Act (CWA) of 1972 and its amendments establish the framework of current U.S. water pollution control. This act is sometimes referred to as the Federal Water Pollution Control Act Amendments after the original 1965 act which the CWA amended. The objective of the CWA is to "restore and maintain the chemical, physical, and biological integrity of the nation's waters." The CWA regulates both nonpoint (runoff from farmlands, roads, and city streets) and point sources (discharges from pipes conveying pollutants to surface waters) of water pollution. The discussions in this chapter are limited to regulation of point sources of pollution to *surface* waters. Discharges to groundwaters are regulated under the Safe Drinking Water Act (SDWA) underground injection control (UIC) program and are beyond the scope of this chapter. The discussion is still quite broad, since all domestic and industrial wastewater discharges are considered point sources under the CWA.

Discharge Permits. The CWA prohibits point source discharges to surface waters unless a *discharge permit* is obtained. Permits typically are issued at the state level and are known as state pollution discharge elimination standards (SPDES) permits. If the state has not taken over this responsibility, national (NPDES) permits are issued by the USEPA. The permit for each point source prescribes allowable discharges in terms of amount and concentration of flows and pollutants. In some cases, an industrial facility may avoid obtaining its own permit by discharging into the local public sewer system. Control of these *indirect discharges* is accomplished by requiring publicly owned treatment works (POTWs) receiving such discharges to enter into formal *pretreatment agreements* with industrial users of the sewer. Such pretreatment agreements are intended to prevent industrial discharges to POTWs which would represent a hazard to the sewer system or its workers, interfere with treatment operations or sewage sludge disposal, or pass through the treatment process causing a violation of the POTW discharge permit.

Permitted Discharges. Prior to the CWA of 1972, point source water pollution control laws were established at the state level and permitted discharge of pollutants on a discharger-by-discharger and water-body-by-water-body basis. This approach worked poorly in most states, demonstrating the need for enforceable *federal minimum effluent standards*. For this reason, the CWA and subsequent amendments have established *technology-based* national minimum pretreatment and discharge standards. Technology-based standards reflect the current state of the art with respect to controlling specific pollutant discharges from specific pollution sources. Since technology-based standards represent the capabilities of typical "well-operated" facilities and are not issued on the basis of in-stream water quality, it is not surprising that in some cases compliance with national minimum standards does not adequately protect all potential uses of a water body. In such cases the uses of the receiving water body are said to be *water quality limited,* and *local discharge limits*, more stringent than national limits, may be imposed. In addition, pretreatment agreements reflect the capabilities of local treatment works with respect to pass through, destruction, or partitioning of nondegradable toxic metal pollutants into POTW sludges. Following the development of local water quality–based limits, these are compared with national minimum technology-based standards and the more restrictive criteria are imposed as discharge permit or pretreatment limitations. Further considerations with respect to setting permit values are presented below.

Industrial dischargers may obtain an industrial NPDES permit and directly discharge wastewater to a receiving water body. Alternatively, where municipal sewer service is available, industries may choose to negotiate a pretreatment agreement with the local sewer authority permitting discharge of industrial wastewater to the sewer system. Most industrial facilities employing either discharge option (direct or indirect) are subject to national technology-based minimum effluent standards. In addition, more-stringent standards may be imposed at the state or local level.

State permits containing discharge constraints more stringent than the national minimum standards are commonly derived from local water quality considerations. Ideally, local water quality–based

standards are developed from science-based use protection criteria, wastewater flow data, and actual receiving water characteristics (such as low-flow conditions for streams). Rational direct discharge permits to streams or rivers can be developed by application of the following procedure: convert use-based, in-stream quality criteria to standards according to local policy; obtain data on upstream water quality and low-flow conditions (conditions at which the waste assimilative capacity of a stream is at a minimum) and determine potential for dilution of the wastewater flow by the receiving water body; compute the allowable water quality–based industrial discharge on a pollutant-by-pollutant basis and compare with national minimum standards; finally, impose permit restrictions according to the most stringent of the two types of standards developed.

Development of water quality–based permits for direct discharges to lakes, estuaries, bays, or oceans follows a procedure similar to that outlined above. However, estimates of mixing patterns and the dilution potential of these water bodies are likely to be more difficult because of the complex flow patterns involved. Water quality–based pretreatment standards for discharges to sewers also follow the same general procedure. However, in this case, the application of use criteria, policy, and potential for in-stream dilution lead to a pollutant-by-pollutant permit for the *local* POTW. Based on this permit and the removal capabilities of the treatment facility, maximum allowable (usually daily) loading of pollutants arriving at the plant is determined on a pollutant-by-pollutant basis. Allowable loads are then allocated to the various municipal and industrial users of the sewer. Setting water quality–based pretreatment standards is obviously complicated by the presence of multiple sewer uses and users. Thus, POTWs receiving significant industrial discharges are required to set up formal programs to negotiate industrial pretreatment agreements and monitor compliance.

Air

Jan F. Kreider

The U.S. Clean Air Act Amendments of 1977 set forth two air quality standard types:

- Primary (protect health)
- Secondary (protect health and prevent other adverse impacts)

Table 16.4.1 summarizes the present primary and secondary standards in the U.S. The amendments also specified for certain geographical areas further standards to prevent significant deterioration (PSD areas). These are standards that are to be considered increments over baseline air quality but are more stringent than secondary or primary standards in most cases. Table 16.4.2 lists these PSD standards.

TABLE 16.4.1 U.S. Federal Primary and Secondary Ambient Air Quality Standards

				Concentration	
Pollutant	**Type of Standard**	**Averaging Time**	**Frequency Parameter**	**μg/m³**	**ppm**
Sulfur oxides (as sulfur dioxide)	Primary	24 hr	Annual maximum	365	0.14
		1 year	Arithmetic mean	80	0.03
	Secondary	3 hr	Annual maximum	1,300	0.5
Particulate matter >10 μm	Primary	24 hr	Annual maximum	150	—
		24 hr	Annual geometric mean	50	—
	Secondary	24 hr	Annual maximum	150	—
		24 hr	Annual geometric mean	50	—
Carbon monoxide	Primary and secondary	1 hr	Annual maximum	40,000	35.0
		8 hr	Annual maximum	10,000	9.0
Ozone	Primary and secondary	1 hr	Annual maximum	235	0.12
Nitrogen dioxide	Primary and secondary	1 year	Arithmetic mean	100	0.05
Lead	Primary and secondary	3 months	Arithmetic mean	1.5	—

TABLE 16.4.2 U.S. Federal PSD Concentration Increments

Pollutant	Increment ($\mu g\ m^3$)
Class I areas	
Particulate matter	
TSP, annual geometric mean	5
TSP, 24-hr maximum	10
Sulfur dioxide	
Annual arithmetic mean	2
24-hr maximum	5
3-hr maximum	25
Nitrogen dioxide	
Annual arithmetic mean	2.5
Class II areas	
Particulate matter	
TSP, annual geometric mean	19
TSP, 24-hr maximum	37
Sulfur dioxide	
Annual arithmetic mean	20
24-hr maximum	91
3-hr maximum	512
Nitrogen dioxide	
Annual arithmetic mean	25
Class III areas	
Particulate matter	
TSP, annual geometric mean	37
TSP, 24-hr maximum	75
Sulfur dioxide	
Annual arithmetic mean	40
24-hr maximum	182
3-hr maximum	700
Nitrogen dioxide	
Annual arithmetic mean	50

16.5 Mitigation of Water and Air Pollution

This section discusses the methods for abating or mitigating air and water pollution burdens on the environment. Because mechanical engineers are most concerned with air pollution control systems and civil engineers deal with water treatment systems, this section emphasizes the air pollution mitigation side.

Overview

Jan F. Kreider

There are several methods for controlling air- or waterborne pollution:

- Process change
- Fuel change
- Pollution removal and disposal
- Pollution prevention

Process change includes everything from modifications that reduce emissions to substitution of a less polluting one for a more polluting one. The latter could be classified as pollution prevention, described shortly. In many cases, for example, in the steel industry, it has proved most economical to replace completely old plants with totally new ones. Likewise, complete substitution has been widely adopted in the pulp and paper industry.

Fuel change as a control strategy is based on the fact that airborne pollutants often are related to the combustion aspects of a process. For example, coal-fired power plants emit SO_2 because coal contains sulfur. Substitution of natural gas for coal eliminates any sulfur in the fuel and, therefore, any oxides of sulfur in the stack gases.

Fuel changes must consider fuel supply, capital cost, and competition for clean fuels among many industries before an engineering design decision can be made. Life cycle methods that consider all parts of the life of a plant are necessary. For example, a nuclear power plant may produce very low emission electricity but after decommissioning may cause long-term waste disposal problems.

Pollution removal is necessary when process or fuel changes cannot provide adequate emission control. A physical, chemical, or electrical feature of the pollutant to be removed must differ from those types of characteristics of the carrying gas or liquid stream. An example is a baghouse or electrostatic precipitator for particulate emissions. Not only collection of pollutants but also disposal of the collected pollutant must be considered by the design engineer. A whole systems viewpoint is necessary for a successful pollution removal design.

Pollutant disposal is governed by different criteria depending on whether it is hazardous or not. Hazardous waste disposal is covered by the Resource Conservation and Recovery Act of 1976, which established the Office of Solid Waste with the USEPA. On the other hand, nonhazardous waste disposal is governed by the states. Table 16.5.1 categorizes the ultimate disposal methods for classes of hazardous wastes. Many of the methods apply for nonhazardous wastes as well.

Pollution prevention is the ultimate solution to abatement. This can be accomplished at the source by various technical means, depending on process specifics. This approach is currently the most commonly used term by the USEPA. The Pollution Prevention Act of 1990 stated these policies:

- Prevent or reduce pollution at the source whenever possible;
- Recycle to the environment pollution that cannot be prevented in a safe manner whenever possible;
- Pollution that can neither be prevented nor recycled should be treated in as environmentally safe a manner as possible;
- Disposal or other release into the environment should be used as only a last resort.

TABLE 16.5.1 Ultimate Waste Disposal Methods

Process	Purpose	Wastes	Problems (Remarks)
Cementation and vitrification	Fixation	Sludges	Expensive
	Immobilization	Liquids	
	Solidification		
Centrifugation	Dewatering	Sludges	—
	Consolidation	Liquids	
Filtration	Dewatering	Sludges	Expensive
	Volume reduction	Liquids	
Thickening (various methods)	Dewatering	Sludges	—
	Volume reduction	Liquids	
Chemical addition (polyelectrolytes)	Precipitation	Sludges	Can be used in conjunction with other processes
	Fixation	Liquids	
	Coagulation		
Submerged combustion	Dewatering	Liquids	Acceptable for aqueous organics
	Major Ultimate Disposal Methods		
Deep well injection	Partial removal from biosphere	Oil field brines; low toxicity, low-persistence wastes; refinery wastes	Monitoring difficulty; need for special geological formations; groundwater contamination
	Storage		
Incineration	Volume reduction	Most organics	If poor process control, unwanted emissions produced
	Toxicity destruction		Can produce NO_x, SO_x, halo acids
Recovery	Reuse	Metals	Sometimes energy prohibitive
		Solvents	
Landfill	Storage	Inert to radioactive	Volatilization
	Major Waste Disposal Methods		
Land application	Isolation		Leaching to groundwater
Land burial	Dispersal		Access to biota
Ocean disposal	Dispersal	Acids, bases	Contact with ocean ecosystem; containers unstable
	Dilution	Explosives	
	Neutralization	Chemical war agents	
	Isolation(?)	Radioactive wastes	
	Minor Disposal Methods		
Biological degradation	Reduction of concentration	Biodegradable organics	Most hazardous wastes do not now qualify
	Oxidation		
Chemical degradation (chlorination)	Conversion	Some persistent pesticides	—
	Oxidation		
Electrolytic processes	Oxidation	Organics	—
Long-term sealed storage	Isolation	Radioactive	How good are containers?
	Storage		
Salt deposit disposal	Isolation	Radioactive	Are salt deposits stable in terms of waste lifetimes?
	Storage		

Air Pollution Control

Jan F. Kreider

The most common method of meeting emission standards in industries that must control air pollution is by *pollution removal*. This section describes the most widely used approaches. Table 16.5.2 lists the important characteristics of control systems for airborne pollutants.

Dry particulate matter is removed by

TABLE 16.5.2 Key Characteristics of Pollution Control Devices and/or Systems

Factor Considered	Characterisitc of Concern
General	Collection efficiency
	Legal limitations such as best available technology
	Initial cost
	Lifetime and salvage value
	Operation and maintenance costs
	Power requirement
	Space requirements and weight
	Materials of construction
	Reliability
	Reputation of manufacturer and guarantees
	Ultimate disposal/use of pollutants
Carrier gas	Temperature
	Pressure
	Humidity
	Density
	Viscosity
	Dewpoint of all condensables
	Corrosiveness
	Inflammability
	Toxicity
Process	Gas flow rate and velocity
	Pollutant concentration
	Variability of gas and pollutant flow rates, temperature, etc.
	Allowable pressure drop
Pollutant (if gaseous)	Corrosiveness
	Inflammability
	Toxicity
	Reactivity
Pollutant (if particulate)	Size range and distribution
	Particle shape
	Agglomeration tendencies
	Corrosiveness
	Abrasiveness
	Hygroscopic tendencies
	Stickiness
	Inflammability
	Toxicity
	Electrical resistivity
	Reactivity

- Filters — baghouse, fixed beds, or mats
- Electrostatic precipitators — plate-type, tube-type; see Figure 16.5.1
- Inertial collectors — cyclones and baffles; see Figure 16.5.2
- Scrubbers — wet or dry

Table 16.5.3 lists the key characteristics of these technologies, where they are best applied, and annual operating, capital, and maintenance costs in $1994. Of course, a design is needed upon which a quotation for final cost studies can be based.

Liquid droplets and mists are controllable by

- Filters — more loosely knit than for dry filters
- Electrostatic precipitators — wetted wall type
- Inertial collectors — cyclones and baffles
- Venturi scrubbers

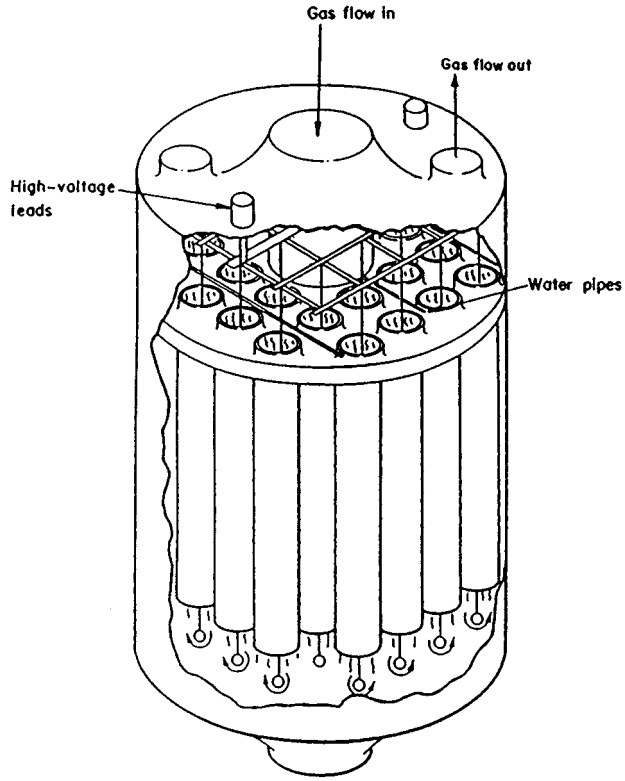


FIGURE 16.5.1 Wet-wall eledtrostatic precipitator with tubular collection electrodes. (From Oglesby, S., Jr., and Nichols, G. B., in *Air Pollution*, 3rd ed., Vol. IV, A. C. Stern, Ed., p.238, Academic Press, New York, 1977. With permission.)

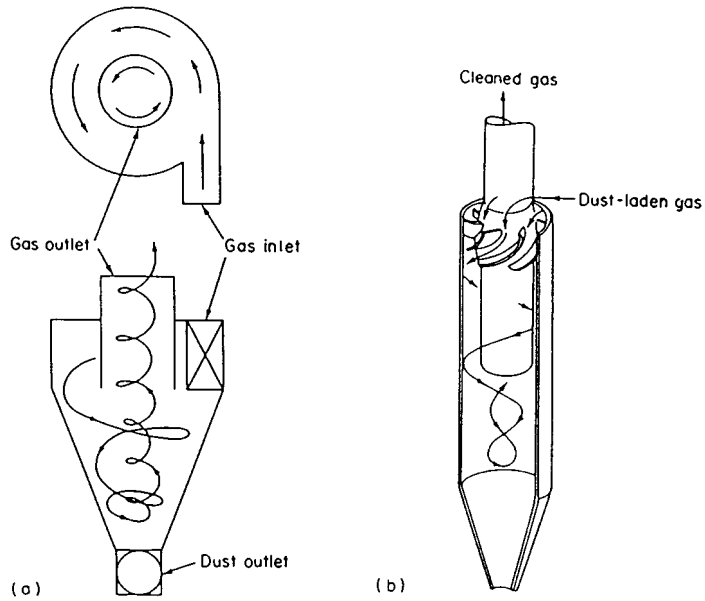


FIGURE 16.5.2 (a) Tangential inlet cyclone. (b) Axial inlet cyclone.

TABLE 16.5.3 Comparison of Particulate Removal Systems

Type of Collector	Particle Size Range (μm)	Removal Efficiency	Space Required	Max. Temp. (°C)	Pressure Drop (cm H_2O)	Annual cost (U.S. $/year/m^3)[a]
Baghouse	0.1–0.1	Fair	Large	80	10	28.00
(cotton bags)	1.0–10.0	Good	Large	80	10	28.00
	10.0–50.0	Excellent	Large	80	10	28.00
Baghouse	0.1–1.0	Fair	Large	120	12	34.00
(Dacron, nylon,	1.0–10.0	Good	Large	120	12	34.00
Orlon)	10.0–50.0	Excellent	Large	120	12	34.00
Baghouse (glass	0.1–1.0	Fair	Large	290	10	42.00
fiber)	1.0–10.0	Good	Large	290	10	42.00
	10.0–50.0	Good	Large	290	10	42.00
Baghouse (Teflon)	0.1–1.0	Fair	Large	260	20	46.00
	1.0–10.0	Good	Large	260	20	46.00
	10.0–50.0	Excellent	Large	260	20	46.00
Electrostatic	0.1–1.0	Excellent	Large	400	1	42.00
precipitator	1.0–10.0	Excellent	Large	400	1	42.00
	10.0–50.0	Good	Large	400	1	42.00
Standard cyclone	0.1–1.0	Poor	Large	400	5	14.00
	1.0–10.0	Poor	Large	400	5	14.00
	10.0–50.0	Good	Large	400	5	14.00
High-efficiency	0.1–1.0	Poor	Moderate	400	12	22.00
cyclone	1.0–10.0	Fair	Moderate	400	12	22.00
	10.0–50.0	Good	Moderate	400	12	22.00
Spray tower	0.1–1.0	Fair	Large	540	5	50.00
	1.0–10.0	Good	Large	540	5	50.00
	10.0–50.0	Good	Large	540	5	50.00
Impingement	0.1–1.0	Fair	Moderate	540	10	46.00
scrubber	1.0–10.0	Good	Moderate	540	10	46.00
	10.0–50.0	Good	Moderate	540	10	46.00
Venturi scrubber	0.1–1.0	Good	Small	540	88	112.00
	1.0–10.0	Excellent	Small	540	88	112.00
	10.0–50.0	Excellent	Small	540	88	112.00
Dry scrubber	0.1–1.0	Fair	Large	500	10	42.00
	1.0–10.0	Good	Large	500	10	42.00
	10.0–50.0	Good	Large	500	10	42.00

[a] Includes water and power cost, maintenance cost, operating cost, capital, and insurace costs (in $ 1994).

Finally, gaseous pollutants can be controlled using

- Absorption materials — wet scrubber, packed tower, or bubble tower; see Figure 16.5.3, for example;
- Adsorption materials — porous media such as activated charcoal, silica gel, and alumina;
- Condensers — widely used in the chemical process industries;
- Chemical conversion to contaminants — oxidize hydrocarbons to CO_2 and H_2O, for example.

Table 16.5.4 compares gaseous pollutant removal systems.

Cataloging of industry-specific designs is beyond the scope of this handbook because of space limitations. The reader is referred to Boubel et al. (1994) for details. One must make an inventory of emissions and then identify the most appropriate control methods based on operating and capital costs considerations. For example, Table 16.5.5 shows common emissions and control methods for the petrochemical industry. The energy, power generation, incineration, ferrous and nonferrous metallurgical, agricultural, and mineral/mining industries produce considerable emissions that must be controlled in accordance with federal regulations.

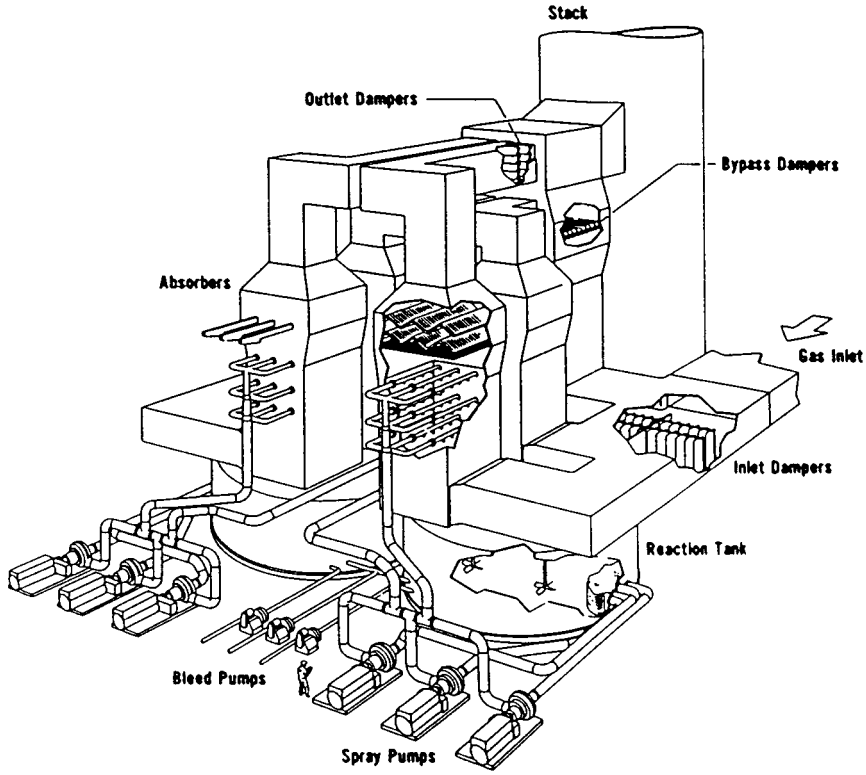


FIGURE 16.5.3 Cutaway drawing of a flue gas desulfurization spray tower absorber. (Courtesy of CE Power Systems, Combustion Engineering, Inc.)

TABLE 16.5.4 Comparison of Gaseous Pollutant Removal Systems

Type of Equipment	Pressure Drop (cm H_2O)	Installed Cost (U.S. \$/m³)	Annual Operating Cost (U.S. \$/m³)
Scrubber	10	9.80	14.00
Absorber	10	10.40	28.00
Condenser	2.5	28.00	7.00
Direct flame afterburner	1.2	8.20	8.40 + gas
Catalytic afterburner	2.5	11.60	28.00 + gas

Water Pollution Control

Nevis Cook and Ronald R. Hewitt Cohen

Control of waterborne pollutants is at present undertaken by nine distinct techniques:

- Biological oxidation
- Chemical oxidation
- Chemical reduction
- Conventional treatment
- Precipitation
- Air stripping
- Activated carbon

TABLE 16.5.5 Air Pollution Emissions and Controls: Petrochemical Processes

Petrochemical Process	Air Pollutant Emissions	Control Methods in Use
Ethylene oxide (most emissions from purge vents)	Ethane, ethylene, ethylene oxide	Catalytic afterburner
Formaldehyde (most emissions from exit gas stream of scrubber)	Formaldehyde, methanol, carbon monoxide, dimethyl ether	Wet scrubber for formaldehyde and methanol only; afterburner for organic vent gases
Phthalic anhydride (most emissions from off-gas from switch condensers)	Organic acids and anhydrides, sulfur dioxide, carbon monoxide, particulate matter	Venturi scrubber followed by cyclone separator and packed countercurrent scrubber
Acrylonitrile (most emissions from exit gas stream from product absorber)	Carbon monoxide, propylene, propane, hydrogen cyanide, acrylonitrile, acetonitrile NO_x from by-product incinerator	Thermal incinerators (gas-fired afterburners or catalytic afterburners)
		None
Carbon black (most emissions from exit gas stream from baghouse, some fugitive particulate)	Hydrogen, carbon monoxide, hydrogen sulfide, sulfur dioxide, methane, acetylene	Waste heat boiler or flare (no control for SO_2)
	Particulate matter (carbon black)	Baghouse
Ethylene dichloride (most emissions from exit gas stream of solvent scrubber)	Carbon monoxide, methane, ethylene, ethane, ethylene dichloride, aromatic solvent	None at present, but could use a waste heat boiler or afterburner, followed by a caustic scrubber for hydrochloric acid generated by combustion

- Ion exchange
- Reverse osmosis

Table 16.5.6 summarizes these approaches and their applications by contaminant type. The letter A in the table indicates the best available technology, whereas a B indicates an alternative that may apply to a subclass of industries. The reader is referred to the notes to the table for further details.

TABLE 16.5.6 General Effectiveness of Treatment Technology vs. Contaminant Type

Constituent	Biological Oxidation	Chemical Oxidation	Chemical Reduction	Conventional Treatment	Precipitation	Air Stripping	Activated Carbon	Ion-Exchange	Reverse Osmosis
BOD	A	—	—	—	—	—	—	—	—
COD	B	B	—	—	—	—	B	—	—
MOs[a]	—	A	—	A	A	—	—	—	I
Turb.[b]	—	—	—	A	B	—	—	I	I
TDS[c]	—	—	—	—	—	—	—	A	A
Ca, Mg	—	—	—	—	A	I	—	A	A
Fe, Mg	—	P	I	B	A	I	—	I	I
NH_3	A	B	—	—	—	B	—	B	B
NO_3^-	—	—	A	—	—	—	—	B	B
Me^{++}[d]	—	—	—	B	A	—	—	A	A
Cr	—	I	P	B	A	—	—	A	A
As, Se	—	P	I	B	A	—	—	A	A
CN	—	A	—	—	—	I	—	—	B
Phenols	B	A	—	—	—	—	B	—	—
SOC[e]	B	B	—	—	—	—	A	—	A
VOC[f]	B	B	—	—	—	A	B	—	B

A = treatment technology commonly applied to reduce contaminant to acceptable levels; perhaps best conventional technology. **B** = treatment technology that has been used to remove a particular contaminant, but might not be fully effective under all conditions. **I** = substance could interfere with efficient removal of other contaminants. **P** = pretreatment required if technology is to be used.

[a] MO = microbiological contaminants including pathogenic bacteria, protozoa, and viruses.

[b] Turb. includes fine colloidal matter and suspended solids, for true suspended solids use sedimentation.

[c] Total dissolved solids (TDS) includes removal of the highly soluble ions: Na^+, K^+, Cl^-, SO_4^{2-}.

[d] Includes the valence +2 transition metals: Cu^{++}, Ni^{++}, Pb^{++}, Zn^{++}.

[e] SOC = synthetic organic compounds, including pesticides.

[f] VOC = volatile organic compounds, including solvents.

16.6 Environmental Modeling

Air Pollution Dispersion Modeling

Paolo Zannetti

To understand air pollution, it is mandatory, at any scale, to simulate correctly the dispersion characteristics of the emissions. Therefore, the role of meteorology is essential. Pollutants are typically transported by two types of flows: an "ordered" flow, which is characterized by average wind speed and direction, and a semirandom, turbulent flow, which is characterized by wind fluctuations. All dispersion models aim at simulating these two components. As further discussed below, dispersion modeling techniques can be categorized into four general classes: (1) Gaussian models; (2) Eulerian grid models; (3) Lagrangian box models; and (4) Lagrangian particle models.

Gaussian Models

All Gaussian models assume that the concentration of pollutants maintains a Gaussian distribution in space. The Gaussian distribution, as illustrated in Figure 16.6.1, is a symmetrical bell-shaped distribution which is described at any given point x by two parameters: the location of the peak (in this case, the centerline of the plume indicated by the segmented line) and the standard deviation (in this case, the spread of the plume mass about its center). Therefore, the dilution rate of the plume is fully characterized by the two standard deviations, σ_y and σ_z, expressed as a function of the downwind distance, x.

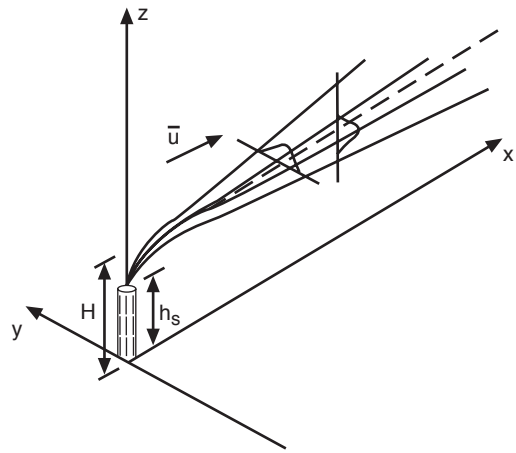


FIGURE 16.6.1 The Gaussian plume in a wind-oriented coordinate system (i.e., the wind is blowing toward the x axis). The plume is released from a source located at (0, 0, h_s) and possesses an initial buoyancy. Therefore, the plume behaves as if it were originated from (0, 0, H), where H is the effective emission height and $\delta h = H - h_s$ is the plume rise. The plume is advected by the average wind speed $\bar{\mathbf{u}}$ and expands in the horizontal and the vertical direction while maintaining a Gaussian distribution along both.

In mathematical notation, the Gaussian plume formula in Figure 16.6.1 can be written as

$$c = \frac{Q}{2\pi\sigma_y\sigma_z\bar{u}} \exp\left[-\frac{1}{2}\left(\frac{y}{\sigma_y}\right)^2\right] \exp\left[-\frac{1}{2}\left(\frac{H-z}{\sigma_z}\right)^2\right] \tag{16.6.1}$$

where c is the concentration computed at the receptor (x, y, z), Q is the emission rate, $\bar{\mathbf{u}}$ is the average horizontal wind speed, H is the effective emission height, and σ_y and σ_z are functions of the downwind

distance, x, with parameters that vary with the meteorological conditions (in fact, the stronger the turbulence intensity of the atmosphere, the larger the growth rate of σ_y and σ_z with x).

As can easily be seen, Equation (16.6.1) refers to a stationary state (i.e., c is not a function of time), uses meteorological parameters that must be considered homogeneous and stationary in the modeled area (i.e., between the source and the receptors at which concentrations are computed), and cannot work in calm conditions where the wind speed approaches zero (in general, the wind speed cannot be less than 1 m/sec when Equation (16.6.1) is applied). In spite of these limitations, the simplicity of the Gaussian approach, its relative ease of use, and, especially, the elevation of this methodology to the quantitative decision-controlling level in the United States (USEPA, 1978) have stimulated research aimed at removing some of the limitations of the Gaussian theory in modeling the real-world situations.

Equation (16.6.1) has been modified and expanded to incorporate, among others, the following factors: ground reflection, multiple reflections, hourly simulations, deposition and decay, chemical transformation, fumigation, complex terrain, gravitational settling, calm conditions, nonstationary and nonhomogenous conditions, and long-term simulations. We summarize below some of these improvements.

Reflection terms can be added to Equation (16.6.1) to account for partial or total reflection of concentration at the ground. Similarly, reflection can be added at the top of **the planetary boundary layer, PBL** (typically, about 500 to 1000 m above the ground). If both reflections are implemented, the plume is trapped inside the PBL. Equation (16.6.1) is generally applied for periods of 1 hr. This allows the incorporation of time-varying emission and meteorological parameters. Chemistry and decay can be incorporated by introducing exponential decay terms (for example, it can be assumed that an emission of primary gaseous SO_2 is transformed into particulate matter SO_4^{2-} at a rate of 1% per hour). Gravitational settling will affect a plume of primary particulate matter. In this case, the plume centerline can be tilted to account for the settling velocity of the particles, which is a function of both particle size and density.

In addition to the Gaussian plume model, Gaussian segment and puff models can be used (Zannetti, 1986a). These models break up the plume into independent elements (plume segments or puffs) whose initial features and time dynamics are a function of time-varying emission and meteorological conditions encountered by the plume elements. These techniques allow us to account properly for nonhomogeneous, nonstationary conditions. Gaussian puff models, in particular, have the additional advantage of being able to simulate calm or low-wind conditions.

Complex terrain conditions affect the plume dynamics, both the motion of the centerline trajectory and the growth of σ_y and σ_z. Finally, the Gaussian plume model equation can be rewritten in a way to simulate long-term concentration averages (e.g., annual averages) by incorporating the joint frequency of occurrence of a predetermined set of emission and meteorological conditions.

Other Models

Eulerian grid models (Lamb [from Longhetto], 1980) simulate pollutant diffusion by superimposing a grid over the computational domain and numerically solving a mass-balance equation (typically, a partial differential equation, PDE, or a set of PDEs) in each grid cell at each time step. This is also called numerical integration. In general, the smaller the grid and time intervals, the more accurate the numerical solution.

Difficulties may be encountered in simulating atmospheric diffusion with the K-theory. In particular, the application of the K-theory to simulate vertical dispersion during daytime, unstable meteorological conditions is highly questionable. To improve the simulation ability of Eulerian grid models, equations of high-order moments of concentration, wind, and temperature fluctuations can be solved simultaneously. This approach is called high-order closure and requires the definition of more-complex, nonlinear relationships between the turbulent fluxes and the concentration fields.

Lagrangian box models are mostly used to perform fast simulations of photochemical smog. These models define a set of "boxes" (e.g., a column of boxes as illustrated in Figure 16.6.2) which are advected horizontally according to the local wind speed and direction. Each box encounters emissions along its trajectory. These emissions inject new pollutants inside the box. A full set of chemical reactions inside

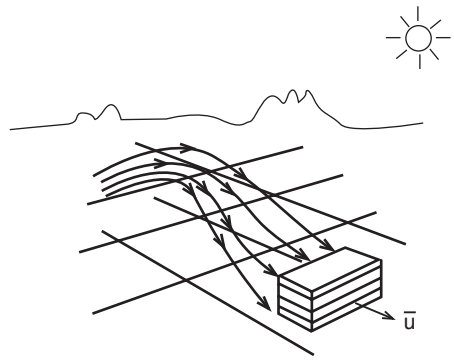


FIGURE 16.6.2 Lagrangian box modeling.

each box allows the simulation of the photochemical smog and the formation of secondary pollutants, such as ozone.

Lagrangian particle models provide a very interesting alternative method for simulating atmospheric diffusion. Particle motion can be produced by both deterministic velocities and semirandom pseudovelocities generated using Monte Carlo techniques. In the latter case, the trajectory of a single particle represents a realization from an infinite set of possible solutions, and, if enough particles are used, important characteristics can be inferred from the computation of particle **ensemble average** properties.

When using Lagrangian particle models to simulate air pollution, pollutant emissions (both gases and particulate matter) are represented by the injection of fictitious particles into the computational domain. Each particle represents a specified amount of pollution and is moved at each time step by a pseudovelocity that is time and space dependent. Zannetti (1990) discusses alternative modeling approaches in more detail than is possible here.

Atmospheric Chemistry

Atmospheric chemistry deals essentially with four major issues (Seigneur, 1987): (1) photochemical smog in sunny, urban areas; (2) **aerosol** chemistry; (3) acidic deposition; and (4) air toxics. Chemical reactions can be simulated in two ways: (1) with simple first-order terms (e.g., a decay term) and (2) with a full chemical reaction scheme.

First-order terms for simulating atmospheric chemistry can be easily incorporated into any of the models previously discussed. For example, a radioactive pollutant with a specified decay rate (or half-life) can be simulated by introducing the following multiplicative term in any concentration equation

$$\exp\left(-t/T\right) \tag{16.6.2}$$

where t is the travel time and T is the time scale of the decay (easily related to the half-life of the chemical species). Similarly, chemical transformation from a primary to a secondary pollutant (e.g., from gaseous SO_2 to sulfate particulate matter, SO_4^{2-}) can be accomplished by introducing two exponential terms— $[\exp(-t/T)]$ and $[1 - \exp(-t/T)]$ — having the effect of simultaneously decreasing the SO_2 concentration and increasing the SO_4^{2-} concentration as time increases. First-order schemes, though relatively simple, may use parameters that are space and time dependent; for example, the SO_2-to- SO_4^{2-} conversion rate may vary as a function of relative humidity and solar radiation.

Photochemical smog, which in the past only affected large cities at low latitudes, such as Los Angeles, has become today the most important and common air pollution problem in urban areas throughout the world. Overall, the photochemical smog reactions can be summarized as

$$\text{ROG} + \text{NO}_x + \text{sunlight} \rightarrow \text{O}_3, \text{PAN}, \text{HCHO}, \text{HNO}_3, \text{PM}, \ldots \tag{16.6.3}$$

where ROG are primary reactive organic gases and NO_x include primary NO and mostly secondary NO_2. This smog includes carbon monoxide (CO), ozone (O_3) formaldehyde (HCHO), peroxyacetyl nitrate (PAN), nitric acid (HNO_3), particulate matter (PM; especially secondary particles, such as nitrates and organic particles), and other products.

A full chemical reaction scheme is required to simulate complex phenomena, such as the photochemical smog described above, and can be incorporated today only inside Eulerian grids or Lagrangian box models. A typical reaction set, applied in each grid cell at each time step, can be written in terms of linear combinations:

$$\sum_{m=1}^{M} r_{nm} C_m \rightarrow \sum_{m=1}^{M} p_{nm} C_m \qquad n = 1, 2, \ldots, N \tag{16.6.4}$$

where M species participate in N reaction steps, c_m is the concentration of the mth species, and r_{nm} and p_{nm} are numerical constants. Each reaction rate is generally expressed as a product of the concentrations of the species involved, with a temperature-dependent rate constant.

The main difficulty in using Equation (16.6.4) to simulate the photochemical smog is the treatment of **organic compounds**. In fact, due to the very large number, organic species cannot all be included explicitly. Three different types of gas-phase chemical mechanisms are generally used: (1) surrogate mechanisms, which use the chemistry of one or two compounds in each class of organics to represent the chemistry of all species in that class; (2) lumped mechanisms, in which the grouping of chemical compounds is done on the bases of their similar structure and reactivity; and (3) the carbon bond approach, which splits each organic molecule into functional groups using the assumption that the reactivity of the molecule is dominated by the chemistry of each functional group. Each classification technique necessarily introduces a simplification and, therefore, a potential simulation error.

Some key reactions involve the **photolysis** of such species as NO_2, HCHO, and nitrous acid (HONO). These one-specie reactions require the calculation of the photolysis rate constant which is a function of, among other things, of solar elevation and temperature.

Aerosol chemistry is particularly difficult to simulate and computationally expensive. However, inclusion of aerosol dynamics within air quality models is of primary importance because of the health effects associated with fine particles in the atmosphere, visibility impairment, and the acid deposition problem. Simple first-order reaction terms can be used to simulate the transformation of SO_2 into sulfates and NO_x into nitrates. These terms can be included in any model. However, a comprehensive simulation of aerosol processes can only be performed within an Eulerian grid or a Lagrangian box model and must include the following fundamental equation of aerosol dynamics (Milford and Russell, 1993) which describes aerosol transport, growth, coagulation, and sedimentation

$$\frac{\delta n}{\delta t} + \nabla \cdot \bar{\mathbf{u}} n + \frac{\delta I}{\delta \nu} = \frac{1}{2}\int_0^{\nu} \beta(\bar{\nu}, \nu - \bar{\nu}) n(\bar{\nu}) n(\nu - \bar{\nu})\, d\bar{\nu} - \int_0^{\infty} \beta(\bar{\nu}, \nu) n(\bar{\nu}) n(\nu)\, d\bar{\nu} - \nabla \cdot \mathbf{C} n \tag{16.6.5}$$

where n is the particle size distribution function, $\bar{\mathbf{u}}$ is the wind velocity, I is the droplet current that describes particle growth and nucleation due to gas-to-particle conversion, ν is the particle volume, β is the rate of particle coagulation, and $\mathbf{C}$ is the sedimentation velocity.

The simulation of heterogeneous and aqueous-phase chemistry is of key importance for regional-scale acid deposition and **stratospheric** ozone models, but is usually neglected in urban photochemical applications where the main goal is the simulation of **tropospheric** ozone.

Deposition

Chemical species are removed from the atmosphere by two mechanisms: reaction and deposition. While chemical reactions may produce new pollutants, deposition is the real process in which the atmosphere cleans itself. Some pollutants are highly reactive and, consequently, have short lifetimes. For example,

ozone has a typical lifetime of 2 min and, therefore, its concentration will drop unless it is continuously regenerated. Other pollutants have longer lifetimes. For example, SO_2 has a typical lifetime of 13 days. Therefore, under certain circumstances, it can easily accumulate during multiday episodes (e.g., the "London" smog). Finally, there are pollutants with very large lifetimes. Because of their low reactivity, they do not cause adverse effects on human health but, nevertheless, can diffuse on a global scale and affect the thermal balance of the Earth. Methane and carbon dioxide are a good example, both with a typical lifetime of 7 years.

Deposition terms can be introduced in any model discussed above. For example, dry deposition can be described by the following formula

$$F_i = V_d c_i \tag{16.6.6}$$

where F_i is the flux of a species i to the ground, c_i is the concentration of the species i at some reference height (e.g., 1 m), and V_d is the deposition velocity. The term V_d has been measured under various meteorological conditions and for a number of surface types (Wesley et al., 1985). Therefore, the calculation of F_i is straightforward.

Wet deposition (i.e., precipitation scavenging) depends upon the intensity and size of raindrops. Fog and cloud droplets can also absorb gases, capture particles, and accelerate chemical reactions. Wet deposition is quantified by computing the wet flux of pollution to the surface. This calculation requires the estimate of the washout coefficient, which can be inferred (Scott, 1982) as a function of storm type and precipitation amounts.

Because of dry and wet deposition, acidic components such as sulfuric acid particles, particulate nitrate, and nitric acid gas are transferred from the atmosphere to the Earth. Areas which are tens and hundreds of kilometers downwind of large SO_2 and NO_x sources (e.g., power plants and smelters) suffer the greatest impact.

Statistical Models

Statistical models are often used in air pollution studies. They include frequency distribution studies, time series analysis, Kalman filters, receptor-modeling techniques, and others. A general distinction between statistical and deterministic approaches is that air pollution deterministic models initiate their calculations at the pollution sources and aim at the establishment of cause/effect relationships, while statistical models are characterized by their direct use of air quality measurements to infer semiempirical relationships. Although very useful, especially for real-time short-term forecasting, statistical models are generally unable to provide cause/effect relationships, with the exception of receptor modeling.

The basic concept of the receptor-modeling approach is the apportionment of the contribution of each source, or group of sources, to the measured concentrations without reconstructing the dispersion pattern and trajectory of the pollutants. Typically, receptor models start with *observed* ambient aerosol concentrations measured at a receptor and seek to apportion the concentrations among several source types (e.g., industrial, transportation, soil, etc.), based on the known chemical composition (i.e., the chemical fractions) of source and receptor materials.

In mathematical notation, the concentration c_{ik} of the species i in the kth sample at a certain monitoring station can be written as

$$c_{ik} = \sum_{j=1}^{p} a_{ij} D_{jk} E_{jk} \tag{16.6.7}$$

where p sources (or groups of sources) are assumed to contribute to c_{ik}, a_{ij} is the fractional amount of the component i in the emission from the jth source, D_{jk} is the atmospheric dispersion term, and E_{jk} is the emission rate (i.e., $D_{jk}E_{jk} = S_{jk}$ is the total contribution of the source j to the kth sample at the receptor location). Dispersion models assume a_{ij}, D_{jk}, and E_{jk} to be known (or obtainable from emission and

meteorological data) and estimate the output c_{ik}. For receptor models, the concentrations c_{ik} and source "profiles" a_{ij} are measured instead, and the $D_{jk}E_{jk}$ products are computed as a model result.

Ground Water Pollution Modeling

Tissa Illangasekare

Saturated Groundwater Flow

The description of the exact movement of fluid particles in a porous medium is difficult (or impossible) as it is not practical to define exactly the flow domain that is described by the geometry of the internal solid surfaces. The problem can be treated at the molecular level, microscopic level, or macroscopic level.

The treatment of the behavior of a system of molecules using theories of classical fluid mechanics is extremely difficult because of the large number of molecules involved and the difficulties in identifying all forces and defining the exact pore geometry. Instead of treating individual molecules, the statistical properties of a very large number of molecules may be inferred from laws governing the motion of individual molecules. Still, this approach will also have similar limitations with respect to the need to define the exact pore geometries. A coarser treatment at the microscopic level where fluid is treated as a continuum is feasible, but most applications in groundwater flow may not require this level of refinement. This approach will also require the accurate definition of the pore geometry.

A coarse level of averaging at the macroscopic level that is referred to as the *representative elementary volume (REV)* is used in most applications in porous media flow analysis. Porous medium is defined as a portion of a space occupied partly by a solid phase (solid matrix) and partly by voids. The voids in general are occupied by one (single phase) or more fluid phases (multiphase). This level of treatment assumes that the solid phase is distributed throughout the problem domain and it should be possible to define an REV such that no matter where we place it within the porous media domain, it will contain both solids and voids.

At the macroscopic scale, the rate at which the water flow in a soil is quantified using a variable that is referred to as the *Darcy velocity* or *specific discharge*. This variable, which has the dimensions of velocity, is defined as the discharge per unit gross area of soil that includes both the pore space and the grains in a flow section. For incompressible fluids a relationship that is referred to as *Darcy's law* expresses the Darcy velocity in terms of a parameter referred to as hydraulic conductivity, K, and the gradient of the piezometric head, h. Darcy's law for saturated flow (a single fluid filling the pore space) in soils is given by

$$\mathbf{q} = -K\nabla h \tag{16.6.8}$$

where $\mathbf{q}$ is the specific discharge or Darcy velocity (L/T) and h is the piezometric head (L).

In groundwater flow, as the velocities are generally very small, the velocity head is neglected and the driving head becomes the sum of the elevation and the pressure heads. The piezometric head is defined as

$$h = h_z + h_p = z + \frac{p}{\rho g} \tag{16.6.9}$$

where ρ is the density of water.

In anisotropic aquifers where the hydraulic conductivity changes with flow direction, $\mathbf{K}$ is a second-rank tensor given as

$$[K] = \begin{bmatrix} K_{xx} & K_{xy} & K_{xz} \\ K_{yx} & K_{yy} & K_{yz} \\ K_{zx} & K_{zy} & K_{zz} \end{bmatrix} \tag{16.6.10}$$

It can be shown that it is always possible to find three mutually orthogonal directions in space such that

$$K_{ij} \neq 0 \quad \text{for all} \quad i = j$$

$$K_{ij} = 0 \quad \text{for all} \quad i \neq j$$

These directions in space are called the principal directions of the anisotropic porous medium. When principal directions are parallel to the axes of the coordinate system, the tensor reduces to a hydraulic conductivity vector given as,

$$[K] = \begin{bmatrix} K_{xx} & 0 & 0 \\ 0 & K_{yy} & 0 \\ 0 & 0 & K_{zz} \end{bmatrix} \tag{16.6.11}$$

For conservation of mass,

$$-\nabla \cdot \rho \mathbf{q} = \frac{\partial n \rho}{\partial t} \tag{16.6.12}$$

For homogeneous, incompressible fluid and a nondeformable porous medium,

$$\nabla \cdot \mathbf{q} = 0 \tag{16.6.13}$$

Introducing head, the mass conservation equation can be written as

$$\nabla \mathbf{q} + S_s \frac{\partial h}{\partial t} = 0 \tag{16.6.14}$$

where S_s is the specific storage, defined as the volume of water added to storage, per unit volume of porous medium, per unit rise in piezometric head. This is given as

$$S_s = \rho g(\alpha + n\beta) \tag{16.6.15}$$

where β is the compressibility of water and α is the compressibility of the soil matrix.

Combining Darcy's law and the equation of mass conservation, the general equation of saturated groundwater flow is obtained as

$$\nabla \cdot K \nabla h = S_s \frac{\partial h}{\partial t} \tag{16.6.16}$$

The initial and boundary value problem obtained by combining the above second-order partial differential equation with the initial head in the aquifer and the head or flux conditions at the aquifer boundary is solved to obtain the unknown head in the aquifer. The head distribution can then be used with Darcy's law to obtain the groundwater flow velocity in the aquifer.

Solute Transport in Groundwater

Transport of dissolved chemicals in aquifers is generally considered to be the result of two processes, namely, advection and dispersion. Advection is the process by which the solute gets transported due to the average motion of water through the intergranular pore spaces of the porous medium. In the mathematical representation of this process in the REV, a macroscopically average velocity that is referred

to as the *average linear pore velocity* is used. An approximate value for this average pore water velocity for granular material based on macroscopic variables is given as

$$v = \frac{q}{n} \tag{16.6.17}$$

where n is the effective porosity. The average liner pore water velocity is also referred to as the *average solution velocity.*

The advective solute transport is given by

$$\mathbf{J}_a = Cq \tag{16.6.18}$$

where $\mathbf{J}_a$ is the vector of solute mass flux (mass per unit time per unit area, $[MT^{-1}L^{-2}]$) and C is the mass concentration of solute per unit volume of the solution (ML^{-3}).

Dispersion is the result of two processes that occur in the pore scale, namely, molecular diffusion and mechanical (or hydrodynamic) mixing. Due to molecular diffusion the solute will move from the high-concentration to low-concentration regions in the fluid phase. Fick's first law modified to account for the presence of the solid phase is used to represent the solute flux as a function of the concentration gradient by

$$J_d = -nD_d \cdot \nabla C \tag{16.6.19}$$

where $\boldsymbol{J_d}$ is diffusive flux, $[MT^{-1}L^{-2}]$, n is the porosity, $\boldsymbol{D_d}$ $[L^2T^{-1}]$, a second-rank tensor is the effective diffusion coefficient of the solute in the porous medium and C is the solute concentration, $[M/L^{-3}]$.

The component of dispersion due to mechanical mixing is the result of velocity variation at the microscopic scale. These velocity variations are the result of three basic mechanisms that occur within the pores: (1) viscous shear forces that produce velocity gradients across flow channels, (2) pores size variations produce pore channels with different sizes transmitting water at different pore velocities, and (3) the changing flow directions due to the tortuosity of the flow channels. The combined effect of these variations results in the solute being mixed at the macroscopic scale and producing mass flux along decreasing concentration gradients. As this process is analogous to the diffusion in the microscopic scale, an equation similar to the Fick's first law is used to describe mass flux due to mechanical mixing. This analogous equation is given as

$$J_m = -nD_m \cdot \nabla C \tag{16.6.20}$$

where J_m is the flux due to mechanical mixing, $[MT^{-1}L^{-2}]$ and D_m is the coefficient of hydrodynamic (mechanical) dispersion.

The total flux due to molecular diffusion and hydrodynamic dispersion is given as

$$J = J_d + J_m \tag{16.6.21}$$

Substituting for J_d and J_m, we have

$$J_m = -n\left(D_d + D_m\right) \cdot \nabla C \tag{16.6.22}$$

Define the dispersion coefficient D as

$$D = D_d + D_m \tag{16.6.23}$$

Equation (16.6.24) reduces to

$$J = -n(D)\cdot\nabla C \tag{16.6.24}$$

In a three-dimensional system the dispersion coefficient is a second-order tensor that takes the form,

$$[D] = \begin{bmatrix} D_{xx} & D_{xy} & D_{xz} \\ D_{yx} & D_{yy} & D_{yz} \\ D_{zx} & D_{zy} & D_{zz} \end{bmatrix} \tag{16.6.25}$$

By orienting the x′ axes of the local coordinates at the aquifer point along the direction of groundwater flow (longitudinal direction), the dispersion coefficient tensor can be reduced to

$$[D] = \begin{bmatrix} D_{x'x'} & 0 & 0 \\ 0 & D_{y'y'} & 0 \\ 0 & 0 & D_{z'z'} \end{bmatrix} \tag{16.6.26}$$

The local axes x', y', and z' are the principal axes of dispersion and $D_{x'x'}$, $D_{y'y'}$, and $D_{z'z'}$ are the principal values of the coefficient of dispersion. These coefficients can be expressed in terms of the longitudinal average linear pore velocity as

$$D_{x'x'} = D_d + \alpha_L \bar{v}, \quad \text{etc.} \tag{16.6.27}$$

where α_L and α_T are longitudinal and lateral dispersion coefficients, respectively. The dispersion coefficients that have dimensions of length can be viewed as characteristic lengths of the porous medium at the pore scale. However, in real applications these values are much larger and have been found to depend on the size of the plume (scale dependent).

Combining advective flux given by Equation (16.6.18) and dispersive flux given by Equation (16.6.24) and applying the principal of mass conservation for the solute, one obtains

$$\frac{\partial(nC)}{\partial t} = \nabla\cdot(nD\cdot\nabla C - Cq) \tag{16.6.28}$$

By substituting the approximation for Darcy's velocity in Equation (16.6.28), the governing equation for solute transport in saturated porous media is obtained as

$$\frac{\partial C}{\partial t} = D\cdot\nabla C - C\bar{v} \tag{16.6.29}$$

The above equation is referred to as the *advection–dispersion equation*. The initial and boundary value problem obtained by combining the above second-order PDE with the initial concentration distribution in the aquifer and the concentration and mass flux at the aquifer boundary is solved to obtain the time and space distribution of the solute concentration in contaminant plumes. It should be noted that to solve the advection–dispersion equation it is necessary to first solve for the groundwater velocities using the groundwater flow equation.

Surface Water Pollution Transport Modeling

Ronald R. Hewitt Cohen and Nevis Cook

The study of water quality modeling bloomed in the late 1960s and through the 1970s. Administrators and bureaucrats with minimal mathematical and science backgrounds were greatly impressed by presentations of pages of PDEs, and model outputs were often treated as absolute truth. As the field and model users matured, it was recognized that a model is just a group of hypotheses about the way the modeler believes the world works, all put in mathematical terms. The limitations, shortcomings, and difficulties with environmental models are well recognized and accepted. Models are now used as *tools* for decision making and planning.

An industrial facility may want to assess how reducing or increasing the mass or concentration of some pollutant in a discharge will impact the receiving waters. It may be that a dramatic, negative impact might be predicted. Reducing the level of pollutant in the discharge may result in little to no improvement to water quality. Obviously, the results of the modeling effort will dictate the level of effort and cost going toward the treatment of the pollutant.

The same facility may be instructed by the USEPA or the state to control discharges such that water quality criteria in the receiving waters are met. Good models can be used to address the question "what are the implications to the receiving waters if various facility process modifications are applied." Thus, a decision could be made as to the process modifications to be focused upon to match the criteria or standards.

It is not anticipated that every industrial facility has an individual with the capabilities to construct a water quality transport model. There are many models available to run on microcomputers and can be obtained through the USEPA.

Impact Pathway Methodology

Ari Rabl and Peter S. Curtiss

A step beyond conventional dispersion modeling includes the physical and economic impacts of air pollution. Rational management of the environment requires an assessment of the damage caused by pollution. The logically correct way to analyze environmental impacts is the impact pathway methodology whose principal steps are the following:

- Specification of the relevant technologies and the environmental burdens they impose (e.g., kilograms per second of particulates emitted by the plant);
- Calculation of increased pollutant concentration in all affected regions (e.g., micrograms per cubic meter of particulates, using models of atmospheric dispersion and chemistry);
- Calculation of physical impacts (e.g., number of cases of asthma due to these particulates, using a dose–response function);
- In some cases a fourth step may be called for: the economic valuation of these impacts (e.g., multiplication by the cost of a case of asthma).

The numbers are summed over all receptors (population, crops, buildings, …) that are affected by this pollutant. Formally, the procedure can be represented as an equation for the incremental damage D due to an incremental quantity Q of a pollutant emitted by the plant

$$D = \sum_i f_{dr,i}\left(f_{\text{disp}-i}(Q)\right) \tag{16.6.30}$$

where $f_{\text{disp}-i}(Q) = c$ = increase in pollutant concentration for receptor i and $f_{\text{dr},i}(c)$ = dose–response function for receptor i; the summation index i runs over all receptors (people, crops, buildings, etc.) of concern.

Which receptors are of concern depends on the circumstances. One can distinguish three kinds of situation:

1. Episodic values (typically for litigation after pollution episodes);
2. Peak values (typically for obtaining a permit for a new plant, by showing that impacts are below a damage threshold or regulatory limit);
3. Expectation values (typically for policy applications such as setting of regulations, by showing that average impacts are acceptable).

For the first two of these the summation will typically be over a limited set of receptors, for instance, the residents in a town. For the third application one will usually want to know the total damage, and the sum should cover all receptors that make a significant contribution to the total.

The notation in Equation (16.6.30) allows the possibility that the impact may be different for different individual receptors. This equation expresses the damage in functional form; hence, this methodology is also known under the name *damage function*. Of course, while this methodology is logically correct, the practical implementation may not always be feasible for lack of appropriate data or models.

The dose–response function

$$Y = f_{dr}(X) \tag{16.6.31}$$

relates the quantity X of a pollutant that affects a receptor (e.g., population) to the physical impact Y on this receptor (e.g., incremental number of deaths). In the narrow sense of the term, X should be the dose actually absorbed by a receptor. But often one uses, as we do in the present section, the term dose–response function in the sense of exposure–response function where X represents the concentration of a pollutant in the ambient air; in that case $f_{dr}(X)$ accounts implicitly for the absorption of the pollutant from the air into the body. Dose–response functions for the classic air pollutants (NO_x, SO_x, O_3, and particulates) are typically of that kind. One can even define aggregated dose–response functions that include more-complicated pathways, for instance, dioxins passing through the food chain, if one interprets the dose–response function to include the aggregated effects of the pathways from a point at the Earth's surface to all final receptors. In the next sections we take a closer look at the major steps of the methodology.

The Source Term

The first step of the impact pathway analysis is relatively straightforward. One identifies the site and circumstances of a pollution source, e.g., the tons of NO per $kWhr_e$ emitted by particular power plant. For the major air pollutants (CO_2, CO, NO, SO_2, VOCs, particulate matter) the emission rates for a given technology are quite well known. For the example of power plants the rate of CO_2 emission is especially well determined. Emissions of CO, NO, SO_2, VOCs, and particulate matter are somewhat less certain, and they can vary with operating conditions. NO emissions, for instance, are likely to increase above the manufacturer's specifications if a selective catalytic reduction unit is not well maintained. There are different grades of oil and coal, and their sulfur content can differ by an order of magnitude; obviously, the emissions of SO_2 depend on the quality of the fuel that will be used. Usually, there are strict regulations that enforce an upper limit on the emissions; due to cost constraints power plants are unlikely to operate significantly below these limits.

The situation is less clear with regard to trace pollutants such as lead and mercury, since their content in different grades of coal can vary by much more than an order of magnitude. Furthermore, some of these pollutants are emitted in such small concentrations that their measurement is difficult. The dirtier the fuel, the greater the uncertainty of emissions. Especially with waste incineration, there has been concern over trace pollutants that are emitted into the air.

Probably the most uncertain emissions are emissions from the disposal and storage of wastes, because they depend on events in the future. Solid waste from coal-fired boilers could be dumped into a simple hole in the ground or it could be placed into an engineered landfill with watertight liners; the possible

impacts will be totally different. There may or may not be a breach of containment, depending on the quality of construction and management and on natural events such as floods or earthquakes. The main risk from a landfill is the leaching of toxic minerals into groundwater; such risk can be kept negligible by proper construction and management.

Transport Modes

Pollutants can be emitted to air, water, or soil. The majority of pollutants are first emitted into the air, even if they later pass into the water or the soil. Therefore, most of this section focuses on atmospheric dispersion. Transport in the soil is difficult to model because it can involve complex processes that depend on the physical and chemical properties of the soil at each site. Furthermore, for new installations, such as new landfills, the emissions into the soil are not known in advance; they depend on the integrity of the containment structure over the indefinite future.

Transport by surface water, i.e., rivers, lakes, and the sea, is relatively simple to analyze if fine geographical resolution is not required. Thus, one can divide these bodies of water into a reasonably small number of compartments that are treated as uniformly mixed. For example, a river may be divided into ten sections. A differential equation with empirical coefficients relates the concentration in a section under consideration to the concentration in the section immediately upstream and to the emission into this section. Sedimentation, removal, and decay processes are included.

Similarly, for the dispersion into marine waters one uses a compartment model where each compartment communicates with one or several neighbors, and the volumes and flow rates are known. For instance, in a model used for the analysis of nuclear power plants (EC, 1995c), the European seas have been divided into 34 compartments.

For dispersion in the atmosphere, in general both physical and chemical processes need to be considered (Seinfeld, 1986; Zannetti, 1990). Some pollutants, e.g., CO_2, CH_4, and ^{133}Xe, are sufficiently inert chemically that only the physical transport needs to be analyzed. Some are moderately reactive and their chemical transformation needs to be taken into account. SO_2, for instance, leads to the formation of SO_3, H_2SO_4 as well as sulfates (the latter from the interaction with NH_3 emitted by, among others, agricultural activities); this can have significant implications for impact analysis on a regional and global scale.* Ozone is a secondary pollutant, formed by the combination of NO_x, VOC, and light, and the chemistry is extremely complex.

Even though the modeling of the physical transport of pollutants is difficult, it is far simpler than weather modeling. The reason is that pollutants can be considered a small admixture, passively transported by the currents of the surrounding medium. Such transport is linear: the increase in concentration at a receptor site is proportional to the emission (the only exception arises from secondary pollutants such as ozone whose formation depends on other variables, coupled through nonlinear phenomena).

Furthermore, for most policy applications one needs only expectation values of environmental impacts. While it is well known that chaotic phenomena in the atmosphere render the prediction of the weather impossible beyond a short time, this does not prevent the prediction of expectation values. The climate is much more certain than the weather. For expectation values of air pollution damage it suffices to know the average motion of the surrounding medium from past observations, by contrast to weather modeling where that very motion needs to be predicted in detail.

Transport in Air

A simple model for atmospheric dispersion is the Gaussian plume, discussed earlier. According to this model the concentration of a pollutant is described by the product of two Gaussian distributions, one for the spread in the vertical direction and one for the spread in the horizontal direction perpendicular to the prevailing wind direction. The plume width parameters are based on empirical correlations and take into account the relevant meteorological conditions.

* For example, ammonia sulfate aerosols can reduce the impact of global warming (Charlson and Wigley, 1994).

The Gaussian plume is considered adequate for the short range, up to tens of kilometers from the source, even for episodic events (Zannetti, 1990). The use of this model at distances beyond 100 km is generally not recommended, although it is acceptable for the prediction of the average values if correction terms are included for reflection at the surface and at the PBL of the Earth, and if the depletion mechanisms (deposition, chemical transformation, radioactive decay) are correctly accounted for. As an example of dispersion software based on a Gaussian plume, one can cite the ISC model of the USEPA (Wackter and Foster, 1987).

For regional modeling most analysts prefer to rely on more-detailed computer simulations, for example the Harwell Trajectory Model (1993) or the EMEP model of the Norwegian Meteorological Service (Barrett, 1992; Sandnes, 1993; Iversen, 1993). The latter model is used for the official allocation of acid rain budgets among the countries of Europe.

A crucial question concerns the geographic range over which the analysis needs to be extended in order to capture most of the impacts. This involves a balance among the rates of emission, of dispersion, and of removal of a pollutant. A look at the results of long-range transport models for SO_2 and NO, for instance, those calculated by EMEP (Sandnes, 1993), shows that these pollutants are transported over hundreds, even thousands of kilometers. This is illustrated in Figure 16.6.3 using the EMEP data for a source at Nantes, assuming uniform receptor density and a linear dose–response function. The range of the analysis must be extended to over 1000 km if one wants to capture 80 to 90% of the total impact. The same holds for any air pollutants with comparable removal rate, as has been confirmed explicitly for radionuclides by Kelly and Jones (1985).

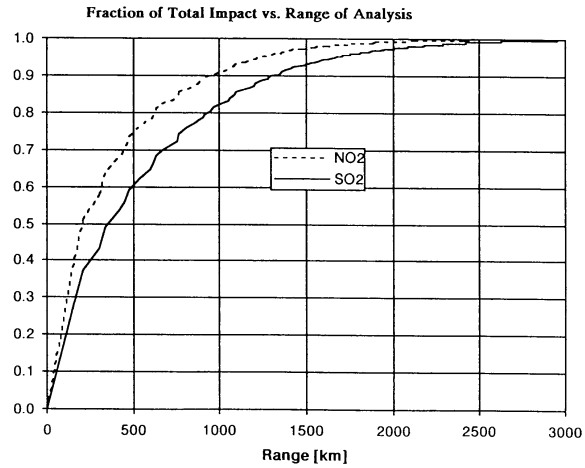


FIGURE 16.6.3 Fraction of total impact vs. range of analysis, for uniform receptor density and linear dose–response function, based on EMEP data (Barrett, 1994). Wiggles are due to discretization.

Secondary Pollutants

Many pollutants are transformed into secondary pollutants by chemical reactions in the atmosphere. For example, the reactions shown in Figure 16.6.4 create acid rain (wet deposition of H_2SO_4) and ammonium sulfate particulates from SO_2.

Another important secondary pollutant is ozone. It is formed when several chemical reactions take place in sequence. The only reaction that forms ozone directly is

$$O + O_2 + M \rightarrow O_3 + M$$

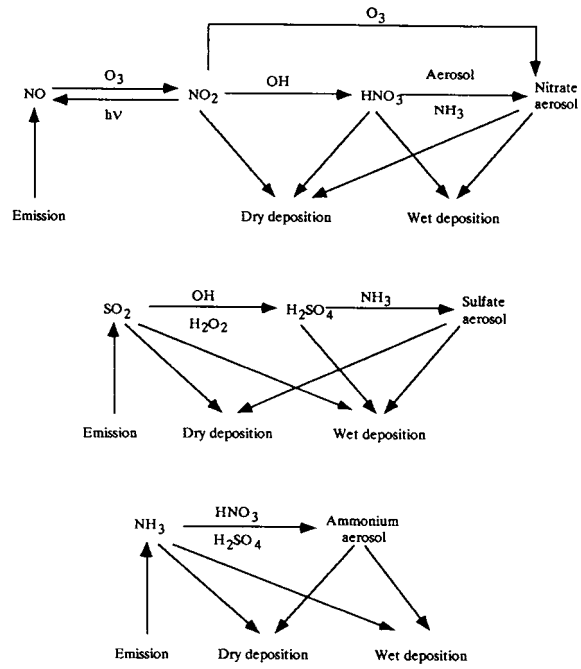


FIGURE 16.6.4 Chemical reactions included in Harwell Trajectory Model. (From EC, 1995c.)

where M is a molecule such as N_2 or O_2 whose participation is necessary to conserve energy and momentum. The oxygen atom involved in the formation of ozone is derived from photolysis of NO_2 under the action of sunlight (indicated by $h\nu$)

$$NO_2 + h\nu \rightarrow O + NO$$

The presence of VOCs is necessary to prevent the ozone formed from being immediately consumed by NO to produce NO_2 in the following reaction:

$$NO + O_3 \rightarrow O_2 + NO_2$$

VOCs enable the transformation of NO into NO_2 without consuming ozone. Finally, note also that NO_2 plays a double role, since, while being necessary to form ozone, it consumes the radicals needed by VOCs to transform NO into NO_2. In fact, an equilibrium is created between these reactions. The concentration of ozone therefore is very dependent on changes in the concentrations of other products, and, due to the complexity of the phenomena, it is observed for example that if VOCs are low (as in the case of an electricity power plant plume), the increase in NO may reduce O_3. Figure 16.6.5 shows the influence of the concentrations of nitrogen oxides and VOCs on the concentration of ozone. In particular, we observe the phenomenon mentioned above: the consequence of an increase in NO on atmospheric ozone depends on the concentration of the organic compounds. The ozone content is also strongly dependent on the $[NO_2]$:[NO] ratio. If this ratio is low, the $[O_3]$ content will remain low.

Dose–Response Functions

Form of the Dose–Response Function. By definition, a dose–response function starts at the origin, and in most cases it increases monotonically with dose *X,* as sketched schematically in Figure 16.6.6. At very high doses the function may level off in S-shaped fashion, implying saturation. Dose-response functions are determined from epidemiological studies or from laboratory studies. Since the latter are

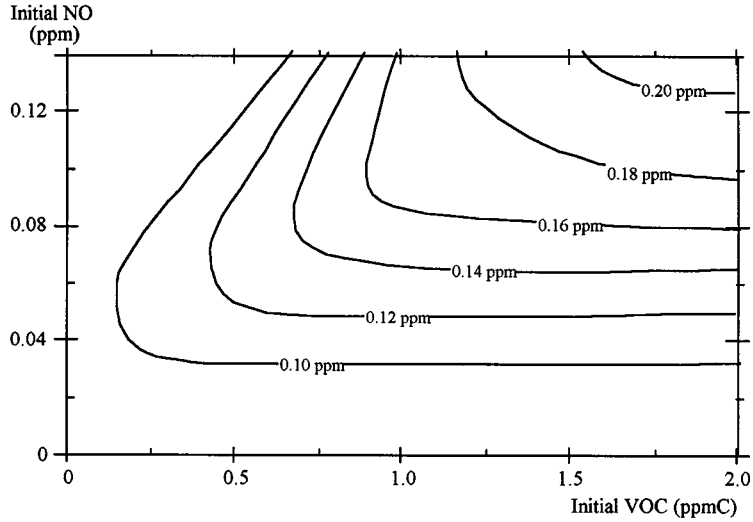


FIGURE 16.6.5 Isopleth plot for the maximum ozone concentration reached during a fixed length of time as a function of initial NO and VOC concentrations. Details of such a plot depend on site and on weather. (From EPRI 1992.)

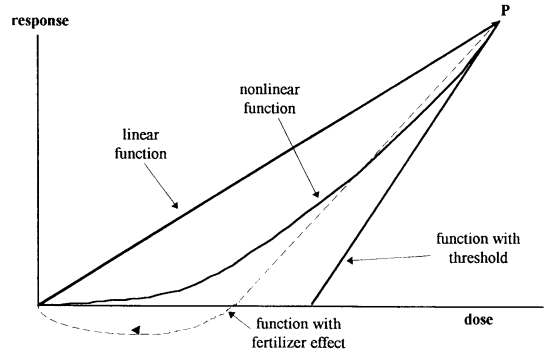


FIGURE 16.6.6 Possible behavior of dose–response functions at low doses: the four functions shown have the same value at P. For the function with threshold the discontinuity in slope at the threshold is a simplification; in reality there is a smooth transition.

mostly limited to animals, the extrapolation to humans introduces large uncertainties. Another major difficulty is that one needs relatively high doses in order to obtain observable nonzero responses in a sample of realistic size; such doses are usually far in excess of the levels one is concerned with in environmental impact studies. Thus, there is a serious problem of how to extrapolate from the observed data toward low doses. Figure 16.6.6 indicates several possibilities. The simplest is the linear model, i.e., a straight line from the origin through the observed data point(s). Cancer from radioactivity is an example. Linearity also seems to be observed for mortality from fine particulates (Dockery et al., 1993; Dockery and Pope, 1994; Lipfert, 1994).

Another possibility is a straight line down to some threshold, and zero effect below that threshold. Thresholds occur when an organism has a natural repair mechanism that can prevent or counteract damage up to a certain limit. Many dose–response functions for noncancer toxicity are of this type.

There is even the possibility of a "fertilizer effect" at low doses, as indicated by the dashed line in Figure 16.6.6. This can be observed, for example, in the dose–response functions for the impact of NO_x and SO_x on crops: a low dose of these pollutants can increase the crop yield; in other words, the damage

is negative. Such a fertilizer effect can occur with pollutants that provide trace elements needed by an organism. It depends on local conditions, in particular the overall balance of nutrients. The fertilizer effect illustrates the link between the understanding of the underlying processes and the choice of the appropriate form for the dose–response function: since N and S are known to be important nutrients for plants, a functional form like the dashed line in Figure 16.6.6 is the most plausible.

If nothing is known about a threshold and a fertilizer effect can be ruled out, the dose–response function could be anywhere between zero and the straight line through the origin, for instance, the curved solid line shown in Figure 16.6.6. *A priori* there is no general rule about the extrapolation to low doses, other than there being no known cases of a dose–response function above the straight line. There is even a case where the same substance causes different cancers according to different dose–response functions, one with and one without threshold. This was established in an experiment (sometimes referred to as the megamouse experiment) in which some 24,000 mice were exposed to the carcinogen 2-acetyl-amino-fluorene at several different dose levels (Frith et al., 1981). The response for liver tumor is linear, whereas the one for bladder tumor has a threshold.

Site Dependence of Marginal Impacts

From here on we limit ourselves to the important case where the dose–response function $f_{dr}(\mathbf{x},c(\mathbf{x}))$ can be approximated by

$$f_{dr}\left(\mathbf{x},c(\mathbf{x})\right) \approx d(\mathbf{x})c(\mathbf{x}) \quad \text{where} \quad d(\mathbf{x}) = \frac{df_{dr}\left(\mathbf{x},c(\mathbf{x})\right)}{dc} \tag{16.6.32}$$

is the slope of the dose–response function. With that assumption one can write the damage in the form

$$D = \int dx \int dy\, r(\mathbf{x}) d(\mathbf{x}) c(\mathbf{x}) \tag{16.6.33}$$

This is obviously exact for any pollutant whose dose–response function is linear, or a straight line with a threshold that is everywhere below the background. It is also valid, regardless of dose–response function, for the evaluation of any marginal impacts, i.e., impacts from small pollutant increments because in that case one can linearize the dose–response function. Since $c(\mathbf{x})$ is linear in the emission, it follows that Equation (16.6.33), and the remainder of this section, are equally applicable to steady-state situations and to emissions that vary with time.

It is instructive to relate the concentration $c(\mathbf{x})$ to the removal rate of the pollutant. There are essentially three mechanisms by which an air pollutant can disappear from the atmosphere (Seinfeld 1986):

1. dry deposition (uptake at the Earth's surface by soil, water, or vegetation);
2. wet deposition (absorption into droplets followed by droplet removal by precipitation);
3. decay or transformation (e.g., decay of radionuclides or chemical transformation of SO_2 to $(NH_4)_2SO_4$).

When evaluating the damage of the original pollutant, this pollutant is no longer counted in the equation once it has been transformed; rather from that point on a different dose–response function comes into play for the secondary pollutant. That issue will be addressed below.

The dry deposition rate is proportional to the concentration $c(\mathbf{x})$ at the Earth's surface, and it is customarily written in the form

$$F_{\text{dry}}(\mathbf{x}) = v_{\text{dry}} c(\mathbf{x}) \tag{16.6.34}$$

where $F_{dry}(\mathbf{x})$ = deposition flux (in kg/m^2·sec) and v_{dry} = dry deposition velocity (m/sec).

Wet deposition and decay or transformation can likewise be characterized in terms of fluxes $F_{wet}(\mathbf{x})$ and $F_{trans}(\mathbf{x})$, defined as the rate at which the pollutant is removed by these mechanisms per square meter and per second. Even though in general these fluxes are not proportional to the surface concentration but rather to the average concentration in the air column above x, we can write the total removal flux,

$$F(\mathbf{x}) = F_{dry}(\mathbf{x}) + F_{wet}(\mathbf{x}) + F_{trans}(\mathbf{x}) \tag{16.6.35}$$

in terms of the surface concentration $c(\mathbf{x})$ as

$$F(\mathbf{x}) = k(\mathbf{x})c(\mathbf{x}) \tag{16.6.36}$$

if we allow the proportionality constant $k(\mathbf{x})$ to vary with $\mathbf{x}$. The units of k are m/sec, and it could be called removal velocity. Using $F(\mathbf{x})$ and $k(\mathbf{x})$, we can write the damage in the form:

$$D = \int dx \int dy\, r(\mathbf{x}) d(\mathbf{x}) F(\mathbf{x})/k(\mathbf{x}) \tag{16.6.37}$$

This equation is exact if we interpret Equation (16.6.36) as the definition of $k(\mathbf{x})$.

If the world were homogeneous, with uniform receptor density $r(\mathbf{x}) = r_{uni}$, uniform dose–response function slope $d(\mathbf{x}) = d_{uni}$, and uniform removal velocity $k(\mathbf{x}) = k_{uni}$, the integral in Equation (16.6.37) would be simply

$$D = D_{uni} = d_{uni} r_{uni} Q/k_{uni} \tag{16.6.38}$$

because the surface integral of the removal flux equals the emission

$$Q = \int dx \int dy\, F(\mathbf{x}) \tag{16.6.39}$$

by conservation of matter.

Even though the assumption $k(\mathbf{x}) = k_{uni}$ may not appear very realistic, especially near a point source, the sensitivity to deviations from uniformity turns out to be surprisingly small, as we will demonstrate below in Figure 16.6.7. The reason is that for typical values of atmospheric dispersion parameters the total impact is dominated by regions sufficiently far from the source that the pollutant can be considered to be vertically well mixed in the PBL, at least as far as expectation values are concerned.

Thus, the simple Equation (16.6.38) can be a useful first estimate, good to an order of magnitude or better, independent of the details of atmospheric dispersion (Curtiss and Rabl, 1996b). It is intuitively plausible that the damage is proportional to the slope d of the dose–response function, to the density r of receptors, and to the emission rate Q. Furthermore, it is inversely proportional to the removal velocity k. If there were no removal mechanism, the pollutant concentration would increase without limit and the damage would be infinite. This approach can also be adapted for the damage due to a secondary pollutant.

To verify the relevance of Equation (16.6.38) we compare it with real site-dependent results, calculated with the PATHWAYS software package (Curtiss and Rabl, 1995; Curtiss and Rabl, 1996c), which carries out an accurate numerical integration of atmospheric dispersion results over geographic data for population and other receptors. To add substance to the results, we consider a specific impact: the increase in mortality due to SO_2 emitted by coal-fired power plants. The dose–response function (based on Sunyer et al., 1996) is linear and can be written in the form:

$$\Delta\text{deaths per year per person} = 5.34 \times \Delta PM_{10}\ \text{concentration}\left(\text{in } g/m^3\right) \qquad (16.6.40)$$

Example

We consider an annual SO_2 output of Q = 1000 ton/year = 30 g/sec. For the atmospheric dispersion we take a curve fit to the EMEP data for a grid cell in the center of France, which yields a removal velocity $k = 0.01$ m/sec. Inserting $Q = 30.9$ g/sec, $k_{uni} = 0.01$ m/sec, $d_{uni} = 5.34$ (deaths/year)/(g/m^3) and $r_{uni} = 1.05 \times 10^{-4}$ m^{-2} (for France) into Equation (16.6.38), we obtain

$$D_{uni} = \frac{5.34\ \text{deaths}/\left(\text{person}\cdot\text{yr}\right)/\left(g/m^3\right) \times 1.05\text{E-}4\ \text{persons}/m^2 \times 30.9\ g/\text{sec}}{0.01\ m/\text{sec}} = 1.78\ \text{deaths/yr}$$

This number is shown as the thick solid line in Figure 16.6.7, where the number of deaths per year is plotted vs. emission height h_e. The points, connected by lines, show the impacts for five specific sites. We have chosen these sites because there are in fact fossil fuel power plants at these sites (the nearest big city, 25 to 50 km away, is indicated in parentheses). Although the real emissions at the different sites are different, here we have assumed the emissions of the Cordemais plant at all sites to bring out the point of the comparison. The impact is about 2 to 6 times larger than D_{uni} for the site near Paris and about 0.4 to 0.5 times D_{uni} for Cordemais, a rural site on the Atlantic Ocean. We also see that there is little variation with stack height.

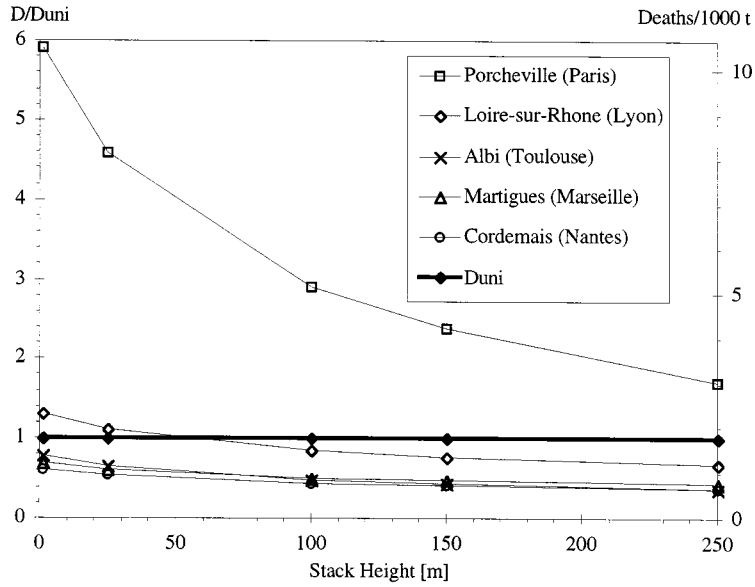


FIGURE 16.6.7 An example of dependence on site and on height of source: mortality from particulate matter from a coal-fired power plant, for five sites in France and for uniform world model (Equation (16.6.38)). Annual SO_2 emission 1000 ton/year.

Analysis of Uncertainties

By contrast to the relatively small uncertainties and normal (Gaussian) frequency distributions typically encountered in science and engineering, the uncertainties in impact analysis are so large that it would be inappropriate to use error intervals that are additively symmetric about the mean. Instead, one should specify multiplicative intervals, in other words, intervals that are additive on a logarithmic scale. The frequency distributions are not symmetric, with implications that may appear counterintuitive to people

not accustomed to them. It is helpful to think in terms of lognormal distributions because they are frequently encountered in impact analysis, analogous to the normal distributions so familiar in the more exact sciences. A variable x has a lognormal distribution if the variable $\ln(x)$ has a normal distribution; in other words, it is normal on a logarithmic scale.

Analogous to the ordinary normal (also known as Gaussian) distribution, which is characterized by two parameters, the mean and the standard deviation, the lognormal distribution can be characterized by the geometric mean and the geometric standard deviation s_G. For this distribution the geometric mean is equal to the median: half of the distribution is above, the other half below the median. The geometric standard has a simple interpretation in terms of the 68% confidence interval (a familiar number because for Gaussian distributions 68% of all values are within one standard deviation of the mean): for a lognormal distribution 68% of the values are within the interval $(1/s_G,s_G)$. Likewise, 95% are within the interval $[(1/s_G)^2,(s_G)^2]$. Note, however, that these values are centered around the median rather than the mean; the lognormal distribution is not symmetric. For impacts of primary air pollutants with relatively well-determined dose–response functions, e.g., mortality due to particulates, s_G may be as small as 3. For other impacts, e.g., cancers due to dioxins, the uncertainties could be an order of a magnitude or more.

It is appropriate to note that technical or scientific uncertainties (e.g., uncertainties of emitted quantities or of dose–response functions) are not the only ones. For long-term impacts, such as cancers caused by radioactive waste, one needs to make assumptions about scenarios for the future: what quantities of radionuclides will leak into the environment and how many people will be affected by them. For the estimation of damage costs, there is also the matter of policy/ethical choice, e.g., about discount rate and value of human life.

16.7 Global Climate Change

Frank Kreith

There is consensus in the scientific community that continuing the emission of CO_2 and other "greenhouse" gases (methane, nitrous oxide, ozone, and chlorofluorocarbons or CFCs) into the atmosphere at current rates will lead to a warming of the Earth. General circulation models of the atmosphere indicate that a doubling of CO_2 concentration in the atmosphere will trap sufficient solar radiation to increase the average global temperature by the middle of the 21st century at least 2°C, but possibly as much as 5°C according to the Congressional Research Service (Morrison, 1989). Although critics of these projections argue that the available models do not accurately portray the potential feedback mechanisms from clouds and oceans to warming, and question the amount, timing, and location of the temperature increase, there is agreement that greenhouse gases trap radiation and increase the global temperature. The latest international scientific assessment concluded that "evidence suggests that there is a discernible human influence on global climate" (Watson et al., 1995).

Despite uncertainty over the magnitude, timing, and potential impacts of global warming, there is general agreement that those policies and technologies that reduce the emission of greenhouse gases and have no adverse economic impact should be implemented now. Highest priorities relate to energy conservation, improved efficiency in energy generation, emission reduction from transportation, reforestation, and phasing out of CFCs (see Chapter 9). These proposed actions represent challenges to many facets of mechanical engineering.

Technical Background

The greenhouse effect is a natural phenomenon and one of the most well established theories in atmospheric science (see Figure 16.7.1). Most of the solar radiation impinging on Earth is in the frequency range below 3 μm. Radiation in this frequency band can readily penetrate Earth's atmosphere and is absorbed by the Earth. But most of the thermal radiation emitted by the Earth is infrared, i.e., in a frequency range above 3 μm, which is partly absorbed and reflected by the atmosphere. Gases in the atmosphere thus prevent part of the radiation emitted by the Earth from escaping into space and trap enough radiation to maintain our climate equilibrium. The atmosphere acts like the glass windows in a greenhouse — hence, the name *greenhouse effect*. Without this effect the Earth would be about 60°F cooler and life as we know it would not exist.

Figure 16.7.2 shows the distribution of the gases in the atmosphere that participate in the greenhouse effect, as well as the industrial sectors emitting them. An increase in the amount of atmospheric greenhouse gases increases the amount of radiation absorbed by the atmosphere and thus reduces the amount of radiation emitted by the Earth that can pass into space. This causes warming of the Earth. Carbon dioxide is the principal greenhouse gas, responsible for over one half of the predicted warming of the Earth. Most of the generation of CO_2 results from the burning of fossil fuels. Its primary sources are electric utilities (33%), transportation (31%), industrial processes (24%), and heating and cooling of buildings (12%). Methane, CFCs, ozone, and nitrous oxide constitute the balance of greenhouse emission. Carbon dioxide emissions have increased globally by 25% since the industrial revolution 200 years ago, and there has been an over 10% increase in the last 30 years alone. In the same period, methane concentration has more than doubled (from 800 to 1700 ppb).

Temperature records compiled by the United Kingdom Meteorological Office indicate that the Earth's average temperate has risen roughly 1°F since 1860 (see Figure 16.7.3). Since CO_2 is the principal greenhouse gas, estimating its growth is important for future prediction of climate change. Although all fossil fuels emit CO_2 when burned, their emission per unit energy produced is not equal. Natural gas is the cleanest fossil fuel. Oil combustion emits between 38 and 43% more CO_2 than natural gas, whereas coal combustion contains from 72 to 95% more CO_2 per unit energy released than natural gas. Ocean

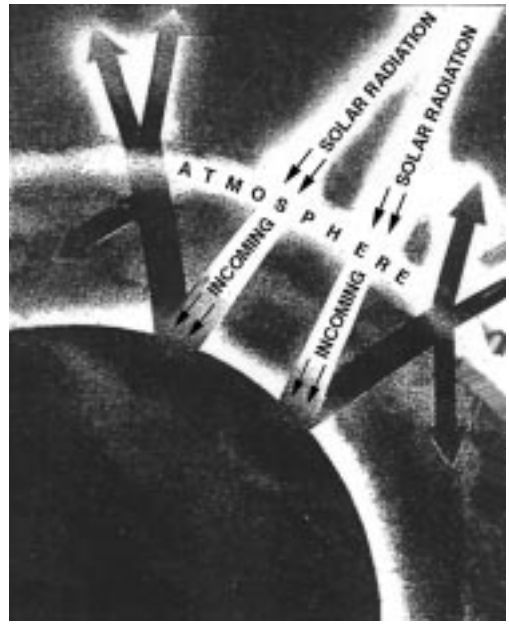


FIGURE 16.7.1 Schematic diagram of the greenhouse effect.

and trees act as carbon sinks because they absorb CO_2. Trees absorb CO_2 during the process of photosynthesis. Deforestation is important because it not only eliminates a mitigating factor of global warming, but when wood is burned, it too emits carbon dioxide directly into the atmosphere.

Potential Impacts of Global Climate Change

In its 1989 report to Congress, the USEPA sought to identify the impacts of global warming on specific geographic areas. The agency used simulations from three general circulation models — Goddard Institute for Space Studies (GISS), Geophysical Fluids Dynamics Laboratory (GFDL), and Oregon State University (OSU) — to project the effects of climate change on four regions: Southeast, Great Lakes, Great Plains, and California. The three models are in general agreement that a doubling of carbon dioxide in the atmosphere will increase the temperature, but they differ in magnitude, predicting between 3 and 5°C. Although they concur that annual precipitation will increase, their regional projections vary — the GFDL model predicts reduced precipitation in the Southeast, the GISS in the Great Plains, and the OSU in California. The report acknowledges, however, that the reliability of these regional predictions is limited.

The USEPA assessment of the potential impact of climate change on natural resources and the environment has serious implications for forests, agriculture, water resources, biodiversity, and sea-level rise. The southern boundary of forest species is expected to move northward and drier soils could alter the plant composition with grasslands and hardwoods replacing commercially valuable conifers. Also, fire and pest disturbances could become more frequent. Agricultural productivity would shift northward, causing economic dislocations for farmers in southern states.

One of the most important effects of global warming is a rise in sea level resulting from melting glaciers. Projections of future sea-level rise range from 0.3 to 1.1 m (1 to 3.5 ft) by the end of the next century. A 1-m rise could flood between 26 and 60% of the nation's coastal wetlands and cost up to $100 billion to protect developed areas that would otherwise be flooded.

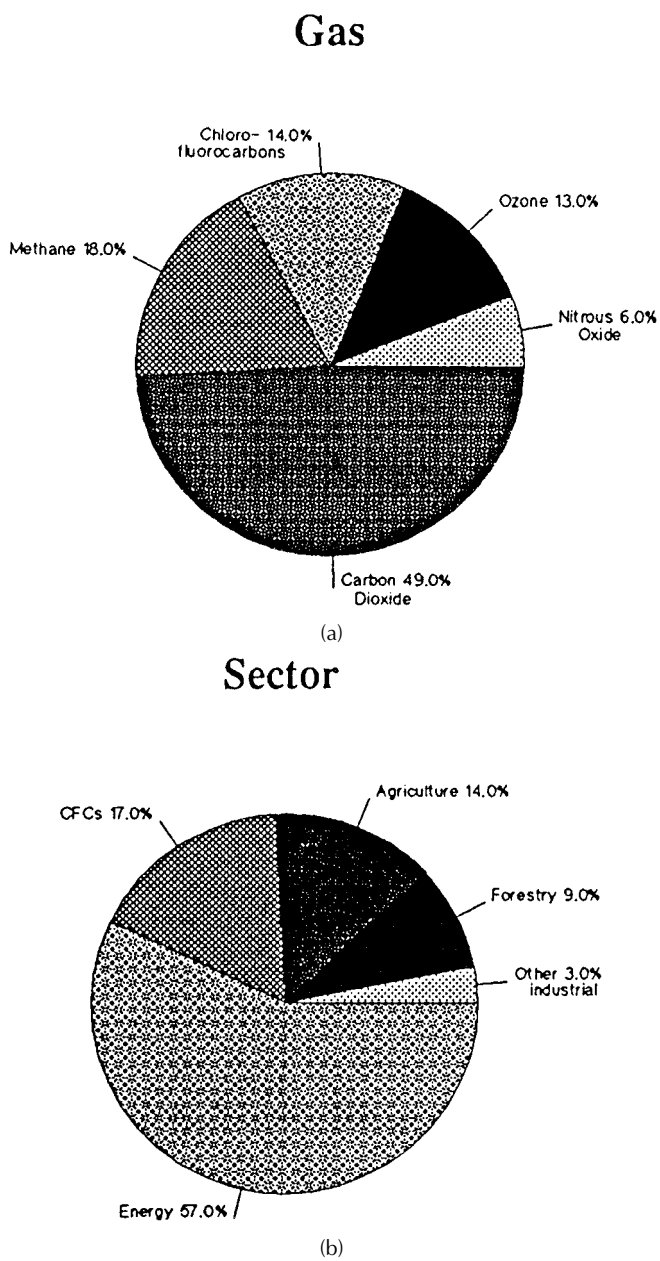


FIGURE 16.7.2 Greenhouse gas profile. (From U.S. Environmental Protection Agency, 1989.)

Mitigation Options for Global Warming

There is general agreement that an initial global warming will result from increases in greenhouse gases in the atmosphere, and the position of the U.S. Government, according to the Office of Science and Technology Policy issued in 1996, is that "we collectively must redouble our efforts in identifying the most cost-effective global approaches to reduce emissions in both the near- and long-term utilizing a broad portfolio of actions." However, the scientific uncertainty regarding the timing and severity of future climate impacts and their consequences creates a policy dilemma of what actions should be taken.

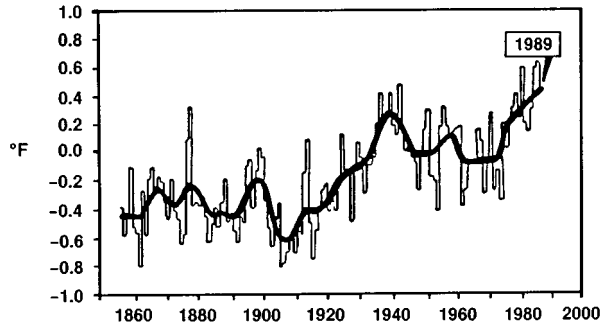


FIGURE 16.7.3 Global temperature change since 1860. Variation of annual surface temperatures for the world. The solid line shows long-term trends. The planet has warmed an average of almost 1° F. (From United Kingdom Meteorological Office, 1989.)

Because measures to limit greenhouse gas emission require efforts lasting many years and could affect the economic health of a country, the costs of climate policy and the technical means of achieving emission reduction must be considered with the focus on obtaining the largest reduction in potential greenhouse warming at the lowest cost to society. To date, CO_2 from combustion of fossil fuels has been the primary focus of attention. A comprehensive look at mitigation options, however, must consider the emission of all greenhouse gases to compare the relative importance of different emissions. The concept of a global warming potential to estimate the CO_2-equivalent emission of each of the major greenhouse gases is useful. Table 16.7.1 shows an estimate of greenhouse gas emissions from various human activities.

Various scenarios of the CO_2-equivalent reduction achievable by various mitigation measures have been studied. The various mitigation measures were grouped into two categories. The "best practice" technology options shown in Table 16.7.2 can be implemented now at no cost or even at net savings, but they are not used because of institutional and other barriers. Additional options shown in Table 16.7.3 either cost money or have benefits that are not readily quantified or face other implementation obstacles. In this study, no forecasts or scenarios for the future were used, but the results were applied to a 1989 base year.

Inspection of Table 16.7.2 indicates that improved energy efficiency in residential and commercial buildings will reduce demand for fossil fuel and, thereby, also reduce emissions. Conservation measures that reduce the use of electric power are the most significant. For example, improved commercial lighting to reduce the energy consumption by about 45% would not only save money, but if installed in commercial buildings, would result in electricity savings of about 10% from space-conditioning requirements.

The potential for reducing electricity consumption in the industrial sector with currently available technology is about 30%. A case study of energy-intensive industries such as steel mills and petrochemical plants indicates that energy savings of the order of 25 to 30% in direct industrial fuel use may be available economically by investing in more-efficient furnaces, energy recovery systems, and other process equipment. Also, increased use of cogeneration technologies would produce cost-effective CO_2 reductions as shown in Table 16.7.2.

The other important sector in which large reductions in greenhouse gas emissions can be achieved is transportation. These reductions in greenhouse gas emissions are largely through improved fuel economy and better transportation management methods that reduce traffic congestion (see Section 10, Transportation).

Mitigation measures outside the energy sector involve landfills, CFC use, agricultural activity, and forests. For example, the collection and combustion of landfill gas could reduce methane emission by about 65%. For a discussion of the other mitigation methods, the reader is referred to the original studies cited.

TABLE 16.7.1 Estimate of Current Greenhouse Gas Emissions from Human Activity

Source	Annual Emissions (Mt/year)[a]		CO_2-Equivalent (Mt/year)[a]	
	World	U.S.	World	U.S.
	CO_2 emissions			
Commercial energy	18,800		18,800	
Tropical deforestation	2,600		2,600	
Other	400		400	
Total CO_2	**21,800**	**4,800**	**21,800**	**4,800**
	CH_4 emissions			
Rice cultivation	110		2,300	
Enteric fermentation	70		1,500	
Fuel production	60		1,300	
Landfills	30		600	
Tropical deforestation	20		400	
Other	30		600	
Total CH_4	**320**	**50**	**6,700**	**1,050**
	CFC emissions			
Total CFCs	**0.6**	**0.3**	**3,200**	**1,640**
	N_2O emissions			
Fertilizer use	1.5		440	
Coal combustion	1.0		290	
Tropical deforestation	0.5		150	
Agricultural wastes	0.4		120	
Land cultivation	0.4		120	
Fuel and industrial biomass	0.2		60	
Total N_2O	**4.0**	**1.4**	**1,180**	**410**
Overall total			**32,880**	**7,900**

[a] Millions of metric tons based on the estimated global warming potential (GWP) for a 100-year averaging time (6): CO_2 = 1, CH_4 = 21, N_2O = 290, CFC-11 = 3500, CFC-12 = 7300 and CFC-113 = 4200. Values give the CO_2-equivalent radiative forcing for an instantaneous injection of 1 kg of gas into the atmosphere. Values for CH_4 include the estimated indirect effects of CO_2 produced. However, the GWP does not incorporate complex couplings with other greenhouse gases such as stratospheric and tropospheric ozone and their precursor emissions. The GWP thus provides only a preliminary basis for comparing diverse mitigation strategies.

From Ruben, E.S., Realistic Mitigation Options for Global Warming, *Science,* 257, 261–266, 1993. With permission.

TABLE 16.7.2 Best-Practice Technology Options Available at Little or No Net Cost that are not Fully Implemented Due to Institutional and Other Barriers

Option	CO_2-Equivalent Reduction[a]	Net Cost[b]
	Residential and commercial energy use	
Electricity efficiency		
1. White roofs and trees[c]	32	–84
2. Res. lighting[d]	39	–79
3. Res. water heating[e]	27	–74
4. Com. water heating[f]	7	–72
5. Com. lighting[g]	117	–71
6. Com. cooking[h]	4	–70
7. Com. cooling[i]	81	–64
8. Com. refrigeration[j]	15	–60
9. Res. appliances[k]	72	–44
10. Res. space heating[l]	74	–39
11. Com. and Ind. space heating[m]	15	–35
12. Com. ventilation[n]	32	1
Oil and gas efficiency[o]	300	–62
Fuel switching[p]	74	–90
Sector total	**890**	**–62 (–78/–47)**
	Industrial energy use	
Electricity efficiency[q]	137	–43
Fuel use efficiency[r]	345	–24
New cogeneration[s]	45	–18
Sector total	**527**	**–28 (–42/–14)**
	Transportation energy	
Light-duty vehicles[t]	251	–40
Heavy-duty trucks[u]	39	–59
Sector total	**290**	**–43 (–21/–75)**
	Power plants	
Coal plants[v]	45	–0
Hydroelectric plants[w]	12	–0
Nuclear plants[x]	42	2
Sector total	**99**	**1 (0/2)**
Landfill gas[y]	230	**1 (0.4/2)**

TABLE 16.7.2 Best-Practice Technology Options Available at Little or No Net Cost that are not Fully Implemented Due to Institutional and Other Barriers (continued)

Option	CO_2-Equivalent Reduction[a]	Net Cost[b]

Numbers with options refer to steps in Figure 1. Com., commercial; Res, residential; Ind, industrial.

[a] Equivalent CO_2 reduction in millions of metric tons based on 1989 fuel and electricity use. [b]Net implemented cost in dollars per ton of CO_2-equivalent. Costs are mid-range estimates based on a 6% real discount rate, constant 1989 dollars. Parentheses give low/high range for average cost reflecting real discount rates of 3 to 10% plus uncertainty across different studies or estimates. [c]Plant shade trees and paint roofs white at 50% of residences to reduce air conditioning use and the urban heat island effect by 25%. [d]Replace incandescent lighting (2.5 inside and 1 outside light bulb per residence) with compact fluorescents to reduce lighting energy consumption by 50%. [e]Efficient tanks, increased insulation, low-flow devices, and alternative water heating systems to improve efficiency by 40 to 70%. [f]Residential measures above, plus heat pumps and heat recovery systems to improve efficiency by 40 to 60%. [g]Replace 100% of commercial light fixtures with compact fluorescent lighting, reflectors, occupancy sensors, and day lighting to reduce lighting energy consumption by 30 to 60%. [h]Additional insulation, seals, improved heating elements, reflective pans, and other measures to increase efficiency by 20 to 30%. [i]Improved heat pumps, chillers, window treatments, and other measures to reduce commercial cooling energy use by 30 to 70%. [j]Improved compressors, air barriers, food case enclosures, and other measures to improve efficiency 20 to 40%. [k]Implementation of new appliance standards for refrigeration and use of no-heat drying cycles in dishwashers to improve efficiency of refrigeration and dishwashers by 10 to 30%. [l]Improved and increased insulation, window glazing, and weather stripping along with increased use of heat pumps and solar heating to reduce energy consumption by 40 to 60%. [m]Use measures similar to residential sector to reduce energy consumption by 20 to 30%. [n]Improved distribution systems, energy-efficient motors, and other measures to improve efficiency 30 to 50%. [o]Efficiency measures similar to those for electricity to reduce fossil fuel energy use by 50%. [p]Switch 10% of building electricity use from electric resistance heat to natural gas heating to improve overall efficiency by 60 to 70%. [q]More efficient motors, electrical drive systems, lighting and industrial process modifications to improve electricity efficiency by 30%. [r]Energy management, waste heat recovery, boiler modifications, and other industrial process enhancements to reduce fuel consumption by 30%. [s]An additional 25,000 MW of co-generation plants to replace existing industrial energy systems. [t]Use existing technology to improve fuel economy to 32.5 mpg (CAFE) with no changes in the existing fleet. [u]Use existing technology to improve fuel economy to 18.2 mpg (CAFE) with no changes in the existing fleet. [v]Improve efficiency of existing plants by 3% through improved plant operation and maintenance. [w]Improve efficiency by 5% through equipment modernization and maintenance. [x]Increase the annual average capacity factor of existing plants from 60 to 65% through improved maintenance and operation. [y]Reduce landfill gas generation by 60 to 65% by collecting and burning in a flare or energy recovery system.

From Ruben, E.S., Realistic Mitigation Options for Global Warming, *Science,* 257, 261–266, 1993. With permission.

TABLE 16.7.3 Additional Mitigation Options that are Costly or that have Significant Other Benefits or Costs that are not Readily Quantified. Some of these Options would Face Serious Implementation Obstacles because of Such Factors

Mitigation Option	CO_2-Equivalent Reduction[a]	Net Cost[b]
Industrial energy use		
Fuel switching[c]	24	60
Transportation energy use		
Demand management[d]	49	–22 (–50/5)
Light-duty vehicle efficiency (change in fleet mix)[e]	53	530 (40/1020)
Aircraft engine efficiency[f]	13	360
Electric supply technology[g]		
Advanced coal[h]	200	280
Natural gas[i]	850	32 (17/46)
Nuclear[j]	1500	49 (28/69)
Hydroelectric	30	38
Biomass	130	36 (29/42)
Wind	30	79 (33/125)
Solar photovoltaic	400	87
Solar thermal	540	160
Sector total[k]	**1780**	**50 (30/70)**
Halocarbon use[l]		
Non-halocarbon substitutes[m]	302	0.02
CFC conservation[n]	509	0.04
HCFC HFC/aerosols, etc.[o]	248	0.6
HFC (chillers)[p]	88	3
HFC (auto air conditioning)[q]	170	5
HFC (refrigerators)	11	11
HCFC (other refrigeration)[r]	67	4
HCFC/HFC (refrigerator insulation)	14	28
Sector total	**1409**	**1.4 (0.9/3)**

TABLE 16.7.3 Additional Mitigation Options that are Costly or that have Significant Other Benefits or Costs that are not Readily Quantified. Some of these Options would Face Serious Implementation Obstacles because of Such Factors (continued)

Mitigation Option	CO_2-Equivalent Reduction[a]	Net Cost[b]
	Domestic agriculture	
Nitrogenous fertilizers[s]	126	2.5
Paddy rice[t]	23	0.5
Ruminant animals[u]	84	2.0
Sector total	**223**	**2 (1/5)**
Reforestation[v]	242	7 (3/10)
	Other options	
New industrial technology[w]	300	?
New transportation fuels[x]	1130	?

[a] Equivalent CO_2 reduction in millions of metric tons based on 1989 fuel and electricity use. [b]Net implemented cost in dollars per ton of CO_2-equivalent. Includes direct costs only. Many of these measures have additional indirect costs that could be significant (see text). Values are mid-range cost estimates based on a 6% real discount rate, constant 1989 dollars. Values in parentheses give low/high range for average cost for real discount rates of 3 to 10% plus uncertainty across different studies or estimates. [c]Switch current coal consumption in industrial plants to natural gas or oil where technically feasible (estimated at 0.6 quadrillion Btu). [d]Estimate 25% of employer-provided parking spaces and tax remaining spaces to reduce solo commuting by 15 to 20%. [e]Improve on-road fuel economy from 32.5 to 46.8 mpg (CAFE) with additional technology measures and downsizing that require changes in the existing fleet mix. [f]Implement improved fan jet and other technologies to improve fuel efficiency by 20%. [g]Potential emission reductions apply only to one technology at a time and are not cumulative. All cost-effectiveness estimates are relative to existing (1989) coal plants. [h]Based on advanced pulverized coal plants. [i]Based on the use of combined cycle systems in place of coal. Co-firing natural gas at existing coal-fired plants has similar costs but lower reduction potential. [j]Based on advanced light-water reactors replacing current fossil-fuel capacity for baseload and intermediate load operation. [k]Based on replacing all fossil fuel plants in the 1989 generating mix. Replacement of coal plants only yields 1470 Mt/year. Remaining potential after maximum demand reductions and plant upgrades is 950 Mt/year. [l]Includes chlorofluorocarbons (CFC), hydrofluorocarbons (HFC), and hydrochlorofluorocarbons (HCFC). [m]Modify or replace existing equipment to use non-CFC materials as cleaning and blowing agents, aerosols, and refrigerants where technically possible. [n]Upgrade equipment and retrain personnel to improve conservation and recycling of CFCs. [o]Substitute cleaning and blowing agents and aerosols with fluorocarbon substitutes. [p]Retrofit or replace all existing chillers to use fluorocarbon substitutes. [q]Replace existing automobile air conditioners with equipment using fluorocarbon substitutes. [r]Replace commercial refrigeration equipment such as used in supermarkets and transportation with equipment using fluorocarbon substitutes. [s]Reduce nitrogenous fertilizer use by 5%. [t]Eliminate all U.S. paddy rice production. [u]Reduce ruminant animal production by 25%. [v]Reforest 28.7 Mha of economically or environmentally marginal crop and pasture lands and nonfederal forest lands. [w]Increase recycling and reduce energy consumption primarily in the primary metals, pulp and paper, chemicals, and petroleum refining industries through new, less energy intensive technology. [x]Based on replacement of highway transport fuels with alternative fuels that emit no greenhouse gases.

From Ruben, E.S., Realistic Mitigation Options for Global Warming, *Science,* 257, 261–266, 1993. With permission.

Defining Terms

Aerosol: Small solid particle or liquid droplet suspended in the air.
Anthropogenic: Man-made.
Deterministic: Dealing with cause and effect.
Ensemble average: Theoretical average, i.e., the value that could be expected as the average from an infinite number of realizations.
Global warming: The warming of the earth as a result of the atmosphere which traps solar radiation. Currently, the term global warming is used to denote the temperature increase due to carbon dioxide, methane and other greenhouse gases.
Greenhouse Gases: Carbon dioxide, methane and other gases resulting from human activities that increase the amount of solar radiation trapped by the atmosphere.
Numerical advection errors: Numerical errors generated by finite-difference solutions of transport terms using an Eulerian grid model.
Organic compounds: Chemical species containing one or more carbon atoms.
Photochemical: Chemical reactions influenced by light.
Photolysis: Chemical decomposition by the action of light.
Planetary boundary layer, PBL: The atmospheric layer which is affected by the momentum and heat fluxes generated by the Earth surface (typically, the first 500–1000 m of the atmosphere).
Stratospheric: Related to the stratosphere, the portion of the atmosphere approximately between 10 to 50 km above the ground.
Tropospheric: Related to the troposphere, the lower level of the atmosphere approximately from the surface to 10 km above.
Wind shear: The change in wind speed and direction as a function of height.

References

APHA, AWWA, and WPCF. 1992. *Standard Methods for the Examination of Water and Wastewater,* 18th ed., American Public Health Association, American Water Works Association, Water Pollution Control Federation, Washington, D.C.

AWWA. 1996, *National Primary Drinking Water Contaminant Standards, Opflow,* Vol. 22, No. 3, March 1996, American Water Works Association, Denver, CO.

Barrett, K. 1992. Dispersion of Nitrogen and Sulfur across Europe from Individual Grid Elements: Marine and Terrestrial Deposition, EMEP/MSC-W Note 3/92. August 1992. Norwegian Meteorological Institute, P.O. Box 43, Blindern, N-0313 Oslo 3.

Boubel, R.W., Fox, D.L., Turner, D.B., and Stern, A.C., 1994. *Fundamentals of Air Pollution*, Academic Press, San Diego, CA, 574 pp.

Cohrssen, J.J. and Covello, V.T. 1989. *Risk Analysis: A Guide to Principles and Methods for Analyzing Health and Environmental Risks,* United States Environmental Protection Agency. Report EPA PB89-137772. Washington, D.C. 20460.

Corey, A.T. 1994. *Mechanics of Immiscible Fluids in Porous Media,* Water Resources Publications, Littleton, CO.

Curtiss, P.S. and A. Rabl, A., 1995. *The PATHWAYS2.0 Impact Analysis Program*. Ecole des Mines de Paris.

Curtiss, P.S., Hernandez, B., Pons, A., Rabl, A., Dreicer, M., Tort, V., Margerie, H., Landrieu, G., Desaigues, B., and Proult, D., 1995. Environmental Impacts and Their Costs: the Nuclear and the Fossil Fuel Cycles, ARMINES (Ecole des Mines), 60 boul. St.-Michel, 75272 Paris CEDEX 06.

Curtiss, P.S. and Rabl, A., 1996a. *PATHWAYS: A Software Package for Calculating Impacts and Costs of Environmental Burdens due to Electricity Production by Nuclear or Fossil Fuels. Program Manual,* Ecole des Mines de Paris, Paris.

Curtiss, P.S. and Rabl., A., 1996b. Impacts of air pollution: general relationships and site dependence. Submitted to *Atmos. Environ.*, 30, 3331–3347.

Curtiss, P.S. and Rabl., A., 1996c. Impact Analysis for Air and Water Pollution: Methodology and Software Implementation. *Environmental Modeling*. Zannetti, P., Ed. Vol. 3, Chap. 13, p.393–426.

de Vera, L. et al. 1980. *Samplers and Sampling Procedures for Hazardous Waste Streams,* EPA-600/2-80-018, USEPA, Washington, D.C.

Dockery, D.W. and Pope, C.A., III, 1994. Acute respiratory effects of particulate air pollution. *Annu. Rev. Public Health*, 15, 107–132.

Dockery, D.W., Pope, C.A., III, Xu, X., Spengler, J.D., Ware, J.H., Fay, M.E., Ferris, B.G., and Speizer, F.E., 1993. An association between air pollution and mortality in six U.S. cities, *New Eng. J. Med.*, 329, 1753–1759.

EC, 1995a. Externalities of Fuel Cycles, "ExternE" Project: Summary Report, Report No. 1. European Commission, Directorate-General XII, Science Research and Development. JOULE programme.

EC, 1995b. Externalities of Fuel Cycles, "ExternE" Project: Coal Fuel Cycle, Report No. 2. European Commission, Directorate-General XII, Science Research and Development. JOULE programme.

EPRI. 1992. *The use of photochemical air quality models for evaluating emission control strategies. A synthesis report*. Prepared for EPRI (Electric Power Research Institute) by ENVAIR, Albany, California.

Frith, C.H., Littlefield, N.A., and Umholtz, R., 1981. Incidence of pulmonary metastases for various neoplasms in BALB/cStCrlfC3H/Nctr female mice fed N-2-fluorenylacetamide, *JNCI,* 703–712.

Heijungs, R. et al. 1992. *Environmental Life Cycle Assessment of Products,* Part 1. *Guide*. Part 2. *Backgrounds,* Centre of Environmental Science, Garenmarkt 1, P.O. Box 9518, 2300 RA Leiden, Netherlands.

Hem, J.D. 1989. Study and Interpretation of the Chemical Characteristics of Natural Water, U.S. Geological Survey Water Supply Paper 2254, U.S. Government Printing Office, Washington, D.C.

Hohmeyer, O. 1988. *Social Costs of Energy Consumption,* Springer-Verlag, New York.

Houghton, J.T., Jenkins, G.J., and Ephrams, J.J. 1990. *Climate Change, the IPCC Scientific Assessment*, Cambridge University Press, Cambridge, MA.

Iversen, T., 1993. Modeled and measured transboundary acidifying pollution in Europe — verification and trends, *Atmos. Environ.*, 27A, 889–920.

Kelly G.N. and Jones, J.A., 1985. The relative importance of collective and individual doses from routine releases of radiocative materials to the atmosphere. *Ann. Rev. Nuclear Energy*, Vol. 12, p.665–673.

Koomey, J., 1990. Comparative Analysis of Monetary Estimates of External Environmental Costs Associated with Combustion of Fossil Fuels, Report LBL-28313. Lawrence Berkeley Laboratory, Berkeley, CA 94720.

LaGrega, M.D., Buckingham, P.L., and Evans, J.C. 1994. *Hazardous Waste Management*, McGraw-Hill New York, 1146 pp.

Lange, R. 1978, ADPIC — a three-dimensional particle-in-cell model for the dispersal of atmospheric pollutants and its comparison to regional tracer studies, *J. Appl. Meteor.,* **17**, 320.

Leeden, F., Troise, F.L., and Todd, D.K., 1990. *The Water Encyclopedia*, Lewis Publishers, Chelsea, MI.

Lipfert, F.W., 1994. *Air Pollution and Community Health: A Critical Review and Data Sourcebook,* Van Nostrand Reinhold, New York.

Longhetto, A., Ed. 1980, *Atmospheric Planetary Boundary Layer Physics,* Elsevier, New York.

Milford, J.B. and Russell, A.G., 1993, Atmospheric models: atmospheric pollutant dynamics, meteorology and climate, *Environmental Modeling*, *Vol. I,* Computational Mechanics Publications, Chapter 2.

Morandi, L., 1992. *Global Climate Change,* NCSL, Denver, CO, November.

Morrison, R.E. 1989. *Global Climate Change,* Congressional Research Service, Library of Congress 1B8905; Washington, D.C.

Nemerow, N.L. 1991. *Industrial and Hazardous Waste Treatment*, Van Nostrand Reinhold, New York.

Nieuwstadt, F.T. and van Dop, H., Eds., 1982, *Atmospheric Turbulence and Air Pollution Modeling,* D. Reidel, Dordrecht, Holland.

ORNL/RFF. 1994a. Fuel Cycle Externalities: Analytical Methods and Issues, Report on the External Costs and Benefits of Fuel Cycles. July 1994. Prepared by Oak Ridge National Laboratory and Resources for the Future, Oak Ridge National Laboratory, Oak Ridge, TN 37831.

ORNL/RFF. 1994b. Estimating Externalities of Coal Fuel Cycles, Report on the External Costs and Benefits of Fuel Cycles. September 1994. Prepared by Oak Ridge National Laboratory and Resources for the Future, Oak Ridge National Laboratory, Oak Ridge, TN 37831.

Ottinger, R.L. et al. 1991. *Environmental Costs of Electricity,* Oceana Publications, New York.

Pescod, M.B. 1992. Wastewater Treatment and Use in Agriculture, FAO Irrigation and Drainage Paper 47, Food and Agriculture Organization of the United Nations, Rome.

Rabl, A., P.S. Curtiss, J.V. Spadaro, B. Hernandez, A. Pons, M. Dreicer, V. Tort, H. Margerie, G. Landrieu, B. Desaigues, and D. Proult, 1996. *Environmental Impacts and Costs: the Nuclear and the Fossil Fuel Cycles*. Report to EC, DG XII, Version 3.0 June 1996. ARMINES (Ecole des Mines), 60 boul. St.-Michel, 75272 Paris CEDEX 06.

Rodricks, J.V. 1992. *Calculated Risks: The Toxicity and Human Health Risks of Chemicals in Our Environment,* Cambridge University Press, Cambridge, U.K.

Rodriguez, D.J., Greenly, G.D., Gresho, P.M., Lange, R., Lawver, B.S., Lawson, L.A., and Walker, H. 1982. User's Guide to the MATHEW/ADPIC Models, Lawrence Livermore National Laboratory Document UASG 82-16, University of California Atmospheric and Geophysical Sciences Division, Livermore, CA.

Ruben, E.S. et al. 1992. Realistic mitigation options for global warming, *Science*, 257, July, 148–149, 261–266.

Sandnes, H. 1993. Calculated Budgets for Airborne Acidifying Components in Europe, EMEP/MSC-W Report 1/93. July 1993. Norwegian Meteorological Institute, P.O. Box 43, Blindern, N-0313 Oslo 3.

Schneider, S.H. 1989. The greenhouse effect: science and policy, *Science*, 243, February, 771–779.

Schwartz J. 1993. Air pollution and daily mortality in Birmingham, Alabama, *Am. J. Epidemiol.,* 137, 1136–1147.

Scientific Perspectives on the Greenhouse Problem, 1989. George C. Marshall Institute, Washington, D.C.

Scott, B.C. 1982. Theoretical estimates for scavenging coefficient for soluble aerosol as function of precipitation type, rate, and altitude, *Atmos. Environ.,* 16, 1735–1762.

Seigneur, C. 1987, Computer simulation of air pollution chemistry, *Environ. Software,* 2, 116.

Seinfeld, J.H. 1986. *Atmospheric Chemistry and Physics of Air Pollution,* John Wiley and Sons, New York.

SETAC. 1992. *Code Practice of Life Cycle Assessment,* Society of Environmental Toxicology and Chemistry, Pensacola, FL.

Sunyer, J., Castellsague, J., Saez, M., Tobias, A., Anto, J.M., 1996. Air pollution and mortality in Barcelona. *J. Epidem. Commun. Hlth.,* 50 (Suppl 1): S76-S80.

USEPA. 1976. *Quality Criteria for Water*, U.S. Government Printing Office, Washington, D.C.

USEPA. 1978, *Guideline on Air Quality Models,* EPA-450/2-78-027, USEPA, Research Triangle Park, NC.

USEPA. 1979 Handbook for Sampling and Sample Preservation of Water and Wastewater, U.S. Environmental Protection Agency, U.S. Government Printing Office, Washington, D.C. EPA-600/4-82-029 (with addendum).

USEPA. 1983. Guidelines for Performing Regulatory Impact Analysis, U.S. Environmental Protection Agency, Office of Policy Analysis. Report EPA-230-01-84-003, reprinted March 1991. Washington, D.C. 20460.

USEPA. 1992. *National Water Quality Inventory — Report to Congress,* EPA 503/9-92/006. Office of Water, USEPA, Washington, D.C.

U.S. Geological Survey can supply long-term, baseline monitoring of water resources in terms of quantity, flow, and quality, and special short term, regional studies.

USEPA. 1991. Technical Support Documents for Water Quality-Based Toxics Control. EPA 505/2-90-001. Office of Water, USEPA, Washington, D.C. Gives access to models such as EXAMS-II (lake, river, estuary for organics); WASP4 or WASP5.x (lake, river, estuary for organics and metals); HSPF (rivers, organics and metals); SARAH-2 (rivers, treatment plant, organic); DYNTOX (river, organics, metals).

Wackter, D.J. and Foster, J.A., 1987. *Industrial source complex (ISC) dispersion model user's guide*. 2nd ed. Vol. 1, EPA 450/4-88-002a. U.S. Environmental Protection Agency, Research Triangle Park, NC.

Watson, R.T., Zinyowera, M., and Moss, R.H. Eds. 1995. *Climate Change, the Second Assessment Report of the Intergovernmental Panel on Climate Change, Vol. 2, Impacts, Adaptations, and Mitigations — Scientific/Technical Analysis*, Cambridge University Press, Cambridge, MA.

Wesley, M.L., Cook, D.R., Hart, R.L., and Speer, R.E. 1985. Measurements and parameterization of particulate sulfur dry deposition over grass, *J. Geophys. Res.*, 90, 2131–2143.

Zannetti, P. 1986a. A new mixed segment-puff approach for dispersion modeling, *Atmos. Environ.,* 20, 1121–1130.

Zannetti, P. 1986b, Monte-Carlo simulation of auto- and cross-correlated turbulent velocity fluctuations (MC-LAGPAR II Model), *Environ. Software,* 1, 26–30.

Zannetti, P. 1990. *Air Pollution Modeling: Theories, Computational Methods and Available Software*, Van Nostrand Reinhold, New York.

Further Information

For a comprehensive textbook on air pollution modeling, see Zannetti, 1990, above. Shorter reviews of air pollution modeling topics can be found in Zannetti, 1989, below, and Milford and Russell, 1993, above. For a comprehensive review of air pollution issues, see Seinfeld, 1986. For a complete discussion on atmospheric chemistry, see Finlayson-Pitts and Pitts, 1986, below.

Finlayson-Pitts, B.J. and Pitts, J.N., Jr. 1986. *Atmospheric Chemistry: Fundamental and Experimental Techniques*, John Wiley & Sons, New York.

Zannetti, P. 1989, Simulating short-term, short-range air quality dispersion phenomena," in *Library of Environmental Control Technology,* Vol. 2, *Air Pollution Control,* P.N. Cheremisinoff, Ed., Gulf Publishing, Houston, Chap. 5.

Park, C.S. and Tippett, D.D. “Engineering Economics and Project Management”
Mechanical Engineering Handbook
Ed. Frank Kreith
Boca Raton: CRC Press LLC, 1999

Engineering Economics and Project Management

Chan S. Park*
Auburn University

Donald D. Tippett
University of Alabama in Huntsville

* Department of Industrial & Systems Engineering, Auburn University, Auburn, AL 36849. Sections 17.1 through 17.6 based on *Contemporary Engineering Economics*, 2nd edition, by Chan S. Park, Addison-Wesley Publishing Company, Reading, MA, 1997.

17.1 Engineering Economic Decisions

Decisions made during the engineering design phase of product development determine the majority (some say 85%) of the costs of manufacturing that product. Thus, a competent engineer in the 21st century must have an understanding of the principles of economics as well as engineering. This chapter examines the most important economic concepts that should be understood by engineers.

Engineers participate in a variety of decision-making processes, ranging from manufacturing to marketing to financing decisions. They must make decisions involving materials, plant facilities, the in-house capabilities of company personnel, and the effective use of capital assets such as buildings and machinery. One of the engineer's primary tasks is to plan for the acquisition of equipment (fixed asset) that will enable the firm to design and produce products economically. These decisions are called *engineering economic decisions.*

17.2 Establishing Economic Equivalence

A typical engineering economic decision involves two dissimilar types of dollar amounts. First, there is the investment, which is usually made in a lump sum at the beginning of the project, a time that for analytical purposes is called today, or time 0. Second, there is a stream of cash benefits that are expected to result from this investment over a period of future years.

In such a fixed asset investment funds are committed today in the expectation of earning a return in the future. In the case of a bank loan, the future return takes the form of interest plus repayment of the principal. This is known as the *loan cash flow.* In the case of the fixed asset, the future return takes the form of cash generated by productive use of the asset. The representation of these future earnings along with the capital expenditures and annual expenses (such as wages, raw materials, operating costs, maintenance costs, and income taxes) is the *project cash flow.* This similarity between the loan cash flow and the project cash flow brings us an important conclusion—that is, first we need to find a way to evaluate a money series occurring at different points in time. Second, if we understand how to evaluate a loan cash flow series, we can use the same concept to evaluate the project cash flow series.

Interest: The Cost of Money

Money left in a savings account earns interest so that the balance over time is greater than the sum of the deposits. In the financial world, money itself is a commodity, and like other goods that are bought and sold, money costs money. The cost of money is established and measured by an *interest rate,* a percentage that is periodically applied and added to an amount (or varying amounts) of money over a specified length of time. When money is borrowed, the interest paid is the charge to the brrower for the use of the lender's property; when money is loaned or invested, the interest earned is the lender's gain from providing a good to another. *Interest,* then, may be defined as the cost of having money available for use.

The operation of interest reflects the fact that money has a time value. This is why amounts of interest depend on lengths of time; interest rates, for example, are typically given in terms of a percentage per year. This principle of the time value of money can be formally defined as follows: the economic value of a sum depends on when it is received. Because money has earning power over time (it can be put to work, earning more money for its owner), a dollar received today has a greater value than a dollar received at some future time.

The changes in the value of a sum of money over time can become extremely significant when we deal with large amounts of money, long periods of time, or high interest rates. For example, at a current annual interest rate of 10%, $1 million will earn $100,000 in interest in a year; thus, waiting a year to receive $1 million clearly involves a significant sacrifice. In deciding among alternative proposals, we must take into account the operation of interest and the time value of money to make valid comparisons of different amounts at various times.

The Elements of Transactions Involving Interest

Many types of transactions involve interest — for example, borrowing or investing money, purchasing machinery on credit — but certain elements are common to all of them:

1. Some initial amount of money, called the *principal* (*P*) in transactions of debt or investment
2. The *interest rate* (*i*), which measures the cost or price of money, expressed as a percentage per period of time
3. A period of time, called the *interest period* (or *compounding period*), that determines how frequently interest is calculated
4. The specified length of time that marks the duration of the transaction and thereby establishes a certain *number of interest periods* (*N*)
5. A *plan for receipts or disbursements* (A_n) that yields a particular cash flow pattern over the length of time (for example, we might have a series of equal monthly payments [A] that repay a loan)
6. A *future amount of money* (F) that results from the cumulative effects of the interest rate over a number of interest periods

Cash Flow Diagrams

It is convenient to represent problems involving the time value of money in graphic form with a cash flow diagram (see Figure 17.2.1), which represents time by a horizontal line marked off with the number of interest periods specified. The cash flows over time are represented by arrows at the relevant periods: upward arrows for positive flows (receipts) and downward arrows for negative flows (disbursements).

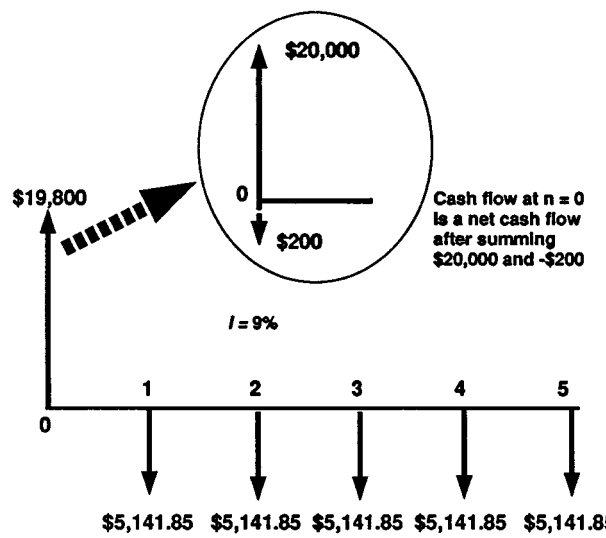


FIGURE 17.2.1 A cash flow diagram for a loan transaction — borrow $20,000 now and pay off the loan with five equal annual installments of $5,141.85. After paying $200 for the loan origination fee, the net amount of financing is $19,800. The borrowing interest rate is 9%.

End-of-Period Convention

In practice, cash flows can occur at the beginning or in the middle of an interest period, or at practically any point in time. One of the simplifying assumptions we make in engineering economic analysis is the *end-of-period convention*, which is the practice of placing all cash flow transactions at the end of an interest period. This assumption relieves us of the responsibility of dealing with the effects of interest within an interest period, which would greatly complicate our calculations.

Compound Interest

Under the compound interest scheme, the interest in each period is based on the total amount owed at the end of the previous period. This total amount includes the original principal plus the accumulated interest that has been left in the account. In this case, you are in effect increasing the deposit amount by the amount of interest earned. In general, if you deposited (invested) P dollars at interest rate i, you would have P + iP = P(1 + i) dollars at the end of one period. With the entire amount (principal and interest) reinvested at the same rate i for another period, you would have, at the end of the second period,

$$P(1+i)+i\left[P(1+i)\right]=P(1+i)(1+i)$$

$$=P(1+i)^2$$

This interest-earning process repeats, and after N periods, the total accumulated value (balance) F will grow to

$$F=P(1+i)^N \tag{17.2.1}$$

Equivalence Calculations

Economic equivalence refers to the fact that a cash flow — whether it is a single payment or a series of payments — can be said to be converted to an *equivalent* cash flow at any point in time; thus, for any sequence of cash flows, we can find an equivalent single cash flow at a given interest rate and a given time.

Equivalence calculations can be viewed as an application of the compound interest relationships developed in Equation 17.2.1. The formula developed for calculating compound interest, $F = P(1 + i)^N$, expresses the equivalence between some present amount, P, and a future amount, F, for a given interest rate, i, and a number of interest periods, N. Therefore, at the end of a 3-year investment period at 8%, $1000 will grow to

$$\$1000(1+0.08)^3=\$1259.71$$

Thus at 8% interest, $1000 received now is equivalent to $1,25l9.71 received in 3 years and we could trade $1000 now for the promise of receiving $1259.71 in 3 years. Example 17.2.1 demonstrates the application of this basic technique.

Example 17.2.1 — Equivalence

Suppose you are offered the alternative of receiving either $3000 at the end of 5 years or P dollars today. There is no question that the $3000 will be paid in full (no risk). Having no current need for the money, you would deposit the P dollars in an account that pays 8% interest. What value of P would make you indifferent in your choice between P dollars today and the promise of $3000 at the end of 5 years from now?

Discussion. Our job is to determine the present amount that is economically equivalent to $3000 in 5 years, given the investment potential of 8% per year. Note that the problem statement assumes that you would exercise your option of using the earning power of your money by depositing it. The “indifference” ascribed to you refers to economic indifference; that is, within a marketplace where 8% is the applicable interest rate, you could trade one cash flow for the other.

Solution. From Equation (17.2.1), we establish

$$\$3000=P(1+0.08)^5$$

Rearranging to solve for P,

$$P = \$3000/(1+0.08)^5 = \$2042$$

Comments. In this example, it is clear that if P is anything less than $2042, you would prefer the promise of $3000 in 5 years to P dollars today; if P were greater than $2042, you would prefer P. It is less obvious that at a lower interest rate, P must be higher to be equivalent to the future amount. For example, at i = 4%, P = $2466.

Interest Formulas

In this section is developed a series of interest formulas for use in more complex comparisons of cash flows. It classifies four major categories of cash flow transactions, develops interest formulas for them, and presents working examples of each type.

Single Cash Flow Formulas

We begin our coverage of interest formulas by considering the simplest cash flows: single cash flows. Given a present sum P invested for N interest periods at interest rate i, what sum will have accumulated at the end of the N periods? You probably noticed quickly that this description matches the case we first encountered in describing compound interest. To solve for F (the future sum) we use Equation (17.2.1):

$$F = P(1+i)^N = P(F/P,i,N)$$

Because of its origin in compound interest calculation, the factor (F/P, i, N), which is read as "find F, given P, i, and N" is known as the *single payment compound amount factor.* Like the concept of equivalence, this factor is one of the foundations of engineering economic analysis. Given this factor, all the other important interest formulas can be derived. This process of finding F is often called the compounding process. (Note the time-scale convention. The first period begins at n = 0 and ends at n = 1.) Thus, in the preceding example, where we had $F = \$1000(1.08)^3$, we can write F = $1000(F/P, 8%, 3). We can directly evaluate the equation or locate the factor value by using the 8% interest table* and finding the factor of 1.2597 in the F/P column for N = 3.

Finding present worth of a future sum is simply the reverse of compounding and is known as *discounting process.* In Equation (17.2.1), we can see that if we were to find a present sum P, given a future sum F, we simply solve for P.

$$P = F\left[\frac{1}{(1+i)^N}\right] = F(P/F,i,N) \qquad (17.2.2)$$

The factor $1/(1 + i)^N$ is known as the *single payment present worth factor* and is designated (P/F, i, N). Tables* have been constructed for the P/F factors for various values of i and N. The interest rate i and the P/F factor are also referred to as *discount rate* and *discounting factor*, respectively. Because using the interest tables is often the easiest way to solve an equation, this factor notation is included for each of the formulas derived in the following sections.

* All standard engineering economy textbooks (such as *Contemporary Engineering Economics* by C. S. Park, Addison Wesley, 1997) provide extensive sets of interest tables. Or you can obtain such interest tables on a World Wide Web site at http://www.eng.auburn.edu/-park/cee.html, which is a textbook web site for *Contemporary Engineering Economics.*

A Stream of Cash Flow Series

A common cash flow transaction involves a series of disbursements or receipts. Familiar situations such as car loans and insurance payments are examples of series payments. Payments of car loans and insurance bills typically involve identical sums paid at regular intervals. However, when there is no clear pattern of payment amounts over a series, one calls the transaction an *uneven cash-flow series*.

The present worth of any stream of payments can be found by calculating the present value of each individual payment and summing the results. Once the present worth is found, one can make other equivalence calculations, such as calculating the future worth by using the interest factors developed in the previous section.

Example 17.2.2 — Present Value of an Uneven Series by Decomposition into Single Payments

Wilson Technology, a growing machine shop, wishes to set aside money now to invest over the next 4 years in automating their customer service department. They now earn 10% on a lump sum deposited, and they wish to withdraw the money in the following increments:

Year 1: $25,000 to purchase a computer and data base software designed for customer service use
Year 2: $3000 to purchase additional hardware to accommodate anticipated growth in use of the system
Year 3: No expenses
Year 4: $5000 to purchase software upgrades

How much money must be deposited now to cover the anticipated payments over the next 4 years?

Discussion. This problem is equivalent to asking what value of P would make you indifferent in your choice between P dollars today and the future expense stream of ($25,000, $3000, $0, $5000). One way to deal with an uneven series of cash flows is to calculate the equivalent present value of each single cash flow and sum the present values to find P. In other words, the cash flow is broken into three parts as shown in Figure 17.2.2.

Solution.

$$P = \$25{,}000(P/F,10\%,1) + \$3000(P/F,10\%,2) + \$5000(P/F,10\%,4)$$

$$= \$28{,}622$$

Cash Flow Series with a Special Pattern

Whenever one can identify patterns in cash flow transactions, one may use them in developing concise expressions for computing either the present or future worth of the series. For this purpose, we will classify cash flow transactions into three categories: (1) equal cash flow series, (2) linear gradient series, and (3) geometric gradient series. To simplify the description of various interest formulas, we will use the following notation:

1. Uniform Series: Probably the most familiar category includes transactions arranged as a series of equal cash flows at regular intervals, known as an *equal-payment series* (or *uniform series*) (Figure 17.2.3a). This describes the cash flows, for example, of the common installment loan contract, which arranges for the repayment of a loan in equal periodic installments. The equal cash flow formulas deal with the equivalence relations of P, F, and A, the constant amount of the cash flows in the series.
2. Linear Gradient Series: While many transactions involve series of cash flows, the amounts are not always uniform: yet they may vary in some regular way. One common pattern of variation occurs when each cash flow in a series increases (or decreases) by a fixed amount (Figure 17.2.3b). A 5-year loan repayment plan might specify, for example, a series of annual payments that

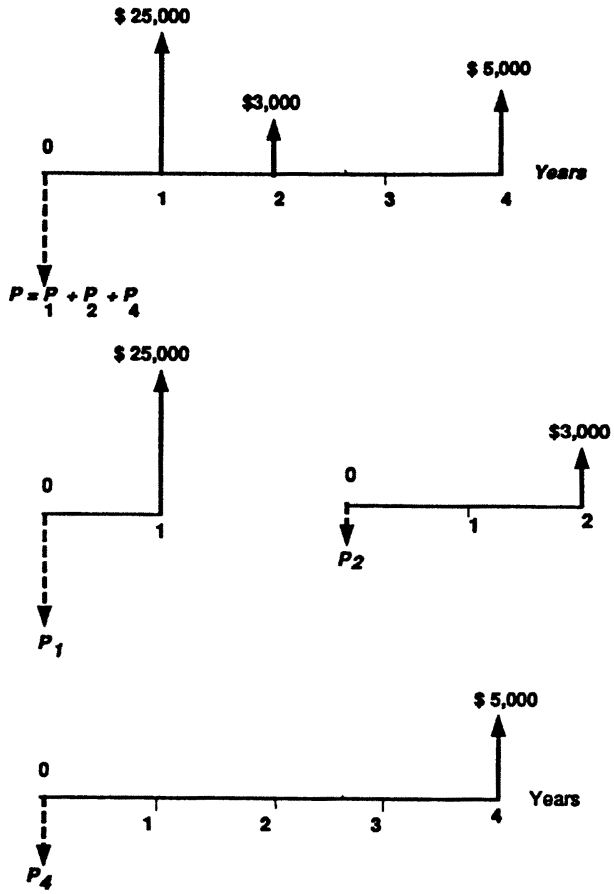


FIGURE 17.2.2 Decomposition of uneven cash flow series into three single-payment transactions. This decomposition allows us to use the single-payment present worth factor.

increased by $50 each year. We call such a cash flow pattern a *linear gradient series* because its cash flow diagram produces an ascending (or descending) straight line. In addition to P, F, and A, the formulas used in such problems involve the constant amount, G, of the change in each cash flow.

3. Geometric Gradient Series: Another kind of gradient series is formed when the series in cash flow is determined, not by some fixed amount like $50, but by some fixed *rate*, expressed as a percentage. For example, in a 5-year financial plan for a project, the cost of a particular raw material might be budgeted to increase at a rate of 4% per year. The curving gradient in the diagram of such a series suggests its name: a *geometric gradient series* (Figure 17.2.3c). In the formulas dealing with such series, the rate of change is represented by a lowercase g.

Table 17.2.1 summarizes the interest formulas and the cash flow situations in which they should be used. For example, the factor notation (F/A, i, N) represents the situation where you want to calculate the equivalent lump-sum future worth (F) for a given uniform payment series (A) over N period at interest rate i. Note that these interest formulas are applicable only when the interest (compounding) period is the same as the payment period. Also in this table we present some useful interest factor relationships. The next two examples illustrate how one might use these interest factors to determine the equivalent cash flow.

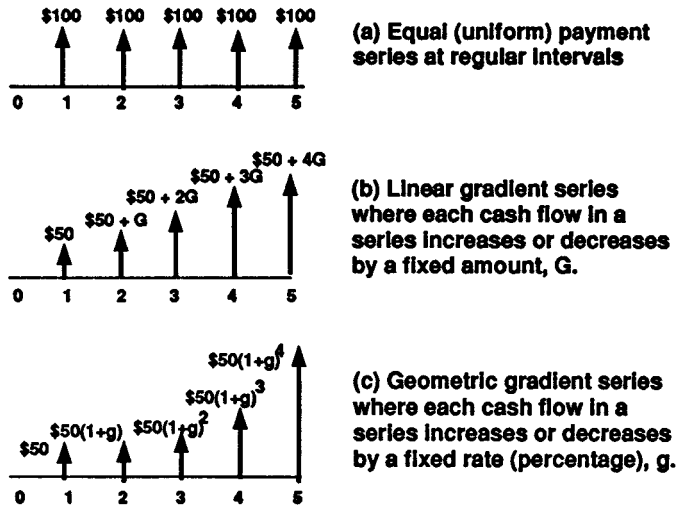


FIGURE 17.2.3 Five types of cash flows: (a) equal (uniform) payment series; (b) linear gradient series; and (c) geometric gradient series.

Example 17.2.3 — Uniform Series: Find F, Given i, A, N

Suppose you make an annual contribution of $3000 to your savings account at the end of each year for 10 years. If your savings account earns 7% interest annually, how much can be withdrawn at the end of 10 years? (See Figure 17.2.4.)

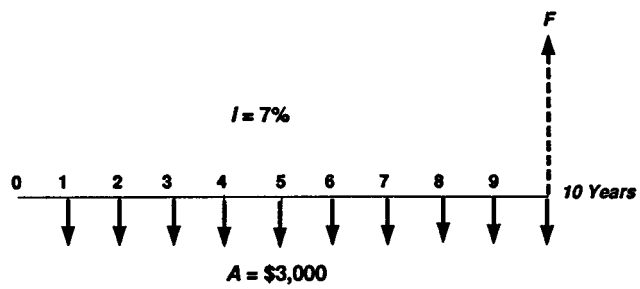


FIGURE 17.2.4 Cash flow diagram (Example 17.2.3).

Solution.

$$F = \$3000(F/A, 7\%, 10)$$

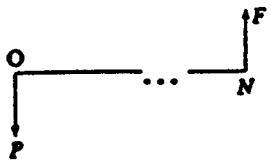
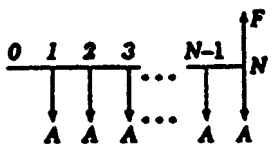
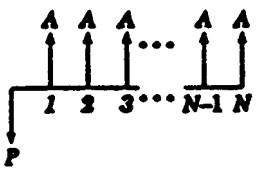
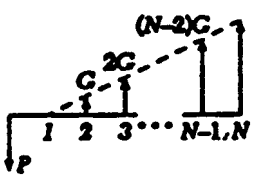
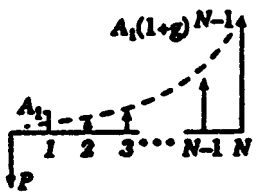
$$= \$3000(13.8164)$$

$$= \$41,449.20$$

Example 17.2.4 — Geometric Gradient: Find P, Given A_1, g, i, N

Ansell Inc., a medical device manufacturer, uses compressed air in solenoids and pressure switches in the machines to control the various mechanical movements. Over the years the manufacturing floor has changed layouts numerous times. With each new layout more piping was added to the compressed air delivery system to accommodate the new locations of the manufacturing machines. None of the extra, unused old pipe was capped or removed; thus the current compressed air delivery system is inefficient and fraught with leaks. Because of the leaks in the current system, the compressor is expected to run

TABLE 17.2.1 Summary of Discrete Compounding Formulas with Discrete Payments

Flow Type	Factor Notation	Formula	Cash Flow Diagram	Factor Relationship
Single	Compound amount ($F/P, i, N$) Present worth ($P/F, i, N$)	$F = P(1+i)^N$ $P = F(1+i)^{-N}$		$(F/P, i, N) = i(F/A, i, N) + 1$ $(P/F, i, N) = 1 - (P/A, i\ N)i$
Equal Payment Series	Compound amount ($F/A, i, N$)	$F = A\left[\frac{(1+i)^N - 1}{i}\right]$		$(A/F, i, N) = (A/P, i, N) - i$
	Sinking fund ($A/F, i, N$)	$A = F\left[\frac{i}{(1+i)^N - 1}\right]$		
	Present worth ($P/A, i, N$)	$P = A\left[\frac{(1+i)^N - 1}{i(1+i)^N}\right]$		$(A/P, i, N) = \frac{i}{1-(P/F, i, N)}$
	Capital recovery ($A/P, i, N$)	$A = P\left[\frac{i(1+i)^N}{(1+i)^N - 1}\right]$		
Gradient Series	Uniform gradient Present worth ($P/G, i, N$)	$P = C\left[\frac{(1+i)^N - iN - 1}{i^2(1+i)^N}\right]$		$(F/G, i, N) = (P/G, i, N)(F/P, i, N)$ $(A/G, i, N) = (P/G, i, N)(A/P, i, N)$
	Geometric gradient Present worth ($P/A_1, g, i, N$)	$P = \begin{cases} A_1\left[\frac{1-(1+g)^N(1+i)^{-N}}{i-g}\right] \\ \frac{NA_1}{1+i} \quad (\text{if } i = g) \end{cases}$		$(F/A_1, g, i, N) = (P/A_1, g, i, N)(F/P, i\ N)$

Adapted from Park, C.S. 1997. *Contemporary Engineering Economics*. Addison-Wesley, Reading, MA. Tables are constructed for various interest factors and you can obtain such interest tables on a World Wide Web site at http://www.eng.auburn.edu/~park/cee.html, which is a textbook web site for *Contemporary Engineering Economics*.

70% of the time that the plant is in operation during the upcoming year, which will require 260 kW/hr of electricity at a rate of $0.05/kW-hr. (The plant runs 250 days a year for 24 hr a day.) If Ansell continues to operate the current air delivery system, the compressor run time will increase by 7% per year for the next 5 years due to ever-deteriorating leaks. (After 5 years, the current system cannot meet the plant's compressed air requirement, so it has to be replaced.) If Ansell decides to replace all of the old piping now, it will cost $28,570. The compressor will still run the same number of days; however, it will run 23% less (or 70% (1 – 0.23) = 53.9% usage during the day) because of the reduced air pressure loss. If Ansell's interest rate is 12%, is it worth fixing now?

Solution.

- Step 1. Calculate the cost of power consumption of the current piping system during the first year. The power consumption is equal to:

$$\begin{aligned}\text{power cost} &= \%\text{ of day operating} \times \text{days operating per year} \times \text{hours per day} \\ &\quad \times \text{kW/hr} \times \$/\text{kW-hr} \\ &= (70\%) \times (250\ \text{days/year}) \times (24\ \text{hr/day}) \times (260\ \text{kW/hr}) \times (\$0.05/\text{kW-hr}) \\ &= \$54,440\end{aligned}$$

- Step 2. Each year the annual power cost will increase at the rate of 7% over the previous year's power cost. Then the anticipated power cost over the 5-year period is summarized in Figure 17.2.5. The equivalent present lump-sum cost at 12% for this geometric gradient series is

$$\begin{aligned}P_{Old} &= \$54,440\left(P/A_1, 7\%, 12\%, 5\right) \\ &= \$54,440\left[\frac{1-(1+0.07)^5(1+0.12)^{-5}}{0.12-0.07}\right] \\ &= \$222,283\end{aligned}$$

- Step 3. If Ansell replaces the current compressed air system with the new one, the annual power cost will be 23% less during the first year and will remain at that level over the next 5 years. The equivalent present lump-sum cost at 12% is

$$\begin{aligned}P_{New} &= \$54,440(1-0.23)\left(P/A, 12\%, 5\right) \\ &= \$41,918.80(3.6048) \\ &= \$151,108\end{aligned}$$

- Step 4. The net cost for not replacing the old system now is $71,175 (= $222,283 – $151,108). Since the new system costs only $28,570, the replacement should be made now.

Nominal and Effective Interest Rates

In all our examples in the previous section, we implicitly assumed that payments are received once a year, or annually. However, some of the most familiar financial transactions in both personal financial matters and engineering economic analysis involve nonannual payments; for example, monthly mortgage payments and daily earnings on savings accounts. Thus, if we are to compare different cash flows with different compounding periods, we need to address them on a common basis. The need to do this has led to the development of the concepts of *nominal interest rate* and *effective interest rate.*

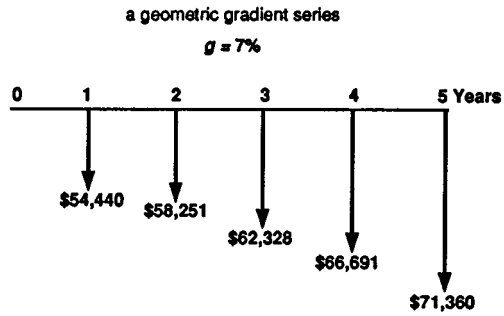


FIGURE 17.2.5 Expected power expenditure over the next 5 years due to deteriorating leaks if no repair is performed (Example 17.2.4).

Nominal Interest Rate

Even if a financial institution uses a unit of time other than a year — a month or quarter, for instance — in calculating interest payments, it usually quotes the interest rate on an annual basis. Many banks, for example, state the interest arrangement for credit cards in this way: "18% compounded monthly." We say 18% is the *nominal interest rate* or *annual percentage rate* (APR), and the compounding frequency is monthly (12). To obtain the interest rate per compounding period, we divide 18% by 12 to obtain 1.5% per month. Therefore, the credit card statement above means that the bank will charge 1.5% interest on unpaid balance for each month.

Although the annual percentage rate, or APR, is commonly used by financial institutions and is familiar to customers, when compounding takes place more frequently than annually, the APR does not explain precisely the amount of interest that will accumulate in a year. To explain the true effect of more frequent compounding on annual interest amounts, we need to introduce the term effective interest rate.

Effective Annual Interest Rate

The *effective interest rate* is the only one that truly represents the interest earned in a year or some other time period. For instance, in our credit card example, the bank will charge 1.5% interest on unpaid balance at the end of each month. Therefore, the 1.5% rate represents an effective interest rate — on a monthly basis, it is the rate that predicts the actual interest payment on your outstanding credit card balance.

Suppose you purchase an appliance on credit at 18% compounded monthly. Unless you pay off the entire amount within a grace period (let's say, a month), any unpaid balance (P) left for a year period would grow to

$$\begin{aligned} F &= P(1+i)^N \\ &= P(1+0.015)^{12} \\ &= 1.1956P \end{aligned}$$

This implies that for each dollar borrowed for 1 year, you owe $1.1956 at the end of the year, including the principal and interest. For each dollar borrowed, you pay an equivalent annual interest of 19.56 cents. In terms of an effective annual interest rate (i_a), we can rewrite the interest payment as a percentage of the principal amount

$$i_a = (1+0.015)^{12} - 1 = 0.1956, \text{ or } 19.56\%$$

Thus, the effective annual interest rate is 19.56%.

Clearly, compounding more frequently increases the amount of interest paid for the year at the same nominal interest rate. We can generalize the result to compute the effective interest rate for *any time duration*. As you will see later, we normally compute the effective interest rate based on payment (transaction) period. For example, cash flow transactions occur quarterly but interest rate is compounded monthly. This quarterly conversion allows us to use the interest formulas in Table 17.2.1. To consider this, we may define the effective interest rate for a given payment period as

$$\begin{aligned} i &= (1 + r/M)^C - 1 \\ &= (1 + r/CK)^C - 1 \end{aligned} \tag{17.2.3}$$

where

M = the number of interest periods per year
C = the number of interest periods per payment period
K = the number of payment periods per year

When M = 1, we have the special case of annual compounding. Substituting M = 1 into Equation (17.2.3), we find it reduces to $i_a = r$. That is, when compounding takes place once annually,* effective interest is equal to nominal interest. Thus, in all our earlier examples, where we considered only annual interest, we were by definition using effective interest rates.

Example 17.2.5 — Calculating Auto Loan Payments

Suppose you want to buy a car priced $22,678.95. The car dealer is willing to offer a financing package with 8.5% annual percentage rate over 48 months. You can afford to make a down payment of $2678.95, so the net amount to be financed is $20,000. What would be the monthly payment?

Solution. The ad does not specify a compounding period, but in automobile financing the interest and the payment periods are almost always both monthly. Thus, the 8.5% APR means 8.5% compounded monthly. In this situation, we can easily compute the monthly payment using the *capital recovery factor* in Table 17.2.1:

$$i = 8.5\%/12 = 0.7083\% \text{ per month}$$

$$N = (12)(4) = 48 \text{ months}$$

$$A = \$20,000(A/P, 0.7083\%, 48) = \$492.97$$

* For an extreme case, we could consider a continuous compounding. As the number of compounding periods (M) becomes very large, the interest rate per compounding period (r/M) becomes very small. As M approaches infinity and r/M approaches zero, we approximate the situation of *continuous compounding*.

$$i = e^{r/K} - 1$$

To calculate the effective annual interest rate for continuous compounding, we set K equal to 1, resulting in:

$$i_a = e^r - 1$$

As an example, the effective annual interest rate for a nominal interest rate of 12% compounded continuously is $i_a = e^{0.12} - 1 = 12.7497\%$.

Example 17.2.6 — Compounding More Frequent Than Payments

Suppose you make equal quarterly deposits of $1000 into a fund that pays interest at a rate of 12% compounded monthly. Find the balance at the end of year 2 (Figure 17.2.6).

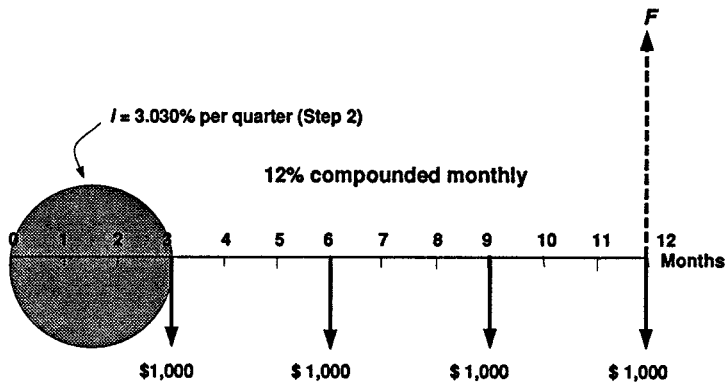


FIGURE 17.2.6 Quarterly deposits with monthly compounding (Example 17.2.6).

Solution. We follow the procedure for noncomparable compounding and payment periods described above.

1. Identify the parameter values for M, K, and C:

$$M = 12 \text{ compounding periods per year}$$

$$K = 4 \text{ payment periods per year}$$

$$C = 3 \text{ interest periods per payment period}$$

2. Use Equation (17.2.3) to compute effective interest per quarter.

$$\text{i} = \left(1 + 0.12/12\right)^3 - 1$$

$$= 3.030\% \text{ per quarter}$$

3. Find the total number of payment periods, N.

$$N = K(\text{number of years}) = 4(2) = 8 \text{ quarters}$$

4. Use i and N in the (F/A, i, N) factor from Table 17.2.1:

$$F = \$1000\left(F/A, 3.030\%, 8\right) = \$8901.81$$

Loss of Purchasing Power

It is important to differentiate between the time value of money as we used it in the previous section and the effects of inflation. The notion that a sum is worth more the earlier it is received can refer to its earning potential over time, to decreases in its value due to inflation over time, or to both. Historically, the general economy has usually fluctuated in such a way that it experiences *inflation*, a loss in the purchasing power of money over time. Inflation means that the cost of an item tends to increase over time or, to put it another way, the same dollar amount buys less of an item over time. *Deflation* is the

opposite of inflation (negative inflation), in that prices decrease over time and hence a specified dollar amount gains in purchasing power. In economic equivalence calculations, we need to consider the change of purchasing power along with the earning power.

The Average Inflation Rate

To account for the effect of varying yearly inflation rates over a period of several years, we can compute a single rate that represents an *average inflation rate.* Since each individual year's inflation rate is based on the pervious year's rate, they have a compounding effect. As an example, suppose we want to calculate the average inflation rate for a 2-year period for a typical item. The first year's inflation rate is 4% and the second year's is 8%, using a base price index of 100.

- Step 1. We find the price at the end of the second year by the process of compounding:

$$\underbrace{100(1+0.04)}_{\qquad}\underbrace{(1+0.08)}_{\qquad\qquad\qquad} = 112.32$$

- Step 2. To find the average inflation rate f over a 2-year period, we establish the following equivalence equation:

$$100(1+f)^2 = 112.32 \leftarrow 100(F/P, f, 2) = 112.32$$

  Solving for f yields

$$f = 5.98\%$$

We can say that the price increases in the last 2 years are equivalent to an average annual percentage rate of 5.98% per year. In other words, our purchasing power decreases at the annual rate of 5.98% over the previous year's dollars. If the average inflation rate is calculated based on the *consumer price index* (CPI), it is known as a *general inflation rate* ($\bar{f}$).

Actual vs. Constant Dollars

To introduce the effect of inflation into our economic analysis, we need to define several inflation-related terms.*

- Actual (current) dollars (A_n): Estimates of future cash flows for year n which take into account any anticipated changes in amount due to inflationary or deflationary effects. Usually these amounts are determined by applying an inflation rate to base year dollar estimates.
- Constant (real) dollars (A'_n): Dollars of constant purchasing power independent of the passage of time. In situations where inflationary effects have been assumed when cash flows were estimated, those estimates can be converted to constant dollars (base year dollars) by adjustment using some readily accepted *general inflation rate*. We will assume that the base year is always time 0 unless we specify otherwise.

* Based on the ANSI Z94 Standards Committee on Industrial Engineering Terminology. 1988. *The Engineering Economist.* Vol. 33(2): 145–171.

Equivalence Calculation under Inflation

In previous sections, our equivalence analyses have taken into consideration changes in the *earning power* of money — that is, interest effects. To factor in changes in *purchasing power* as well — that is, inflation — we may use either (1) constant dollar analysis or (2) actual dollar analysis. Either method will produce the same solution; however, each uses a different interest rate and procedure.

There are two types of interest rate for equivalence calculation: (1) the market interest rate and (2) the inflation-free interest rate. The interest rate that is applicable depends on the assumptions used in estimating the cash flow.

- Market interest rate (i): This interest rate takes into account the combined effects of the earning value of capital (earning power) and any anticipated inflation or deflation (purchasing power). Virtually all interest rates stated by financial institutions for loans and savings accounts are market interest rates. Most firms use a market interest rate (also known as *inflation-adjusted rate of return* [*discount rate*]) in evaluating their investment projects.
- Inflation-free interest rate (i′): An estimate of the true earning power of money when inflation effects have been removed. This rate is commonly known as *real interest rate* and can be computed if the market interest rate and inflation rate are known.

In calculating any cash flow equivalence, we need to identify the nature of project cash flows. There are three common cases:

Case 1. All cash flow elements are estimated in constant dollars. Then, to find the equivalent present value of a cash flow series in constant dollars, use the inflation-free interest rate.

Case 2. All cash flow elements are estimated in actual dollars. Then, use the market interest rate to find the equivalent worth of the cash flow series in actual dollars.

Case 3. Some of the cash flow elements are estimated in constant dollars and others are estimated in actual dollars. In this situation, we simply convert all cash flow elements into one type — either constant or actual dollars. Then we proceed with either constant-dollar analysis for case 1 or actual-dollar analysis for case 2.

Removing the effect of inflation by deflating the actual dollars with $\bar{f}$ and finding the equivalent worth of these constant dollars by using the inflation-free interest rate can be greatly streamlined by the efficiency of the *adjusted-discount method*, which performs deflation and discounting in one step. Mathematically we can combine this two-step procedure into one by

$$i = i' + \bar{f} + i'\bar{f} \tag{17.2.4}$$

This implies that the market interest rate is a function of two terms, i′, $\bar{f}$. Note that if there is no inflationary effect, the two interest rates are the same ($\bar{f} = 0 \rightarrow$ i = i′). As either i′ or $\bar{f}$ increases, i also increases. For example, we can easily observe that when prices are increasing due to inflation, bond rates climb, because lenders (that is anyone who invests in a money-market fund, bond, or certificate of deposit) demand higher rates to protect themselves against the eroding value of their dollars. If inflation were at 3%, you might be satisfied with an interest rate of 7% on a bond because your return would more than beat inflation. If inflation were running at 10%, however, you would not buy a 7% bond; you might insist instead on a return of at least 14%. On the other hand, when prices are coming down, or at least are stable, lenders do not fear the loss of purchasing power with the loans they make, so they are satisfied to lend at lower interest rates.

17.3 Measures of Project Worth

This section shows how to compare alternatives on equal basis and select the wisest alternative from an economic standpoint. The three common measures based on cash flow equivalence are (1) equivalent present worth, (2) equivalent annual worth, and (3) rate of return. The present worth represents a measure of future cash flow relative to the time point "now" with provisions that account for earning opportunities. Annual worth is a measure of the cash flow in terms of the equivalent equal payments on an annual basis. The third measure is based on *yield* or percentage.

Describing Project Cash Flows

When a company purchases a fixed asset such as equipment, it makes an investment. The company commits funds today in the expectation of earning a return on those funds in the future. Such an investment is similar to that made by a bank when it lends money. For the bank loan, the future cash flow consists of interest plus repayment of the principal. For the fixed asset, the future return is in the form of cash flows from the profitable use of the asset. In evaluating a capital investment, we are concerned only with those cash flows that result directly from the investment. These cash flows, called *differential* or *incremental cash flows*, represent the change in the firm's total cash flow that occurs as a direct result of the investment.

We must also recognize that one of the most important parts of the capital budgeting process is the estimation of the relevant cash flows. For all examples in this section, however, net cash flows can be viewed as before-tax values or after-tax values for which tax effects have been recalculated. Since some organizations (e.g., governments and nonprofit organizations) are not taxable, the before-tax situation can be a valid base for that type of economic evaluation. This view will allow us to focus on our main area of concern, the economic evaluation of an investment project. The procedures for determining after-tax net cash flows in taxable situations are developed in Section 17.4.

Example 17.3.1 — Identifying Project Cash Flows

Merco Inc., a machinery builder in Louisville, KY, is considering making an investment of $1,250,000 in a complete structural-beam-fabrication system. The increased productivity resulting from the installation of the drilling system is central to the justification. Merco estimates the following figures as a basis for calculating productivity:

- Increased fabricated steel production: 2000 tons/year
- Average sales price/ton fabricated steel: $2566.50/ton
- Labor rate: $10.50/hr
- Tons of steel produced in a year: 15,000 tons
- Cost of steel per ton (2205 lb): $1950/ton
- Number of workers on layout, holemaking, sawing, and material handling: 17
- Additional maintenance cost: $128,500 per year

With the cost of steel at $1950/ton and the direct labor cost of fabricating 1 lb at 10 cents, the cost of producing a ton of fabricated steel is about $2170.50. With a selling price of $2566.50/ton, the resulting contribution to overhead and profit becomes $396/ton. Assuming that Merco will be able to sustain an increased production of 2000 tons per year by purchasing the system, the projected additional contribution has been estimated to be 2000 tons $\times$ $396 = $792,000.

Since the drilling system has the capacity to fabricate the full range of structural steel, two workers can run the system, one on the saw and the other on the drill. A third operator is required as a crane operator for loading and unloading materials. Merco estimates that to do the equivalent work of these

three workers with conventional manufacture requires, on the average, an additional 14 people for centerpunching, holemaking with radial or magnetic drill, and material handling. This translates into a labor savings in the amount of $294,000 per year ($10.50 × 40 hr/week × 50 weeks/year × 14). The system can last for 15 years with an estimated after-tax salvage value of $80,000. The expected annual corporate income taxes would amount to $226,000. Determine the net cash flow from undertaking the investment. Determine the net cash flows from the project over the service life.

Solution. The net investment cost as well as savings are as follows:

- Cash inflows:
  - Increased annual revenue: $792,000
  - Projected annual net savings in labor: $294,000
  - Projected after-tax salvage value at the end of year 15: $80,000
- Cash outflows:
  - Project investment cost: $1,250,000
  - Projected increase in annual maintenance cost: $128,500
  - Projected increase in corporate income taxes: $226,000

Now we are ready to summarize a cash flow table as follows:

Year	Cash Inflows	Cash Outflows	Net Cash Flows
0	0	$1,250,000	–$1,250,000
1	1,086,000	354,500	731,500
2	1,086,000	354,500	731,500
⋮	⋮	⋮	⋮
15	1,086,000 + 80,000	354,500	811,500

The project's cash flow diagram is shown in Figure 17.3.1.

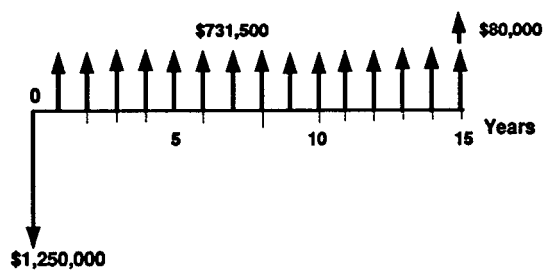


FIGURE 17.3.1 Cash flow diagram (Example 17.3.1).

Assuming these cost savings and cash flow estimates are correct, should management give the go-ahead for installation of the system? If management has decided not to install the fabrication system, what do they do with the $1,250,000 (assuming they have it in the first place)? The company could buy $1,250,000 of Treasury bonds. Or it could invest the amount in other cost-saving projects. How would the company compare cash flows that differ both in timing and amount for the alternatives it is considering? This is an extremely important question because virtually every engineering investment decision involves a comparison of alternatives. These are the types of questions this section is designed to help you answer.

Present Worth Analysis

Until the 1950s, the paycheck method* was widely used as a means of making investment decisions. As flows in this method were recognized, however, business people began to search for methods to improve project evaluations. This led to the development of discounted cash flow techniques (DCF), which take into account the time value of money. One of the DCFs is the net present worth method (NPW). A capital investment problem is essentially one of determining whether the anticipated cash inflows from a proposed project are sufficiently attractive to invest funds in the project. In developing the NPW criterion, we will use the concept of cash flow equivalence discussed in Section 17.2. Usually, the most convenient point at which to calculate the equivalent values is often time 0. Under the NPW criterion, the present worth of all cash inflows is compared against the present worth of all cash outflows that are associated with an investment project. The difference between the present worth of these cash flows, called the *net present worth* (NPW), determines whether or not the project is an acceptable investment. When two or more projects are under consideration. NPW analysis further allows us to select the best project by comparing their NPW figures.

We will first summarize the basic procedure for applying the present worth criterion to a typical investment project.

- Determine the interest rate that the firm wishes to earn on its investments. This represents an interest rate at which the firm can always invest the money in its *investment pool*. We often refer to this interest rate as either a *required rate of return* or a *minimum attractive rate of return* (MARR). Usually this selection will be a policy decision by top management. It is possible for the MARR to change over the life of a project, but for now we will use a single rate of interest in calculating NPW.
- Estimate the service life of the project.**
- Determine the net cash flows (net cash flow = cash inflow – cash outflow).
- Find the present worth of each net cash flow at the MARR. Add up these present worth figures; their sum is defined as the project's NPW.
- Here, a positive NPW means the equivalent worth of inflows are greater than the equivalent worth of outflows, so project makes a profit. Therefore, if the PW(i) is positive for a single project, the project should be accepted; if negative, it should be rejected. The decision rule is

$$\text{If } PW(i) > 0, \text{ accept the investment}$$

$$\text{If } PW(i) = 0, \text{ remain indifferent}$$

$$\text{If } PW(i) < 0, \text{ reject the investment}$$

Note that the decision rule is for a single project evaluation where you can estimate the revenues as well as costs associated with the project. As you will find later, when you are comparing alternatives with

* One of the primary concerns of most business people is whether and when the money invested in a project can be recovered. The *payback method* screens projects on the basis of how long it takes for net receipts to equal investment outlays. A common standard used in determining whether or not to pursue a project is that no project may be considered unless its payback period is shorter than some specified period of time. If the payback period is within the acceptable range, a formal project evaluation (such as the present worth analysis) may begin. It is important to remember that payback screening is not an *end* itself, but rather a method of screening out certain obvious unacceptable investment alternatives before progressing to an analysis of potentially acceptable ones. But the much-used payback method of equipment screening has a number of serious drawbacks. The principal objection to the payback method is its failure to measure profitability; that is, there is no "profit" made during the payback period. Simply measuring how long it will take to recover the initial investment outlay contributes little to gauging the earning power of a project.

the same revenues, you can compare them based on the cost only. In this situation (because you are minimizing costs, rather than maximizing profits), you should accept the project that results in smallest, or least negative, NPW.

Example 17.3.2 — Net Present Worth

Consider the investment cash flows associated with the metal fabrication project in Example 17.3.1. If the firm's MARR is 15%, compute the NPW of this project. Is this project acceptable?

Solution. Since the fabrication project requires an initial investment of $1,250,000 at n = 0 followed by the 15 equal annual savings of $731,500, and $80,000 salvage value at the end of 15 years, we can easily determine the NPW as follows:

$$PW(15\%)_{outflow} = \$1,250,000$$

$$\begin{aligned} PW(15\%)_{inflow} &= \$731,500(P/A,15\%,15) + \$80,000(P/F,15\%,15) \\ &= \$4,284,259 \end{aligned}$$

Then, the NPW of the project is

$$\begin{aligned} PW(15\%) &= PW(15\%)_{in\ flow} - PW(15\%)_{out\ flow} \\ &= \$4,284,259 - \$1,250,000 \\ &= \$3,034,259 \end{aligned}$$

Since PW (15%) > 0, the project would be acceptable.

Annual Equivalent Method

The annual equivalent worth (AE) criterion is a basis for measuring investment worth by determining equal payments on an annual basis. Knowing that we can convert any lump-sum cash amount into a series of equal annual payments, we may first find the NPW for the original series and then multiply the NPW by the capital recovery factor:

** Another special case of the PW criterion is useful when the life of a proposed project is *perpetual* or the planning horizon is extremely long. The process of computing the PW cost for this infinite series is referred to as the *capitalization* of project cost. The cost is known as the *capitalized cost.* It represents the amount of money that must be invested today to yield a certain return A at the end of each and every period forever, assuming an interest rate of i. Observe the limit of the uniform series present worth factor as N approaches infinity:

$$\lim_{N\to\infty}(P/A,i,N) = \lim_{N\to\infty}\left[\frac{(1+i)^N - 1}{i(1+i)^N}\right] = \frac{1}{i}$$

Thus, it follows that

$$PW(i) = A(P/A,i,N\to\infty) = \frac{A}{i} \tag{17.5}$$

Another way of looking at this, PW(i) dollars today, is to ask what constant income stream could be generated by this in perpetuity. Clearly, the answer is A = iPW(i). If withdrawals were greater than A, they could be eating into the principal, which would eventually reduce to 0.

$$AE(i) = PW(i)(A/P, i, N) \tag{17.3.1}$$

The accept-reject decision rule for a single *revenue* project is

If $AE(i) > 0$, accept the investment

If $AE(i) = 0$, remain indifferent

If $AE(i) < 0$, reject the investment

Notice that the factor (A/P, i, N) in Table 17.2.1 is positive for $-1 < i < \infty$. This indicates that the AE(i) value will be positive if and only if PW(i) is positive. In other words, accepting a project that has a positive AE(i) value is equivalent to accepting a project that has a positive PW(i) value. Therefore, the AE criterion should provide a basis for evaluating a project that is consistent with the NPW criterion.

As with the present worth analysis, when you are comparing mutually exclusive *service* projects whose revenues are the same, you may compare them based on *cost* only. In this situation, you will select the alternative with the minimum annual equivalent cost (or least negative annual equivalent worth).

Unit Profit/Cost Calculation

There are many situations in which we want to know the unit profit (or cost) of operating an asset. A general procedure to obtain such a unit profit or cost figure involves the following two steps:

- Determine the number of units to be produced (or serviced) each year over the life of the asset.
- Identify the cash flow series associated with the production or service over the life of the asset.
- Calculate the net present worth of the project cash flow series at a given interest rate and then determine the equivalent annual worth.
- Divide the equivalent annual worth by the number of units to be produced or serviced during each year. When you have the number of units varying each year, you may need to convert them into equivalent annual units.

To illustrate the procedure, we will consider Example 17.3.3, where the annual equivalent concept can be useful in estimating the savings per machine hour for a proposed machine acquisition.

Example 17.3.3 — Unit Profit per Machine Hour

Tiger Machine Tool Company is considering the proposed acquisition of a new metal-cutting machine. The required initial investment of $75,000 and the projected cash benefits and annual operating hours over the 3-year project life are as follows.

End of Year	Net Cash Flow	Operating Hours
0	–$75,000	
1	24,400	2,000
2	27,340	2,000
3	55,760	2,000

Compute the equivalent savings per machine hour at i = 15%.

Solution. Bringing each flow to its equivalent at time zero, we find

$$PW(15\%) = -\$75,000 + \$24,400(P/F, 15\%, 1) + \$27,340(P/F, 15\%, 2) + \$55,760(P/F, 15\%, 3)$$

$$= \$3553$$

Since the project results in a surplus of $3553, the project would be acceptable. We first compute the annual equivalent savings from the use of the machine. Since we already know the NPW of the project, we obtain the AE by

$$AE(15\%) = \$3553(A/P,15\%,3) = \$1556$$

With an annual usage of 2000 hr, the equivalent savings per machine hour would be

$$\text{Savings per machine hour} = \$1556/\$2000 \text{ hr} = \$0.78/\text{hr}$$

Comments. Note that we cannot simply divide the NPW amount ($3553) by the total number of machine hours over the 3-year period (6000 hr), or $0.59/hr. This $0.59 figure represents the *instant savings* in present worth for each hourly use of the equipment, but does not consider the time over which the savings occur. Once we have the annual equivalent worth, we can divide by the desired time unit if the compounding period is 1 year. If the compounding period is shorter, then the equivalent worth should be calculated for the compounding period.

Rate of Return Analysis

Along with the NPW and AE, the third primary measure of investment worth is based on yield, known as *rate of return*. The NPW measure is easy to calculate and apply. Nevertheless, many engineers and financial managers prefer rate of return analysis to the NPW method because they find it intuitively more appealing to analyze investments in terms of percentage rates of return than in dollars of NPW.

Internal Rate of Return

Many different terms refer to rate of return, including yield (that is, the yield to maturity, commonly used in bond valuation), internal rate of return, and marginal efficiency of capital. In this section, we will define *internal rate of return* as the break-even interest rate, i^*, which equates the present worth of a project's cash outflows to the present worth of its cash inflows, or

$$\begin{aligned} PW(i^*) &= PW_{\text{cash inflows}} - PW_{\text{cash outflows}} \\ &= 0 \end{aligned}$$

Note that the NPW expression is equivalent to

$$PW(i^*) = \frac{A_0}{(1+i^*)^0} + \frac{A_1}{(1+i^*)^1} + \cdots + \frac{A_N}{(1+i^*)^N} = 0 \qquad (17.3.2)$$

Here we know the value of A_n for all n, but not the value of i^*. Since it is the only unknown, we can solve for i^*. There will inevitably be N values of i^* that satisfy this equation. In most project cash flows you would be able to find a unique positive i^* that satisfies Equation (17.3.2). However, you may encounter some cash flow that cannot be solved for a single rate of return greater than –100%. By the nature of the NPW function in Equation (17.3.2), it is certainly possible to have more than one rate of return for a certain type of cash flow.* (For some cash flows, we may not find any rate of return at all.)

Finding the IRR

We don't need laborious manual calculations to find i^*. Many financial calculators have built-in functions for calculating i^*. It is also worth noting here that many spreadsheet packages have i^* functions, which solve Equation (17.3.2) very rapidly. This is normally done by entering the cash flows through a computer

keyboard or by reading a cash flow data file. As an alternative, you could try the trial-and-error method to locate the break-even interest that makes the net present worth equal to zero.

Accept/Reject Decision Rules

Why are we interested in finding the particular interest rate that equates a project's cost with the present worth of its receipts? Again, we may easily answer this by examining Figure 17.3.2. In this figure, we notice two important characteristics of the NPW profile. First, as we compute the project's PW(i) at a varying interest rate (i), we see that the NPW becomes positive for $i < i^*$, indicating that the project would be acceptable under the PW analysis for those values of i. Second, the NPW becomes negative for $i > i^*$, indicating that the project is unacceptable for those values of i. Therefore, the i^* serves as a *break-even* interest rate. By knowing this break-even rate, we will be able to make an accept/reject decision that is consistent with the NPW analysis.

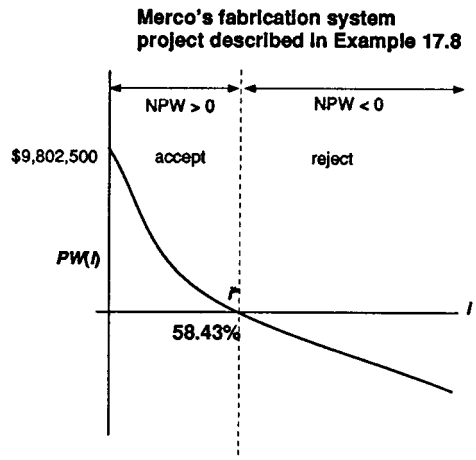


FIGURE 17.3.2 A net present worth profile for the cash flow series given in Figure 17.3.1 at varying interest rates. The project breaks even at 58.43% so that the NPW will be positive as long as the discount rate is less than 58.43%.

At the MARR the company will more than break even. Thus, the IRR becomes a useful gauge against which to judge project acceptability, and the decision rule for a simple project is

---

* When applied to projects that require investments at the outset followed by a series of cash inflows (or a simple project), the i^* provides an unambiguous criterion for measuring profitability. However, when multiple rates of return occur, none of them is an accurate portrayal of project acceptability or profitability. Clearly, then, we should place a high priority on discovering this situation early in our analysis of a project's cash flows. The quickest way to predict multiple i^*s is to generate a NPW profile and check to see if it crosses the horizontal axis more than once.

In addition to the NPW profile, there are good — although somewhat more complex — analytical methods for predicting multiple i^*s. Perhaps more importantly, there is a good method, which uses a *cost of capital*, of refining our analysis when we do discover multiple i^*s. Use of a cost of capital allows us to calculate a single accurate rate of return (also known as *return on invested capital*); it is covered in *Contemporary Engineering Economics*, C.S. Park, Addison-Wesley, 1997. If you choose to avoid these more complex applications of rate-of-return techniques, you must at a minimum be able to predict multiple i^*s via the NPW profile and, when they occur, select an alternative method such as NPW or AE analysis for determining project acceptability.

If IRR > MARR, accept the project

If IRR = MARR, remain indifferent

If IRR < MARR, reject the project

Note that this decision rule is designed to be applied for a single project evaluation. When we have to compare mutually exclusive investment projects, we need to apply the incremental analysis, as we shall see in a later section.

Example 17.3.4 — Rate of Return Analysis

Reconsider the fabrication investment project in Example 17.3.1. (a) What is the projected IRR on this fabrication investment? (b) If Merco's MARR is known to be 15%, is this investment justifiable?

Solution.

- (a) The net present worth expression as a function of interest rate (i) is

$$PW(i) = -\$1,250,000 + \$731,500(P/A,i,15) + \$80,000(P/F,i,15)$$
$$= 0$$

  Using Excel's financial function (IRR), we find the IRR to be 58.43%. (See Figure 17.3.2.) Merco will recover the initial investment fully and also earn 58.43% interest on its invested capital.
- (b) If Merco does not undertake the project, the $1,250,000 would remain in the firm's investment pool and continue to earn only 15% interest. The IRR figure far exceeds the Merco's MARR, indicating that the fabrication system project is an economically attractive one. Merco's management believes that, over a broad base of structural products, there is no doubt that the installation of its fabricating system would result in a significant savings, even after considering some potential deviations from the estimates used in the analysis.

Mutually Exclusive Alternatives

Until now, we have considered situations in which only one project was under consideration, and we were determining whether to pursue it, based on whether its present worth or rate of return met our MARR requirements. We were making an accept or reject decision about a *single* project.

In the real world of engineering practice, however, it is more typical for us to have two or more choices of projects for accomplishing a business objective. *Mutually exclusive* means that any one of several alternatives will fulfill the same need and that selecting one alternative means that the others will be excluded.

Analysis Period

The *analysis period* is the time span over which the economic effects of an investment will be evaluated. The analysis period may also be called the *study period* or *planning horizon*. The length of the analysis period may be determined in several ways: it may be a predetermined amount of time set by company policy, or it may be either implied or explicit in the need the company is trying to fulfill. In either of these situations, we consider the analysis period to be a *required service period*. When no required service period is stated at the outset, the analyst must choose an appropriate analysis period over which to study the alternative investment projects. In such a case, one convenient choice of analysis period is the period of useful life of the investment project.

When useful life of the investment project does not match the analysis or required service period, we must make adjustments in our analysis. A further complication, when we are considering two or more mutually exclusive projects, is that the investments themselves may have differing useful lives. We must

compare projects with different useful lives over an *equal time span*, which may require further adjustments in our analysis.

Analysis Period Equals Project Lives

Let's begin our analysis with the simplest situation where the project lives equal the analysis period. In this case, we compute the NPW for each project and select the one with highest NPW for revenue projects or least negative NPW for service projects. Example 17.3.5 will illustrate this point.

Example 17.3.5 — Two Mutually Exclusive Alternatives

A pilot wants to start her own company to airlift goods to the Commonwealth of Independent States (formerly the U.S.S.R.) during their transition to a free-market economy. To economize the start-up business, she decides to purchase only one plane and fly it herself. She has two mutually exclusive options: an old aircraft (A1) or a new jet (A2) with which she expects to have higher purchase costs, but higher revenues as well because of its larger payload. In either case, she expects to fold up business in 3 years because of competition from larger companies. The cash flows for the two mutually exclusive alternatives are given in thousand dollars:

n	A1	A2
0	–3,000	–$12,000
1	1,350	4,200
2	1,800	6,225
3	1,500	6,330

Assuming that there is no do-nothing alternative, which project would she select at MARR = 10%?

Solution. Since the required service period is 3 years, we should select the analysis period of 3 years. Since the analysis period coincides with the project lives, we simply compute the NPW value for each option. The equivalent NPW figures at i = 10% would be as follows:

- For A1:

$$PW(15\%)_{A1} = -\$3000 + \$1350(P/F,10\%,1) + \$1800(P/F,10\%,2) + \$1500(P/F,10\%,3)$$
$$= \$842$$

- For A2:

$$PW(15\%)_{A2} = -\$12,000 + \$4200(P/F,10\%,1) + \$6225(P/F,10\%,2) + \$6330(P/F,10\%,3)$$
$$= \$1719$$

Clearly, A2 is the most economical option.

Project Lives Differ from a Specified Analysis Period

Often project lives do not match the required analysis period and/or do not match each other. For example, two machines may perform exactly the same function, but one lasts longer than the other and both of them last longer than the analysis period for which they are being considered. We are then left with some unused portion of the equipment, which we include as salvage value in our analysis. Salvage value is the amount of money for which the equipment could be sold after its service or the dollar measure of its remaining usefulness.

When project lives are shorter than the required service period, we must consider how, at the end of the project lives, we will satisfy the rest of the required service period. Replacement projects — additional

projects to be implemented when the initial project has reached the limits of its useful life — are needed in such a case. Sufficient replacement projects must be analyzed to match or exceed the required service period.

To simplify our analysis, we sometimes assume that the replacement project will be exactly the same as the initial project, with the same corresponding costs and benefits. However, this assumption is not necessary. For example, depending on our forecasting skills, we may decide that a different kind of technology — in the form of equipment, materials, or processes — is a preferable replacement. Whether we select exactly the same alternative or a new technology as the replacement project, we are ultimately likely to have some unused portion of the equipment to consider as salvage value at the end of the required service period. On the other hand, if a required service period is relatively short, we may decide to lease the necessary equipment or subcontract the remaining work for the duration of the analysis period. In this case, we can probably exactly match our analysis period and not worry about salvage values.

Example 17.3.6 — Present Worth Comparison — Project Lives Shorter Than Analysis Period

The Smith Novelty Company, a mail-order firm, wants to install an automatic mailing system to handle product announcements and invoices. The firm has a choice between two different types of machines. The two machines are designed differently but have identical capacities and do exactly the same job. The $12,500 semiautomatic model A will last 3 years with a salvage value of $2000, while the fully automatic model B will cost $15,000 and last 4 years with a salvage value of $1500. The expected cash flows for the two machines including maintenance, salvage value, and tax effects are as follows:

n	Model A	Model B
0	–12,500	–$15,000
1	–5,000	–4,000
2	–5,000	–4,000
3	–5,000 + 2,000	–4,000
4		–4,000 + 1,500
5		

As business grows to a certain level, neither of the models can handle the expanded volume at the end of year 5. If that happens, a fully computerized mail-order system will need to be installed to handle the increased business volume. With this scenario, which model should the firm select at MARR = 15%?

Solution. Since both models have a shorter life than the required service period (5 years), we need to make an explicit assumption of how the service requirement is to be met. Suppose that the company considers leasing comparable equipment that has an annual lease payment of $6000 (after taxes) for the remaining required service period. In this case, the cash flow would look like Figure 17.3.3.

n	*Model A*	*Model B*
0	–12,500	–$15,000
1	–5,000	–4,000
2	–5,000	–4,000
3	–3,000	–4,000
4	**–6,000** –5,000	–2,500
5	**–6,000** –5,000	**–6,000** –5,000

Here the bold figures represent the annual lease payments. (It costs $6000 to lease the equipment and $5000 to operate anually. Other maintenance will be paid by the leasing company.) Note that both alternatives now have the same required service period of 5 years. Therefore, we can use NPW analysis.

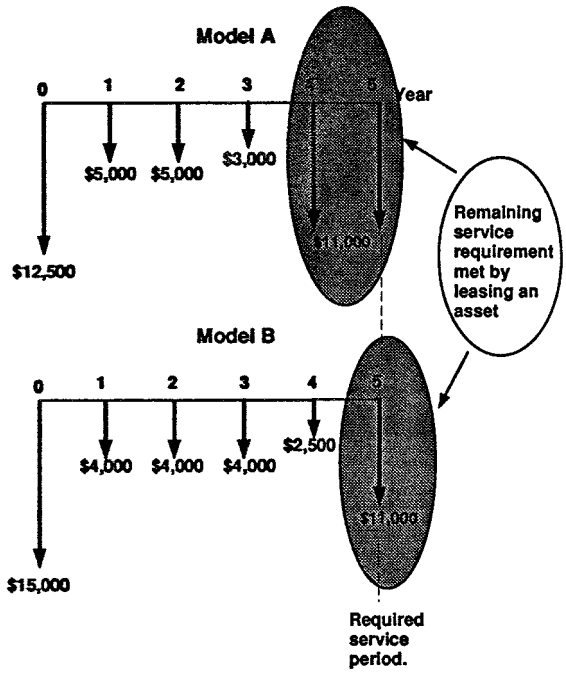


FIGURE 17.3.3 Comparison for unequal-lived projects when the required service period is longer then the individual project life (Example 17.3.6.)

$$PW(15\%)_A = -\$12,500 - \$5000(P/A,15\%,2) + \$3000(P/F,15\%,3)$$
$$- \$11,000(P/A,15\%,2)(P/F,15\%,3)$$
$$= -\$34,359$$

$$PW(15\%)_B = -\$15,000 - \$4000(P/A,15\%,3) - \$2500(P/F,15\%,4) - \$11,000(P/F,15\%,5)$$
$$= -\$32,747$$

Since these are service projects, model B is the better choice.

Flaws in Project Ranking by IRR. Under NPW, the mutually exclusive project with the highest worth figure was preferred. Unfortunately, the analogy does not carry over to IRR analysis. The project with the highest IRR may *not* be the preferred alternative. To illustrate the flaws of comparing IRRs to choose from mutually exclusive projects, suppose you have two mutually exclusive alternatives, each with a 1-year service life; one requires an investment of \$1000 with a return of \$2000 and the other requires \$5000 with a return of \$7000. You already obtained the IRRs and NPWs at MARR = 10% as follows:

n	A1	A2
0	–\$1000	–\$5000
1	222000	7000
IRR	100%	40%
PW(10%)	\$818	\$1364

Would you prefer the first project simply because you expect a higher rate of return?

We can see that A2 is preferred over A1 by the NPW measure. On the other hand, the IRR measure gives a *numerically* higher rating for A1. This inconsistency in ranking is due to the fact that the NPW is an *absolute (dollar)* measure of investment worth, while the IRR is a *relative (percentage)* measure and cannot be applied in the same way. That is, the IRR measure ignores the *scale* of the investment. Therefore, the answer is no; instead, you would prefer the second project with the lower rate of return, but higher NPW. The NPW measure would lead to that choice, but comparison of IRRs would rank the smaller project higher. Another approach, called *incremental analysis*, is needed.

Rate of Return on Incremental Investment

In our previous ranking example, the more costly option requires an incremental investment of $4000 at an incremental return of $5000. If you decide to take the more costly option, certainly you would be interested in knowing that this additional investment can be justified at the MARR. The 10% of MARR value implies that you can always earn that rate from other investment sources — $4400 at the end of 1 year for $4000 investment. However, by investing the additional $4000 in the second option, you would make an additional $5000, which is equivalent to earning at the rate of 25%. Therefore, the incremental investment can be justified.

Now we can generalize the decision rule for comparing mutually exclusive projects. For a pair of mutually exclusive projects (A, B), rate of return analysis is done by computing the *internal rate of return on incremental investment* (IRR_Δ) between the projects. Since we want to consider increments of investment, we compute the cash flow for the difference between the projects by subtracting the cash flow for the lower investment-cost project (A) from that of the higher investment-cost project (B). Then, the decision rule is select B, if $IRR_{B\text{-}A} > MARR$. Otherwise select A.

Example 17.3.7 — IRR on Incremental Investment: Two Alternatives

Reconsider the two mutually exclusive projects in Example 17.3.5.

n	B1	B2	B2–B1
0	–$3,000	–$12,000	–$9,000
1	1,350	4,200	2,850
2	1,800	6,225	4,425
3	1,500	6,330	4,830
IRR	25%	17.43%	

Since B1 is the lower cost investment project, we compute the incremental cash flow for B2–B1. Then we compute the IRR on this increment of investment by solving

$$-\$9000 + \$2850(P/F,i,1) + \$4425(P/F,i,2) + \$4830(P/F,i,3) = 0$$

We obtain $i^*_{B2\text{-}B1} = 15\%$. Since $IRR_{B2\text{-}B1} > MARR$, we select B2, which is consistent with the NPW analysis.

Comments. Why did we choose to look at the increment B2-B1 instead of B1-B2? We want the increment to have investment during at least some part of the time span so that we can calculate an IRR. Subtracting the lower initial investment project from the higher guarantees that the first increment will be investment flow. Ignoring the investment ranking, we might end up with an increment which involves borrowing cash flow and has no internal rate of return. This is the case for B1-B2. ($i^*_{B1\text{-}B2}$ is also 15%, not –15%.) If we erroneously compare this i^* with MARR, we might have accepted project B1 over B2.

17.4 Cash Flow Projections

With the purchase of any fixed asset such as equipment, we need to estimate the profits (more precisely, cash flows) that the asset will generate during its service period. An inaccurate estimate of asset needs can have serious consequences. If you invest too much in assets, you incur unnecessarily heavy expenses. Spending too little on fixed assets also is harmful, for then the firm's equipment may be too obsolete to produce products competitively and without an adequate capacity you may lose a portion of your market share to rival firms. Regaining lost customers involves heavy marketing expenses and may even require price reductions and/or product improvements, all of which are costly. We will begin this section by giving an overview on how a company determines its operating profit.

Operating Profit — Net Income

Firms invest in a project because they expect it to increase their wealth. If the project does this — if project revenues exceed project costs — we say it has generated a *profit*, or *income*. If the project reduces the owner's wealth, we say that the project has resulted in a *loss*. One of the most important roles of the accounting function within an organization is to measure the amount of profit or loss a project generates each year or in any other relevant time period. Any profit generated will be taxed. The accounting measure of a project's after-tax profit during a particular time period is known as *net income*.

Accountants measure the net income of a specified operating period by subtracting expenses from revenues for that period. These terms can be defined as follows:

1. The *project revenue* is the income earned by a business as a result of providing products or services to outsiders. Revenue comes from sales of merchandise to customers and from fees earned by services performed for clients or others.
2. The *project expenses* incurred are the cost of doing business to generate the revenues of that period. Some common expenses are the cost of the goods sold (labor, material, inventory, and supplies), depreciation, the cost of employees' salaries, the operating cost (such as the cost of renting a building and the cost of insurance coverage), and income taxes.

The business expenses listed above are all accounted for in a straightforward fashion on the income statement and balance sheet: the amount paid by the organization for each item would translate dollar for dollar into expenses in financial reports for the period. One additional category of expenses, the purchase of new assets, is treated by depreciating the total cost gradually over time. Because it plays a role in reducing taxable income, depreciation accounting is of special concern to a company.

Accounting Depreciation

The acquisition of fixed assets is an important activity for a business organization, whether the organization is starting up or acquiring new assets to remain competitive. The systematic allocation of the initial cost of an asset in parts over a time known as its depreciable life is what we mean by *accounting depreciation*. Because accounting depreciation is the standard of the business world, we sometimes refer to it more generally as *asset depreciation*.

The process of depreciating an asset requires that we make several preliminary determinations:

1. *What can be depreciable?* Depreciable property includes buildings, machinery, equipment, and vehicles. Inventories are not depreciable property because they are held primarily for sale to customers in the ordinary course of business. If an asset has no definite service life, the asset cannot be depreciated. For example, *you can never depreciate land*.
2. *What cost base should be used in asset depreciation?* The *cost base* of an asset represents the total cost that is claimed as an expense over an asset's life, that is, the sum of the annual depreciation expenses. The cost base generally includes the actual cost of the asset and all the other incidental expenses, such as freight, site preparation, and installation. This total cost, rather

than the cost of the asset only, must be the depreciation base charged as an expense over the asset's life.

3. *What is the asset's value at the end of its useful life?* The salvage value is an asset's value at the end of its life; it is the amount eventually recovered through sale, trade-in, or salvage. The eventual salvage value of an asset must be estimated when the depreciation schedule for the asset is established. If this estimate subsequently proves to be inaccurate, then an adjustment must be made.
4. *What is the depreciable life of the asset?* Historically, depreciation accounting included choosing a depreciable life that was based on the service life of an asset. Determining the service life of an asset, however, was often very difficult, and the uncertainty of these estimates often led to disputes between taxpayers and the IRS. To alleviate the problems, the IRS published guidelines on lives for categories of assets known as *Asset Depreciation Ranges*, or ADRs. These guidelines specified a range of lives for classes of assets based on historical data, and taxpayers were free to choose a depreciable life within the specified range for a given asset.
5. *What method of depreciation do we choose?* Companies generally calculate depreciation one way when figuring taxes and another way when reporting income (profit) to investors: (1) they use the straight-line method (or declining balance or sum-of-years' digits) for investors and (2) they use the fastest rate permitted by law (known as "modified accelerated cost recovery system [MACRS]") for tax purposes. Under the straight-line method, for an asset with a 5-year life which costs $10,000 and has a $1000 salvage value, the annual depreciation charge is ($10,000 – $1000)/5 = $1800. For tax purposes, Congress created several classes of assets, each with a more or less arbitrarily prescribed life called a *recovery period* or *class life*. The depreciable base is not adjusted for salvage value, which is the estimated market value of the asset at the end of its useful life. Table 17.4.1 describes what types of property fit into the different class life groups and the allowed depreciation percentages. Congress developed these recovery allowance percentages based on the declining balance method, with a switch to straight-line depreciation at some point in the asset's life. The MACRS recovery percentages as shown in Table 17.4.1 also employ the half-year convention — that is, they assume that all assets are put into service at midyear and, hence, generate a half-year's depreciation.

Corporate Income Taxes

Corporate taxable income is defined as follows:

$$\text{taxable income} = \text{gross income (revenues)} - (\text{cost of goods sold} + \text{depreciation} + \text{operating expenses})$$

Once taxable income is calculated, income taxes are determined by

$$\text{income taxes} = (\text{tax rate}) \times (\text{taxable income})$$

The corporate tax rate structure for 1996 is relatively simple. There are four basic rate brackets (ranging from 15 to 35%) plus two surtax rates (5 and 3%) based on taxable incomes, and businesses with lower taxable incomes continue to be taxed at lower rates than those with higher taxable incomes.

Tax Treatment of Gains or Losses for Depreciable Assets

When a depreciable asset used in business is sold for an amount different from its book value, this gain or loss has an important effect on income taxes. An asset's book value at any given time is determined by

TABLE 17.4.1 MACRS Depreciation Schedules for Personal Properties with Half-Year Convention

	Personal Property						
Year	Class Depreciation Rate	3 200% DB	5 200% DB	7 200% DB	10 200% DB	15 150% DB	20 150% DB
1		33.33	20.00	14.29	10.00	5.00	3.750
2		44.45	32.00	24.49	18.00	9.50	7.219
3		14.81*	19.20	17.49	14.40	8.55	6.677
4		7.41	11.52*	12.49	11.52	7.70	6.177
5			11.52	8.93*	9.22	6.93	5.713
6			5.76	8.92	7.37	6.23	5.285
7				8.93	6.55*	5.90*	4.888
8				4.46	6.55	5.90	4.522
9					6.56	5.91	4.462*
10					6.55	5.90	4.461
11					3.28	5.91	4.462
12						5.90	4.461
13						5.91	4.462
14						5.90	4.461
15						5.91	4.462
16						2.95	4.461
17							4.462
18							4.461
19							4.462
20							4.461
21							2.231

Note: "*" denotes year to switch from declining balance to straight line.

Property Class	Applicable Property
3-year	Machine tools
5-year	Automobiles, light trucks, high-tech equipment
7-year	Manufacturing equipment, office furniture
10-year	Vessels, barges, tugs, railroad cars
15-year	Utility property (water, telephone)
20-year	Municipal sewers, electrical power plant

$$\text{book value} = \text{cost base} - \text{total amount of depreciation}$$

The gain or loss is found by

$$\text{gains (losses)} = \text{salvage value} - \text{book value}$$

where the salvage value represents the proceeds from the sale less any selling expense or removal cost.

These gains, commonly known as *depreciation recapture*, are taxed as ordinary income. In the unlikely event that if an asset is sold for an amount greater than its initial cost, the gains (salvage value – book value) are divided into two parts (ordinary gains and capital gains) for tax purposes:

$$\text{gains} = \text{salvage value} - \text{book value}$$

$$= \underbrace{(\text{salvage value} - \text{cost base})}_{\text{capital gains}} + \underbrace{(\text{cost base} - \text{book value})}_{\text{ordinary gains}}$$

Capital gains are normally taxed at a different rate from that of ordinary gains.

After-Tax Cash Flow Analysis

In developing an after-tax flow, we are concerned only with those cash flows that result directly from the project. These cash flows, called *incremental cash flows*, represent the change in the firm's total cash flows that occurs as a direct result of undertaking the project. There are several elements that contribute toward the project cash flows. In preparing the cash flow statement which shows sources and uses of cash in project undertaking, we may group them into three areas: (1) cash flow elements associated with operations, (2) cash flow elements associated with investment activities (capital expenditures), and (3) cash flow elements associated with project financing (such as borrowing). The main purpose of this grouping is to provide information about the operating, investing, and financing activities of a project.

- *Operating Activities:* In general, cash flows from operations include current sales revenue, cost of goods sold, operating expense, and income taxes. Cash flows from operations should generally reflect the cash effects of transactions entering into the determination of net income. The interest portion of a loan repayment is a deductible expense when determining net income, and it is included in the operating activities. Since we will usually look only at yearly flows, it is logical to express all cash flows on a yearly basis. We can determine the net cash flow from operations either (1) based on net income or (2) based on cash flow by computing income taxes directly. When we use net income as the starting point for cash flow determination, we should add any noncash expenses (mainly depreciation) back to net income to compute the net cash flow. Thus,

$$\text{net operating cash flow} = \text{net income} + \text{noncash expenses}$$

$$(\text{depreciation})$$

Approach 1	**Approach 2**
Cash revenues (savings)	Cash revenues (savings)
–Cost of goods sold	–Cost of goods sold
–Depreciation	
–Operating expense	–Operating expense
–Interest expense	–Interest expense
Taxable income	
–Income taxes	–Income taxes
Net income + depreciation	Operating cash flow

In business practice, accountants usually prepare the cash flow statements based on the net income, namely, using Approach 1, whereas Approach 2 is commonly used in many traditional engineering economic texts. If you learn only Approach 2, it is more than likely that you need to be retrained to learn Approach 1 to communicate with the financing and accounting professionals within your organization. Therefore, we will use the income statement approach (Approach 1) whenever possible throughout this section.

- *Investing Activities:* Three types of investment flows are associated with buying a piece of equipment: the original investment, salvage value at the end of its useful life, and working capital investment* or recovery.** We will assume that our outflow for both capital investment and working capital investment is as if they take place in year 0. It is quite possible that both investments will not occur instantaneously, but rather over a few months as the project gets into gear; we could then use year 1 as an investment year. (Capital expenditures may occur over several years before a large investment project becomes fully operational. In this case, we should enter all expenditures as they occur.) For a small project, either method of timing these flows is satisfactory, because the numerical differences are likely to be insignificant.
- *Financing Activities:* Cash flows classified as financing activities include (1) the amount of borrowing and (2) repayment of principal. Recall that interest payments are tax deductible expenses so that they are usually classified as operating, not financing, activities.

Then, net cash flow for a given year is simply the sum of the net cash flows from these three activities.

Example 17.4.1 — Developing a Cash Flow Statement

A computerized machining center has been proposed for a small tool manufacturing company. If the new system costing $125,000 is installed, it will generate annual revenues of $100,000 and require $20,000 in annual labor, $12,000 in annual material expenses, and another $8000 in annual overhead (power and utility) expenses. The automation facility would be classified as a 7-year MACRS property. The company expects to phase out the facility at the end of 5 years, at which time it will be sold for $50,000. The machining center will require an investment of $23,331 in working capital (mainly spare parts inventory), which will be recovered in full amount at the end of the project life. Assume that $62,500 of the $125,000 paid for the investment is obtained through a bank loan which is to be repaid in equal annual installments at 10% interest over 5 years. The remaining $62,500 will be provided by equity (for example, from retained earnings). Find the year-by-year after-tax net cash flow for the project at a 40% marginal tax rate based on the net income (Approach 1), and determine the after-tax net present worth of the project at the company's MARR of 15%.

Discussion. We will use the business convention that no signs (positive or negative) are used in preparing the income statement, except in the situation where we have a negative taxable income or tax savings. In this situation we will use ( ). However, in preparing the cash flow statement, we will observe explicitly the sign convention: a positive sign indicating cash inflow, a negative sign or ( ) indicating cash outflow.

Solution. Before presenting the cash flow table, we need some preliminary calculations. The following notes explain the essential items in Table 17.4.2.

- *Depreciation Calculation*:
  1. If it is held for all 7 years, we depreciate a 7-year property in respective percentages of 14.29, 24.49, 17.49, 12.49, 8.93, 8.92, 8.93, and 4.46% (see Table 17.4.1).
  2. If the asset is sold at the end of the fifth tax year (during the recovery period), the applicable depreciation amounts would be $17,863, $30,613, $21,863, $15,613, and $5581. Since the

* Normally, additional inventories are required to support a new operation and increased accounts receivable. These increases in current assets must be financed. On the other hand, accounts payable will also increase as a result of business expansion and this will reduce the net cash needed to finance inventories and accounts receivable. The difference is known as *net change in working capital.* If this difference is positive, an investment in working must be made.

** As the project approaches termination, inventories are sold off and not replaced, and account receivables are eventually collected. As this change occurs, the firm experiences an end-of-project cash flow (known as *working capital recovery*) that is equal to the net working capital investments that were made when the project was begun.

TABLE 17.4.2 Cash Flow Statement for the Automated Machining Center Project with Debt Financing (Example 17.4.1)

	0	1	2	3	4	5
Income Statement						
Revenues		$100,000	$100,000	$100,000	$100,000	$100,000
Labor		20,000	20,000	20,000	20,000	20,000
Material		12,000	12,000	12,000	12,000	12,000
Overhead		8,000	8,000	8,000	8,000	8,000
Depreciation		17,863	30,613	21,863	15,613	5,581
Debt interest		6,250	5,226	4,100	2,861	1,499
Taxable income		$35,887	$24,161	$34,037	$41,526	$52,920
Income taxes (40%)		14,355	9,664	13,615	16,610	21,168
Net income		$21,532	$14,497	$20,433	$24,916	$31,752
Cash Flow Statement						
Operating activities						
Net income		21,532	14,497	20,422	24,916	31,752
Depreciation		17,863	30,613	21,863	15,613	5,581
Investment activities						
Investment	(125,000)					
Salvage						50,000
Gains tax						(6,613)
Working capital	(23,331)					23,331
Financing activities						
Borrowed funds	62,500					
Principal repayment		(10,237)	(11,261)	(12,387)	(13,626)	(14,988)
Net cash flow	$(85,831)	$29,158	$33,849	$29,898	$26,903	$89,063

asset is disposed of in the fifth tax year, the last year's depreciation, which would ordinarily be $11,163, is halved due to the half-year convention.

- *Salvage Value and Gain Taxes*: In year 5, we must deal with two aspects of the asset's disposal — salvage value and gains (both ordinary as well as capital). We list the estimated salvage value as a positive cash flow. Taxable gains are calculated as follows:
  1. The total depreciation in years 1 to 5 is $17,863 + $30,613 + $21,863 + $15,613 + $5581 = $91,533.
  2. The book value at the end of period 5 is the cost base minus the total depreciation, or $125,000 – $91,533 = $33,467.
  3. The gains on the sale are the salvage value minus the book value, or $50,000 – $33,467 = $16,533. (The salvage value is less than the cost base, so all the gain is ordinary.)
  4. The tax on the ordinary gains is $16,633 × 40% = $6,613. This is the amount that is placed in the table under "gains tax".
- *Interest and principal repayment*: We first need to compute the size of the annual installments:

$$\$62,500(A/P,10\%,5) = \$16,487$$

Next, we determine the repayment schedule of the loan by itemizing both the interest and the principal represented in each annual repayment as follows:

The interest payments are listed under the income statement and the principal payments are listed under the cash flow statement as shown in Table 17.4.2.

Year	Beginning Balance	Interest Payment	Principal Payment	Ending Balance
1	$62,500	$6,250	$10,237	$52,263
2	52,263	5,226	11,261	41,002
3	41,002	4,100	12,387	28,615
4	28,615	2,861	13,626	14,989
5	14,989	1,499	14,988	0

- *Investment Analysis:* Once we obtain the project's after-tax net cash flows, we can determine their equivalent present worth at the firm's interest rate of 15%. The present value equivalent of the after-tax cash flow series is

$$PW(15\%) = -\$85,351 + \$29,158(P/F,15\%,1) + \ldots + \$89,063(P/F,15\%,5)$$
$$= \$44,439$$

This means that investing $125,000 in this automated facility would bring in enough revenue to recover the initial investment and the cost of funds with a surplus of $44,439.

Effects of Inflation on Project Cash Flows

We now introduce inflation into estimating project cash flows. We are especially interested in two elements of project cash flows — depreciation expenses and interest expenses — that are essentially immune to the effects of inflation. We also consider the complication of how to proceed when multiple price indexes have been used in generating various project cash flows.

Because depreciation expenses (as well as loan repayment) are calculated on some base-year purchase amount, they do not increase over time to keep pace with inflation. Thus, they lose some of their value to defer taxes as inflation drives up the general price level and hence taxable income. Similarly, salvage values of depreciable assets can increase with the general inflation rate and, because any gains on salvage values are taxable, they can result in increased taxes.

Example 17.4.2 — After-Tax Cash Flows under Inflation

A construction firm is offered a fixed-price contract for a 5-year period. The firm will be paid $23,500 (actual dollars) per year for the contract period. In order to accept the contract, the firm must purchase equipment costing $15,000 and requiring $13,000 (constant dollars) per year to operate. The equipment qualifies for 5-year MACRS depreciation and is expected to have a salvage value of $1000 at the end of 5 years. Use a tax rate of 40% and an inflation-free interest rate of 20%. If the general inflation rate $(\bar{f})$ is expected to average 5% over the next 5 years, salvage value at the annual rate of 5%, and operating expenses at 8% per year for the project duration, should the contractor accept the contract?

Solution. The analysis of this situation should explicitly consider inflation, since the contractor is being offered a fixed-price contract and cannot increase the fee to compensate for increased costs. In this problem, as is common practice, we assume that all estimated costs are expressed in today's dollars, and actual costs will be increased due to inflation.

The Excel spreadsheet analysis is shown in Table 17.4.3, where the first five rows are designated as the input fields. Then column C in both the income statement and the cash flow statement is reserved for entering the specific inflation rate for the item listed in that row. Since the O&M costs are responsive to inflation, we can automate the cell entries by using the following cell formulas:

TABLE 17.4.3 Excel's Spreadsheet Application to "What If" Questions — What General Inflation Rate Does the Project Break Even? (Example 17.15)

	A	B	C	D	E	F	G	H	I
1									
2		INPUT:	O&M Cost	$ 13,000			General Inflation Rate	5.00%	
3			Salvage	$ 1,000			Inflation-Free Interest	20.00%	
4			Contract $	$ 23,000			Market Interest Rate	26.00%	
5			Investment	$ 15,000			Income Tax Rate	40.00%	
6									
7			Inflation Rate	0	1	2	3	4	5
8		Income Statement							
9									
10		Revenues			$ 23,500	$ 23,500	$ 23,500	$ 23,500	$ 23,500
11		Expenses:							
12		O&M	8%		14,040	15,163	16,376	17,686	19,101
13		Depreciation			3,000	4,800	2,880	1,728	864
15		Taxable Income			$ 6,460	$ 3,537	$ 4,244	$ 4,086	$ 3,535
16		Income Taxes (40%)			2,584	1,415	1,697	1,634	1,414
18		Net Income			$ 3,876	$ 2,122	$ 2,546	$ 2,451	$ 2,121
19									
20		Cash Flow Statement							
21									
22		Operating Activities:							
23		Net Income			3,876	2,122	2,546	2,451	2,121
24		Depreciation			3,000	4,800	2,880	1,728	864
25		Investment Activities:							
26		Investment		(15,000)					
27		Salvage	5%						1,276
28		Gains Tax							181
30		Net Cash Flow		$ (15,000)	$ 6,876	$ 6,922	$ 5,426	$ 4,179	$ 4,442
31		(in actual dollars)							
32		Net Cash Flow		$ (15,000)	$6,549	$6,279	$4,687	$3,438	$3,480
33		(in constant dollars)							
34		Equ. Present Worth		$ (15,000)	$5,457	$4,360	$2,713	$1,658	$1,399
36		Net Present Worth	$587						
37		Internal Rate of Return	22.03%						

$$\text{cell E12} = \$D\$2^{*}(1+\$C\$12)$$

$$\text{cell F12} = E12^{*}(1+\$C\$12)$$

$$\vdots \quad \vdots$$

$$\text{cell I12} = H12^{*}(1+\$C\$12)$$

Note that the salvage value in cell I27 has been increased at 5%. (When no inflation rate is given for a specific cost category, it is normally assumed that the general inflation rate holds.) Gains taxes on disposal are based on the difference between book value and salvage value, in this case, $1276 – $1728 = ($452) or loss of $452. This results in a tax savings of ($452) (0.40) = $181, as shown in cell I28.

In row 12, operating expenses have been increased at 8%. As explained in the section "Loss of Purchasing Power," actual-dollar cash flows are converted to constant-dollar flows by "deflating" at the general inflation rate. The IRR is calculated at 22.03%, based on the constant-dollar cash flows. This is compared to the MARR of 20%, the inflation-free interest rate, indicating that this is an acceptable contract as long as the 5 and 8% escalation rates are correct.

Since no one can predict inflation rates in any precise manner, it is always wise to investigate the effects of changes in these rates. This is very easy to do with a spreadsheet. For example, we can easily determine the value of the general inflation rate at which this project exactly earns 20%. The value of $\bar{f}$ was adjusted manually until the IRR was exactly 20%. This would not be a good project if $\bar{f}$ were greater than 6.77% (cell H2). (Note that at the exact value of $\bar{f}$ (6.771%), the NPW in cell C36 will be zero or its rate of return would be 20%.) Similarly, you can vary the O&M cost and salvage value to see how the project's profitability changes.

17.5 Sensitivity and Risk Analysis

In previous sections, the cash flows from projects were assumed to be known with complete certainty, and our analysis was concerned with measuring the economic worth of projects and selecting the best investment projects. Although these results can provide reasonable decision bases for many investment situations, we should certainly consider the more usual situation where forecasts of cash flows are subject to some degree of uncertainty. In this situation, management rarely has precise expectations regarding the future cash flows to be derived from a particular project. In fact, the best that the firm can reasonably expect to do is to estimate the range of possible future costs and benefits and the relative chances of achieving a certain return on the investment. We can use the term risk in describing an investment project whose cash flow is not known in advance with absolute certainty, but for which an array of alternative outcomes and their probabilities (odds) are known. We will also use the term project risk to refer to the variability in a project's NPW. A greater project risk means that there is a greater variability in the project's NPW, or simply saying that *risk is the potential for loss*. We may begin analyzing project risk by first determining the uncertainty inherent in a project's cash flows. We can consider risk in a number of ways ranging from making informal judgments to calculating complex economic and statistical analyses. In this section, we will introduce three methods of describing project risk: (1) sensitivity analysis, (2) scenario analysis, and (3) risk analysis. We shall explain each method with a single example (Boston Metal Company).

Project Risk

The decision to make a major capital investment such as the introduction of a new product requires cash flow information over the life of the project. The profitability estimate of the investment depends on cash flow estimations, which are generally uncertain. Many cash flow elements (such as demand) are subject to substantial uncertainty. The common approach is to make single-number "best estimates" for each of the uncertain factors and then to calculate measures of profitability such as NPW or rate of return for the project. This approach has two drawbacks:

1. There is no guarantee that the "best estimates" will ever match actual values.
2. There is no way to measure the risk associated with the investment or the project risk. In particular, the manager has no way of determining either the probability that the project will lose money or the probability that it will generate very large profits.

Because cash flows can be so difficult to estimate accurately, project managers frequently consider a range of possible values for cash flow elements. If there is a range of possible values for individual cash flows, it follows that there is a range of possible values for the NPW of a given project. Clearly, the analyst will want to try to gauge the probability and reliability of individual cash flows occurring and, consequently, the level of certainty about achieving the overall project worth.

Sensitivity Analysis

One way to glean a sense of the possible outcomes of an investment is to perform a sensitivity analysis. This analysis determines the effect on NPW of variations in the input variables (such as revenues,

operating cost, and salvage value) used to estimate after-tax cash flows. A *sensitivity analysis* reveals how much the NPW will change in response to a given change in an input variable. In a calculation of cash flows, some items have a greater influence on the final result than others. In some problems, we may easily identify the most significant item. For example, the estimate of sales volume is often a major factor in a problem in which the quantity sold varies among the alternatives. In other problems, we may want to locate the items that have an important influence on the final results so that they can be subjected to special scrutiny.

Sensitivity analysis is sometimes called "what-if" analysis because it answers questions such as: What if incremental sales are only 1000 units, rather than 2000 units? Then what will the NPW be? Sensitivity analysis begins with a base-case situation, which is developed using the most-likely values for each input. Then we change the specific variable of interest by several specific percentages above and below the most-likely value, holding other variables constant. Next, we calculate a new NPW for each of these values. A convenient and useful way to present the results of a sensitivity analysis is to plot *sensitivity graphs*. The slopes of the lines show how sensitive the NPW is to changes in each of the inputs: the steeper the slope, the more sensitive the NPW is to a change in the particular variable. It identifies the crucial variables that affect the final outcome most. We will use Example 17.5.1 to illustrate the concept of sensitivity analysis.

Example 17.5.1 — Sensitivity Analysis

Boston Metal Company (BMC), a small manufacturer of fabricated metal parts, must decide whether to enter the competition to be the supplier of transmission housings for Gulf Electric. Gulf Electric has its own in-house manufacturing facility to produce transmission housings, but it has almost reached its maximum production capacity. Therefore, Gulf is looking for an outside supplier. To compete, the firm must design a new fixture for the production process and purchase a new forge. The new forge would cost $125,000, including tooling costs for the transmission housings. If BMC gets the order, it may be able to sell as many as 2000 units per year to Gulf Electric for $50 each, and the variable production costs,* such as direct labor and direct material costs, will be $15 per unit. The increase in fixed costs** other than depreciation will amount to $10,000 per year. The firm expects that the proposed transmission-housings project would have about a 5-year product life. The firm also estimates that the amount ordered by Gulf Electric for the first year will be ordered in each of the subsequent 4 years. (Due to the nature of contracted production, the annual demand and unit price would remain the same over the project after the contract is signed.) The initial investment can be depreciated on a MACRS basis over the 7-year period, and the marginal income tax rate is expected to remain at 40%. At the end of 5 years, the forge machine is expected to retain a market value of about 35% of the original investment. Based on this information, the engineering and marketing staffs have prepared the cash flow forecasts shown in Table 17.5.1. Since NPW is positive ($40,169) at the 15% opportunity cost of capital (MARR), the project appears to be worth undertaking.

However, BMC's managers are uneasy about this project because there are too many uncertain elements that have not been considered in the analysis. If decided, BMC must make the investment in the forging machine to provide some samples with Gulf Electric as a part of the bidding process. If Gulf Electric does not like BMC's sample, BMC stands to lose the entire investment in the forging machine. On the other hand, if Gulf likes BMC's sample but it is overpriced, BMC would be under pressure to bring the price in line with competing firms. There is even a possibility that BMC would get a smaller order as Gulf may utilize their overtime capacity to produce some extra units. They also are not certain about the variable and fixed cost figures. Recognizing these uncertainties, the managers want to assess the various potential future outcomes before making a final decision. (a) Perform a sensitivity analysis to each variable and (b) develop a sensitivity graph.

---

* Expenses that change in direct proportion to the change in volume of sales or production.

** Expenses that do not vary as the volume of sales or production changes. For example, property taxes, insurance, depreciation, and rent are usually fixed expenses.

TABLE 17.5.1 A Typical Excel Worksheet Design to Perform Sensitivity Analyses (Example 17.5.1)

	A	B	C	D	E	F	G	H
1								
2		Input Data (Base):			Sensitivity Analysis:			
3		Unit Price ($)	50.00					
4		Demand	2000		Category		% Change	
5		Var. cost ($/unit)	15.00		Unit price		0%	
6		Fixed cost ($)	10000		Demand		0%	
7		Salvage ($)	40000		Var. cost (unit)		0%	
8		Tax rate (%)	40%		Fixed cost		0%	
9		MARR (%)	15%		Salvage		0%	
10								
11					Output (NPW)		$40,169	
12								
13			0	1	2	3	4	5
14		Income Statement						
15								
16		Revenues:						
17		Unit Price		$ 50.00	$ 50.00	$ 50.00	$ 50.00	$ 50.00
18		Demand (units)		2000	2000	2000	2000	2000
19		Sales Revenue		$ 100,000	$ 100,000	$ 100,000	$ 100,000	$ 100,000
20		Expenses:						
21		Unit Variable Cost		$ 15	$ 15	$ 15	$ 15	$ 15
22		Variable Cost		30,000	30,000	30,000	30,000	30,000
23		Fixed Cost		10,000	10,000	10,000	10,000	10,000
24		Depreciation		17,863	30,613	21,863	15,613	5,581
26		Taxable Income		$ 42,137	$ 29,387	$ 38,137	$ 44,387	$ 54,419
27		Income Taxes (40%)		16,855	11,755	15,255	17,755	21,768
29		Net Income		$ 25,282	$ 17,632	$ 22,882	$ 26,632	$ 32,651
30								
31		Cash Flow Statement						
32								
33		Operating Activities:						
34		Net Income		25,282	17,632	22,882	26,632	32,651
35		Depreciation		17,863	30,613	21,863	15,613	5,581
36		Investment Activities:						
37		Investment	(125,000)					
38		Salvage						40,000
39		Gains Tax						(2,613)
41		Net Cash Flow	$ (125,000)	$ 43,145	$ 48,245	$ 44,745	$ 42,245	$ 75,619

Discussion. Table 17.5.1 shows BMC's expected cash flows — but there is no guarantee that they will indeed materialize. BMC is not particularly confident in is revenue forecasts. The managers think that if competing firms enter the market, BMC will lose a substantial portion of the projected revenues. Before undertaking the project described, the company wants to identify the key variables that determine whether the project will succeed or fail. The marketing department has estimated revenue as follows:

$$\text{annual revenue} = (\text{product demand})(\text{unit price})$$

$$= (2000)(\$50) = \$100,000$$

The engineering department has estimated variable costs such as labor and material per unit at $15. Since the projected sales volume is 2000 units per years, the total variable cost is $30,000.

Having defined the unit sales, unit price, unit variable cost, and fixed cost, we conduct a sensitivity analysis with respect to these key input variables. This is done by varying each of the estimates by a

given percentage and determining what effect the variation in that item has on the final results. If the effect is large, the result is sensitive to that item. Our objective is to locate the most sensitive item(s).

Solution.

(a) *Sensitivity analysis:* We begin the sensitivity analysis with a "base-case" situation, which reflects the best estimate (expected value) for each input variable. In developing Table 17.5.2, we changed a given variable by 20% in 5% increments, above and below the base-case value, and calculated new NPWs, holding other variables constant. The values for both sales and operating costs were the expected, or base-case, values, and the resulting $40,169 is called the base-case NPW. Now we ask a series of "what-if" questions: What if sales are 20% below the expected level? What if operating costs rise? What if the unit price drops from $50 to $45? Table 17.5.2 summarizes the results of varying the values of the key input variables.

(b) *Sensitivity graph:* Figure 17.5.1 shows the transmission project's sensitivity graphs for six of the key input variables. The base-case NPW is plotted on the ordinate of the graph at the value 1.0 on the abscissa. Next, the value of product demand is reduced to 0.95 of its base-case value, and the NPW is recomputed with all other variables held at their base-case value. We repeat the process by either decreasing or increasing the relative deviation from the base case. The lines for the variable unit price, variable unit cost, fixed costs, and salvage value are obtained in the same manner. In Figure 17.5.1, we see that the project's NPW is very sensitive to changes in product demand and unit price, is fairly sensitive to changes in the variable costs, and is relatively insensitive to changes in the fixed cost and the salvage value.

Graphic displays such as those in Figure 17.5.1 provide a useful means to communicate the relative sensitivities of the different variables on the corresponding NPW value. However, the sensitivity graph does not explain any variable interactions among the variables or the likelihood of realizing any specific deviation from the base case. Certainly, it is conceivable that an answer might not be very sensitive to changes in either of the two items, but very sensitive to combined changes in them.

Scenario Analysis

Although sensitivity analysis is useful, it does have limitations. It is often difficult to specify precisely the relationship between a particular variable and the NPW. The relationship is further complicated by interdependencies among the variables. Holding operating costs constant while varying unit sales may ease the analysis, but in reality, operating costs do not behave in that manner. Yet it may complicate the analysis too much to permit movement in more than one variable at a time. A scenario analysis is a technique that does consider the sensitivity of NPW both to changes in key variables and to the range of likely variable values. For example, the decision-maker may consider two extreme cases, a "worst-case" scenario (low unit sales, low unit price, high variable cost per unit, high fixed cost, and so on) and a "best-case" scenario. The NPWs under the worst and best conditions are then calculated and compared to the expected, or base-case, NPW. Example 17.5.2 will illustrate a plausible scenario analysis for the BMC's transmission-housings project.

Example 17.5.2 — Scenario Analysis

Consider again BMC's transmission-housing project in Example 17.5.2. Assume that the company's managers are fairly confident of their estimates of all the projects's cash flow variables except that for unit sales. Further, assume that they regard a decline in unit sales to below 1600 or a rise above 2400 as extremely unlikely. Thus, decremental annual sales of 400 units define the lower bound, or the worst-case scenario, whereas incremental annual sales of 400 units define the upper bound, or the best-case scenario. (Remember that the most-likely value was 2000 in annual unit sales.) Discuss the worst- and best-case scenarios, assuming that the unit sales for all 5 years would be equal.

TABLE 17.5.2 Sensitivity Analysis for Six Key Input Variables (Example 17.5.1)

Deviation	-20%	-15%	-10%	-5%	Base 0%	5%	10%	15%	20%
Unit price	$ (57)	$ 9,999	$ 20,055	$ 30,111	$ 40,169	$ 50,225	$ 60,281	$ 70,337	$ 80,393
Demand	$ 12,010	$ 19,049	$ 26,088	$ 33,130	$ 40,169	$ 47,208	$ 54,247	$ 61,286	$ 68,325
Variable cost	$ 52,236	$ 49,219	$ 46,202	$ 43,186	$ 40,169	$ 37,152	$ 34,135	$ 31,118	$ 28,101
Fixed cost	$ 44,191	$ 43,185	$ 42,179	$ 41,175	$ 40,169	$ 39,163	$ 38,157	$ 37,151	$ 36,145
Salvage value	$ 37,782	$ 38,378	$ 38,974	$ 39,573	$ 40,169	$ 40,765	$ 41,361	$ 41,957	$ 42,553
MARR	$ 53,588	$ 50,075	$ 46,670	$ 43,369	$ 40,169	$ 37,064	$ 34,051	$ 32,128	$ 28,289

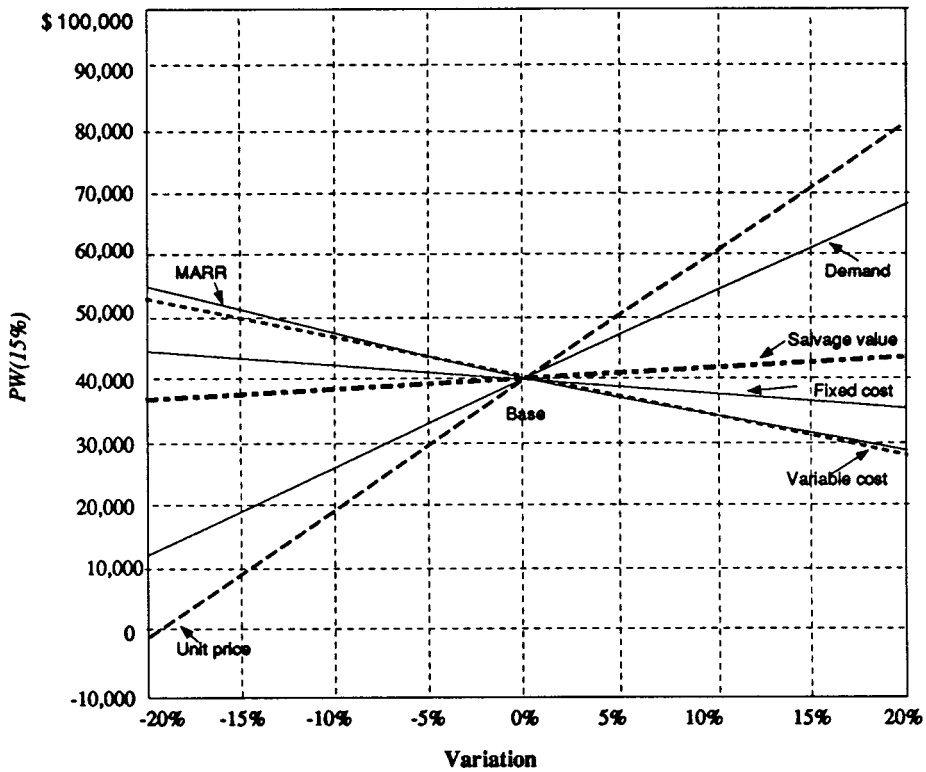


FIGURE 17.5.1 Sensitivity graph — BMC's transmission-housings project (Example 17.5.1).

Discussion. To carry out the scenario analysis, we ask the marketing and engineering staffs to give optimistic (best-case) and pessimistic (worst-case) estimates for the key variables. Then we use the worst-case variable values to obtain the worst-case NPW and the best-case variable values to obtain the best-case NPW. Table 17.5.3 summarizes the results of our analysis. We see that the base case produces a positive NPW, the worst case produces a negative NPW, and the best case produces a large positive NPW.

TABLE 17.5.3 Scenario Analysis for BMC (Example 17.5.2)

Variable Considered	Worst-Case Scenario	Most-Likely-Case Scenario	Best-Case Scenario
Unit demand	1,600	2,000	2,400
Unit price ($)	48	50	53
Variable cost ($)	17	15	12
Fixed cost ($)	11,000	10,000	8,000
Salvage value ($)	30,000	40,000	50,000
PW (15%)	–$5,856	$40,169	$104,295

By just looking at the results in Table 17.5.3, it is not easy to interpret scenario analysis or to make a decision based on it. For example, we can say that there is a chance of losing money on the project, but we do not yet have a specific probability for this possibility. Clearly, we need estimates of the probabilities of occurrence of the worst case, the best case, the base case (most likely), and all the other possibilities. This need leads us directly to the next step, developing a probability distribution. If we can predict the effects on the NPW of variations in the parameters, why should we not assign a probability distribution to the possible outcomes of each parameter and combine these distributions in some way

to produce a probability distribution for the possible outcomes of the NPW? We shall consider this issue in the next section.

Risk Analysis

Quantitative statements about risk are given as numerical probabilities, or as values for likelihoods (odds) of occurrence. Probabilities are given as decimal fractions in the interval 0.0 to 1.0. An event or outcome that is certain to occur has a probability of 1.0. As the probability of an event approaches 0, the event becomes increasingly less likely to occur. The assignment of probabilities to the various outcomes of an investment project is generally called *risk analysis*. In this section, we shall assume that the analyst has available the probabilities (likelihoods) of future events from either previous experience in a similar project or a market survey. The use of probability information can provide management with a range of possible outcomes and the likelihood of achieving different goals under each investment alternative.

Procedure for Developing an NPW Distribution

To develop the NPW distribution, we may follow these steps:

1. Identify all cash flows elements that are random variables. (A *random variable*) is a variable that can have more than one possible value.)
2. Express the NPW (or cash flow series) as a function of random variables.
3. Determine the probability distribution for each random variable.
4. Determine the joint events and their probabilities.
5. Evaluate the NPW equation at these joint events.
6. Order the NPW values in increasing order of NPW.

These steps can best be illustrated by Example 17.5.3.

Example 17.5.3 — Developing an NPW Probability Distribution

Consider the BMC's transmission-housing project. If the unit sales (X) and unit price (Y) were to vary with the following probabilities, determine the NPW probability distribution. Here we assume the situation where both random variables are independent. In other words, observing a typical outcome for random variable X does not have any influence on predicting the outcome for random variable Y.

Demand (X)	Probability	Unit Price (Y)	Probability
1600	0.20	$48	0.30
2000	0.60	$50	0.50
2400	0.20	$53	0.20

Discussion. If the product demand X and the unit price Y are independent random variables, the PW (15%) will also be a random variable. To determine the NPW distribution, we need to consider all the combinations of possible outcomes. The first possibility is the event where x = 1600 and y = $48. With these values as input in Table 17.5.1, we compute the resulting NPW outcome to be $5574. Since X and Y are considered to be independent random variables, the probability of this joint event is

$$\begin{aligned} P(x = 1600, y = \$48) &= P(x = 1600)\,P(y = \$48) \\ &= (0.20)(0.30) \\ &= 0.06 \end{aligned}$$

TABLE 17.5.4 The NPW Probability Distribution with Independent Random Variables (Example 17.5.3)

Event No.	x	y	P(x,y)	Cumulative Joint Probability	NPW
1	1600	$48.00	0.06	0.06	$5,574
2	1600	$50.00	0.10	0.16	$12,010
3	1600	$53.00	0.04	0.20	$21,664
4	2000	$48.00	0.18	0.38	$32,123
5	2000	$50.00	0.30	0.68	$40,168
6	2000	$53.00	0.12	0.80	$52,236
7	2400	$48.00	0.06	0.86	$58,672
8	2400	$50.00	0.10	0.96	$68,326
9	2400	$53.00	0.04	1.00	$82,808

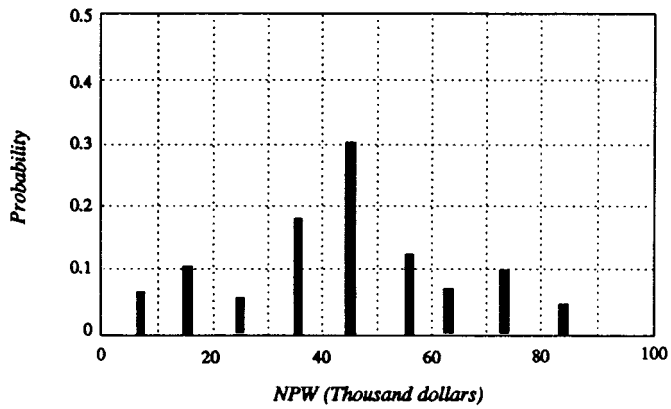


FIGURE 17.5.2 NPW probability distributions: when X and Y are independent (Example 17.5.3).

There are eight other possible joint outcomes. Substituting these pairs of values in Table 17.5.1, we obtain the NPWs and their joint probabilities in Table 17.5.4 and its NPW distribution as depicted in Figure 17.5.2.

Solution. The NPW probability distribution in Table 17.5.4 indicates that the project's NPW varies between $5574 and $82,808, but that there is no loss under any of the circumstances examined. From the cumulative distribution, we further observe that there is a 0.38 probability that the project would realize an NPW less than that forecast for the base-case situation ($40,168). On the other hand, there is a 0.32 probability that the NPW will be greater than this value. Certainly, the probability distribution provides much more information on the likelihood of each possible event, as compared with the scenario analysis presented in Table 17.5.3.

Expected Value and Variance

We have developed a probability distribution for the NPW by considering the random cash flows. As we have observed, the probability distribution helps us to see what the data imply in terms of the risk of the project. With the NPW distribution defined, we can further summarize the probabilistic information — the mean and the variance.

The expected value (also called the *mean*) is a weighted average value of the random variable (μ) where the weighting factors are the probabilities of occurrence. (See Table 17.5.5.) The expected value of a distribution tells us important information about the "average" or expected value of a random variable such as the NPW, but it does not tell us anything about the variability on either side of the

TABLE 17.5.5 Calculation of the Mean and Variance of NPW Distribution (Example 17.5.3)

Event No.	x	y	P(x,y)	Cumulative Joint Probability	NPW	Weighted NPW
1	1600	$48.00	0.06	0.06	$5,574	$334
2	1600	$50.00	0.10	0.16	$12,010	$1,201
3	1600	$53.00	0.04	0.20	$21,664	$867
4	2000	$48.00	0.18	0.38	$32,123	$5,782
5	2000	$50.00	0.30	0.68	$40,168	$12,050
6	2000	$53.00	0.12	0.80	$52,236	$6,268
7	2400	$48.00	0.06	0.86	$58,672	$3,520
8	2400	$50.00	0.10	0.96	$68,326	$6,833
9	2400	$53.00	0.04	1.00	$82,808	$3,312
						E(NPW) = $40,168

Event No.	x	y	P(x,y)	NPW	(NPW – E[NPW])^2	Weighted (NPW – E[NPW])^2
1	1600	$48.00	0.06	$5,574	1,196,769,744	71,806,185
2	1600	$50.00	0.10	$12,010	792,884,227	79,288,423
3	1600	$53.00	0.04	$21,664	342,396,536	13,695,861
4	2000	$48.00	0.18	$32,123	64,725,243	11,650,544
5	2000	$50.00	0.30	$40,168	0	0
6	2000	$53.00	0.12	$52,236	145,631,797	17,475,816
7	2400	$48.00	0.06	$58,672	342,396,536	20,543,792
8	2400	$50.00	0.10	$68,326	792,884,227	79,288,423
9	2400	$53.00	0.04	$82,808	1,818,132,077	72,725,283
						Var[PW(15%)] = 366,474,326
						Standard Deviation = $19,144

expected value. Will the range of possible values of the random variable be very small, with all the values located at or near the expected value?

Another measure that we need in analyzing probabilistic situations is a measure of the risk due to the variability of the outcomes. There are several measures of the variation of a set of numbers that are used in statistical analysis — the range and the variance (or standard deviation), among others. The variance and the standard deviation are used most commonly in the analysis of risk situations. We will use Var[X] or σ_x^2 to denote the variance, and σ_x to denote the standard deviation of random variable X. (If there is only one random variable in an analysis, we normally omit the subscript.) The variance tells us the degree of spread, or dispersion, of the distribution on either side of the mean value. As the variance increases, the spread of the distribution increases; the smaller the variance, the narrower the spread about the expected value.

To determine the variance, we first calculate the deviation of each possible outcome x_j from the expected value $(x_j - \mu)$, then raise the result to the second power and multiply it by the probability of x_j occurring (that is, p_j). The summation of all these products serves as a measure of the distribution's variability. To be most useful, any measure of risk should have a definite value (unit). One such measure is the standard deviation. To calculate the standard deviation, we take the positive square root of Var[X], which is measured in the same units as is X. The standard deviation is a probability-weighted deviation (more precisely, square root of sum of squared deviations) from the expected value. Thus, it gives us an idea of how far above or below the expected value the actual value is likely to be. For most probability distributions, the actual value will be observed within the $\pm 3\sigma$ range. In our BMC project, we obtain the variance of the NPW distribution, assuming independence between X and Y, as shown in Table 17.5.5.

Decision Rule

Once the expected value has been located from the NPW distribution, it can be used to make an accept-reject decision, in much the same way that a single NPW is used when a single possible outcome is considered for an investment project. The decision rule is called the *expected value criterion*, and using it we may accept a single project if its expected NPW value is positive. In the case of mutually exclusive alternatives, we select the one with the highest expected NPW. The use of expected NPW has an advantage over the use of a point estimate, such as the likely value, because it includes all the possible cash flow events and their probabilities.

The justification for the use of the expected value criterion is based on the *law of large numbers*, which states that if many repetitions of an experiment are performed, the average outcome will tend toward the expected value. This justification may seem to negate the usefulness of the expected value criterion in economic analysis, since most often in project evaluation we are concerned with a single, nonrepeatable "experiment" — that is, investment alternative. However, if a firm adopts the expected value criterion as a standard decision rule for *all* its investment alternatives, over the long term the law of large numbers predicts that accepted projects will tend to meet their expected values. Individual projects may succeed or fail, but the average project result will tend to meet the firm's standard for economic success.

The expected-value criterion is simple and straightforward to use, but it fails to reflect the variability of investment outcome. Certainly, we can enrich our decision by incorporating the variability information along with the expected value. Since the variance represents the dispersion of the distribution, it is desirable to minimize it. In other words, the smaller the variance, the less the variability (the potential for loss) associated with the NPW. Therefore, when we compare the mutually exclusive projects, we may select the alternative with the smaller variance if its expected value is the same as, or larger than, those of other alternatives. In cases where there are no clear-cut preferences, the ultimate choice will depend on the decision-maker's trade-offs — how much he or she is willing to take the variability to achieve a higher expected value. In other words, the challenge is to decide what level of risk you are willing to accept and then, having decided on your risk tolerance, to understand the implications of that choice.

17.6 Design Economics

Engineers frequently have to come up with a minimum-cost solution when they have two or more cost components that are affected differently by the same design element. That is, for a single design variable, some costs increase while others decrease. Another valuable extension of the AE analysis in the section on "Annual Equivalent Method," is that these cost components be expressed in equivalent annual form so that we can identify which cost component we need to control to obtain the minimum cost solution to various engineering design problems.

Capital Costs vs. Operating Costs

The AE method is sometimes called the annual equivalent cost method when only costs are involved. In this case, there are two kinds of costs that revenues must cover: *operating costs* and *capital costs*. Operating costs are incurred by the operation of physical plant or equipment to provide service; they include such items as labor and raw materials. Capital costs are incurred by purchasing the assets used in production and service. Normally, capital costs are nonrecurring (that is, one-time costs), whereas operating costs recur as long as the asset is owned.

Because operating costs recur over the life of a project, they tend to be estimated on an annual basis anyway, so for the purposes of an annual equivalent cost analysis, they require no special calculation on our part. However, because capital costs tend to be one-time costs in conducting an annual equivalent

cost analysis, we must translate this one-time cost into its annual equivalent over the life of the project. The annual equivalent of a capital cost is given a special name: *capital recovery cost*, designated CR(i).

There are two general monetary transactions associated with the purchase and eventual retirement of a capital asset, its initial cost (I) and its salvage value (S). Taking into account these sums, we calculate the capital recovery cost as

$$CR(i) = I(A/P,i,N) - S(A/F,i,N) \tag{17.6.1}$$

Recalling the algebraic relationships between factors in Table 17.1.1, notice that the (A/P,*i*,N) factor can be expressed as

$$(A/P,i,N) = (A/F,i,N) + i$$

Then, we may rewrite the CR(*i*) as

$$\begin{aligned} CR(i) &= I(A/P,i,N) - S\left[(A/P,i,N) - i\right] \\ &= (I - S)(A/P,i,N) + iS \end{aligned} \tag{17.6.2}$$

Minimum-Cost Function

When the equivalent annual total cost of a design variable is a function of increasing (O&M costs) and decreasing cost components (capital costs), we usually can find the optimal value that will minimize its cost.

$$AE(i) = a + bx + \frac{c}{x} \tag{17.6.3}$$

where x is a common design variable, and a, b, and c are constants.

To find the value of the common design variable that minimizes the AE(*i*), we need to take the first derivative, equate the result to zero, and solve for x.

$$\begin{aligned} \frac{dAE(i)}{dx} &= b - \frac{c}{x^2} \\ &= 0 \\ x^* &= \sqrt{\frac{c}{b}} \end{aligned} \tag{17.6.4}$$

The value x^* is the minimum cost point for the design alternative. To illustrate the optimization concept, we will consider selecting an optimal pipe size.

Example 17.6.1 — Economical Pipe Size

As a result of the conflict in the Persian Gulf, Kuwait is studying the feasibility of running a steel pipeline across the Arabian Peninsula to the Red Sea. The pipeline will be designed to handle 3 million barrels of crude oil per day at optimum conditions. The length of the line will be 600 miles. Calculate the optimum pipeline diameter that will be used for 20 years for the following data at i = 10%:

- Pumping power = 1.333QΔP/1,980,000 horsepower
- Q = volume flow rate, cubic ft/hr

- $\Delta P = 128Q\mu L/g\pi D^4$, pressure drop lb/sq ft
- L = pipe length, ft
- D = inside pipe diameter, ft
- t = 0.01 D, pipeline wall thickness, ft
- μ = 8500 lb/hr ft, oil viscosity
- g = 32.2 × 12,960,000 ft/hr^2
- Power cost, $0.015 per horsepower hour
- Oil cost, $18 per barrel
- Pipeline cost, $1.00 per pound of steel
- Pump and motor costs, $195 per horsepower

The salvage value of the steel after 20 years is assumed to be zero considering the cost of removal. (See Figure 17.6.1 for relationship between D and t.)

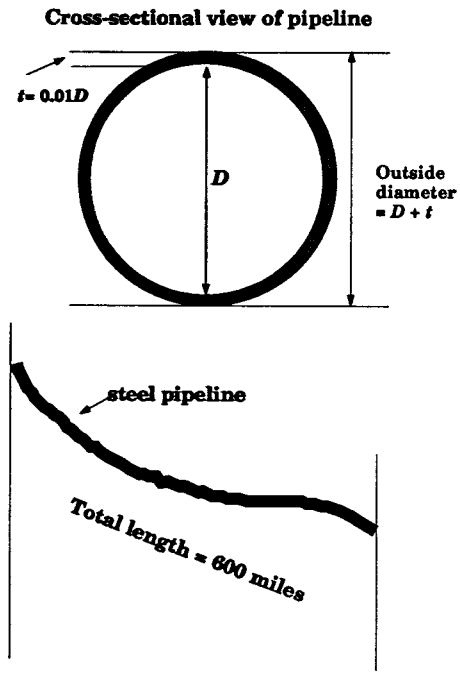


FIGURE 17.6.1 Designing economical pipe size to handle 3 million barrels of crude oil per day (Example 17.6.1).

Discussion. In general, when progressively larger-size pipe is used to carry a given fluid at a given volume flow rate, the energy required to move the fluid will progressively decrease. However, as we increase the pipe size, the cost to construct the pipe will increase. In practice, to obtain the best pipe size for a particular situation, you may choose a reasonable, but small, starting size. Compute the energy cost of pumping fluid through this size and the total construction cost. Compare the difference in energy cost with the difference in construction cost. When the savings in energy cost exceed the added construction cost, you may repeat the process with progressively larger pipe sizes until the added construction cost exceeds the savings in energy cost. As soon as this happens the best pipe size to use in the particular application is identified. However, we can simplify this search process by using the minimum cost concept as explained through Equations (17.6.3) and (17.6.4).

Solution. We will solve the pipe sizing problem in 8 steps:

1. Since the pipe size will be measured in inches, we will assume the following basic conversion units:
   - 1 mile = 5280 ft
   - 600 miles = 600 × 5280 = 3,168,000 ft
   - 1 barrel = 42 U.S. gal
   - 1 barrel = 42 gal × 231 in.^{3}/gal = 9702 in.^{3}
   - 1 barrel = 9702 in.^{3}/12^3 = 5.6146 ft^3
   - Density of steel = 490.75 lb/ft^3
2. To determine the operating cost in pumping oil, we first need to determine power (electricity) required to pump oil:
   - Volume flow rate per hour:

$$\begin{aligned} Q &= 3{,}000{,}000 \text{ barrels/day} \times 5.6146 \text{ ft}^3\text{/barrel} \\ &= 16{,}843{,}800 \text{ ft}^3\text{/day} \\ &= 701{,}825 \text{ ft}^3\text{/hr} \end{aligned}$$

   - Pressure drop:

$$\begin{aligned} \Delta P &= \frac{128 Q \mu L}{g \pi D^4} \\ &= \frac{128 \times 701{,}825 \times 8500 \times 3{,}168{,}000}{32.2 \times 12{,}960{,}000 \times 3.14159 D^4} \\ &= \frac{1{,}845{,}153{,}595}{D^4} \text{ lb/ft}^2 \end{aligned}$$

   - Pumping power required to boost the pressure drop:

$$\begin{aligned} \text{power} &= \frac{1.333 Q \Delta P}{1{,}980{,}000} \\ &= \frac{1.333 \times 701{,}825 \times \dfrac{1{,}845{,}153{,}595}{D^4}}{1{,}980{,}000} \\ &= \frac{871{,}818{,}975}{D^4} \text{ hp} \end{aligned}$$

   - Power cost to pump oil:

$$\begin{aligned} \text{power cost} &= \frac{871{,}818{,}975}{D^4} \text{ hp} \times \$0.015\text{/hp.hr} \times 24 \text{ hr/day} \times 365 \text{ days/year} \\ &= \frac{\$114{,}557{,}013{,}315}{D^4} \Big/ \text{year} \end{aligned}$$

3. Pump and motor cost calculation

$$\text{pump and motor cost} = \frac{871,818,975}{D^4} \times \$195/\text{hp}$$

$$= \frac{\$170,004,700,125}{D^4}$$

4. Required amount and cost of steel

$$\text{cross-sectional area} = \frac{3.14159\left[(0.51D)^2 - (0.50D)^2\right]}{4}$$

$$= 0.032D^2 \text{ ft}^2$$

$$\text{total volume of pipe} = 0.032D^2 \text{ ft}^2 \times 3,168,000 \text{ ft}$$

$$= 101,376D^2 \text{ ft}^3$$

$$\text{total weight of steel} = 101,376D^2 \text{ ft}^3 \times 490.75 \text{ lb}/\text{ft}^3$$

$$= 49,750,272D^2 \text{ lb}$$

$$\text{total pipeline cost} = \$1.00/\text{lb} \times 49,750,272D^2 \text{ lb}$$

$$= \$49,750,272D^2$$

5. Annual equivalent cost calculation

$$\text{capital cost} = \left(\$49,750,272D^2 + \frac{\$170,004,700,125}{D^4}\right)(A/P,10\%,20)$$

$$= 5,843,648D^2 + \frac{19,968,752,076}{D^4}$$

$$\text{annual power cost} = \frac{\$114,557,013,315}{D^4}$$

6. Economical pipe size

$$AE(10\%) = 5,843,648D^2 + \frac{19,968,752,076}{D^4} + \frac{\$114,557,013,315}{D^4}$$

To find the optimal pipe size (D) that results in the minimum annual equivalent cost, we take the first derivative of AE(10%) with respect to D, equate the result to zero, and solve for D.

$$\frac{dAE(10\%)}{dD} = 11,687,297D - \frac{538,103,061,567}{D^5}$$

$$= 0$$

$$11,687,297D^6 = 538,103,061,567$$

$$D^6 = 46,041.70$$

$$D^* = \mathbf{5.9868 \text{ ft}}$$

Note that velocity in a pipe should be ideally no more than around 10 ft/sec due to friction wearing in the pipe. To check to see if the answer is reasonable, we may compute

$$Q = \text{velocity} \times \text{pipe inner area}$$

$$701{,}825 \text{ ft}^3/\text{hr} \times \frac{1}{3{,}600} \text{hr/sec} = V \frac{3.14159(5.9868)^2}{4}$$

$$V = 6.93 \text{ ft/sec}$$

which is less than 10 ft/sec. Therefore, the optimal answer as calculated can be practical.

7. Equivalent annual cost at optimal pipe size
   - Capital cost:

$$\text{capital cost} = \left[\$49{,}750{,}272(5.9868)^2 + \frac{170{,}004{,}700{,}125}{5.9868^4}\right](A/P, 10\%, 20)$$

$$= 5{,}843{,}648(5.9868)^2 + \frac{19{,}968{,}752{,}076}{5.9868^4}$$

$$= \mathbf{\$224{,}991{,}039}$$

   - Annual power cost:

$$\text{annual power cost} = \frac{114{,}557{,}013{,}315}{5.9868^4}$$

$$= \mathbf{\$89{,}174{,}911}$$

   - Total annual equivalent cost:

$$\text{total annual cost} = \$224{,}991{,}039 + \$89{,}174{,}911$$

$$= \mathbf{\$314{,}165{,}950}$$

8. Total annual oil revenue

$$\text{annual oil revenue} = \$18/\text{bbl} \times 3{,}000{,}000 \text{ bbls/day} \times 365 \text{ days/year}$$

$$= \$19{,}710{,}000{,}000/\text{year}$$

There are enough revenues to offset the capital as well as operating cost.

Comments. A variety of other criteria exists for choosing pipe size for a particular fluid transfer application. For example, low velocity may be required where there are erosion or corrosion concerns. Alternatively, higher velocities may be desirable for slurries where setting is a concern. Constructional ease will also weight significantly in the choice of pipe size. A small pipe size may not accommodate the head and flow requirements efficiently, whereas space limitations may prohibit selecting large pipe size.

17.7 Project Management

Donald D. Tippett

Engineers, Projects, and Project Management

For a majority of engineers, project work is a way of life throughout their professional careers. Projects build bridges, put men in space, design and install computer-integrated manufacturing systems, perform R&D on state-of-the art materials, and pursue a thousand other objectives in organizations large and small across the world.

A project has a specific set of objectives and a definite schedule, budget, and set of performance requirements. It has a finite life span and is usually a team effort led by a recognized project manager.

Project management is planning, organizing, leading, and controlling organizational resources to successfully complete a project.

Project Planning

The two most important facets of project management are planning and team building (to be discussed in the section on team building).

Project Charter. The first step in a sound planning process is the issuance of a project charter. A project charter is a formal establishing of the project by the upper management chartering authority (a project is not really a project unless it has management's sanction). As a minimum, the charter should: (1) designate the purpose of the project, (2) establish the general organizational format for the project, (3) appoint a project manager, and (4) state management's support for the project.

All projects should be formally launched with some sort of charter. For small in-house projects, it may be only a brief memo. Larger projects will have much more extensive charters. A charter puts the organization on notice that the project is being launched and that it has management's acknowledgment and commitment. The charter is the project manager's action authority to perform the project.

What is Project Planning?

- Planning is determining what needs to be done, by whom, and by when.
- Planning is decision making based upon futurity.
- Planning is selecting enterprise objectives, and establishing policies/procedures/programs necessary for achieving them.
- Planning is an iterative process and must be performed throughout the life of the project (Kerzner, 1998).

Each of these statements adds to an understanding of what project planning is. But why plan? There are a number of reasons:

- The planning process serves as a simulation of the project in the planners' minds before significant resources are expended. Thus false starts are avoided and problems are identified and resolved when solutions are the easiest.
- The planning process serves as a means of communications, informing everyone on the project what will be expected of them.
- Plans identify inconsistency and risk.
- Plans provide the basis for action.

Plans must be flexible, and provisions should be provided to update them as needed throughout the project. They should include reasonable contingencies and usually require several iterations before they are acceptable.

Whenever possible, the people who will do the work should plan the work. This develops a sense of ownership of the tasks among those who will actually execute them (Rosenau, 1992).

Senior managers tend to apply pressure on the project manager to cut the planning process prematurely short. However, cutting the planning process short is usually false economy. As much as 50% of direct labor hours and dollars can be spent before project execution begins, and companies that spend significantly less usually find quality problems during execution (Kerzner, 1998).

Project Planning Steps. The steps in sound project planning are:

Establish objectives
Develop statement of work
Prepare detailed specifications
Establish project milestones
Create the work breakdown structure
Establish detailed schedules
Establish detailed budgets
Define responsibilities
Replan as required

Objectives. Project objectives are the basis for all project planning. Objectives must be clearly stated, specific, measurable, and achievable. They should also be realistic and agreed upon by all parties concerned. Objectives should be consistent with the resources available and with other organizational plans and projects.

Statement of Work/Specifications. The statement of work (SOW) specifically states what the project will do for the customer. For large projects it will be a formal part of the contract. For projects within the organization it might be a memorandum. The SOW specifies the reason for establishing the project, the desired end results, and the performance, budget, and schedule goals. It may also include specific acceptance criteria, as well as a management section which discusses client relationships, management philosophies, project organization and personnel reporting, contingencies, and communications. The establishment of a satisfactory statement of work may require several iterations. However, it is critical that all parties have a detailed understanding of what is to be done before work proceeds.

Depending upon the size of the project, detailed specifications may be listed separately, or included as part of the SOW. They define the standards which must be met with respect to materials, labor, prices, equipment, support, etc. It is important to firmly establish detailed specifications up front, as small changes in specifications can have major effects on budgets and schedules.

Project Milestones. Project milestones establish the start and end dates of the project (if known), as well as all the significant subsets. Milestones are designated points in time by which certain specific project tasks/accomplishments are to be completed. The set of milestones forms a group of waypoints which provide a basis for status assessment, management reviews, and replanning the project. Because they are simple to understand, they are an important tool to gauge progress.

Work Breakdown Structure. The work breakdown structure (WBS) is a systematic way of defining the component parts of the project. The WBS breaks down the work to be accomplished into a logical hierarchy of sets, subsets, etc., so that it is put into "bite-sized" chunks which can be easily understood by planners, schedulers, workers, and managers. More specifically, breaking the project's work into manageable pieces:

- Facilitates the planning process by describing the total job in terms of discrete tasks.
- Allows assignment of these tasks to specific groups and individuals.
- Forms the framework upon which the budget and schedule are built, since each WBS task can easily be estimated in terms of labor, duration, and other required resources.

- Facilitates the chore of scheduling, supervision, and project control and information systems by defining exactly who is responsible for each task and establishing a target completion time.
- Simplifies purchasing, scheduling, and staging of required materials.

In creating a WBS, the following principles should be kept in mind:

- Routine or repetitive work should not be excessively subdivided.
- Subdivide each block of work to a level which is useful to project management; all blocks do not have to have the same number of subdivisions.
- Each task should be easily understood and should have an identifiable schedule.
- The cost of the smallest subtask should be significant.

It is essential that a work breakdown structure be created. The WBS forms the skeleton upon which the plan, schedule, budget, and control system are built. In addition, a completed WBS is the basis of entering project information into any of the myriad of computerized planning and scheduling systems now being used by virtually all project managers.

Scheduling/Budgeting/Responsibilities/Replanning. The remaining components of project planning — developing a schedule and budget, and assigning responsibilities — naturally follow the creation of the WBS. Once the project's work has been separated into a set of manageable subtasks, it is a straightforward process to set about estimating the time and resources required to complete each one (see the section on estimating). These estimates are then compiled into a project budget. Similarly, responsibility for each WBS component can be assigned to individuals and groups as appropriate and compatible with available resources. While the scheduling process involves a few more intermediate steps (see the next section), it too is a logical extension of the work breakdown structure. Much of the project plan depends on the WBS. Therefore project managers should pay close attention to it.

Since plans relate to future events but are based upon current knowledge, it follows that plans will need to change during the course of the project. As the project team learns lessons along the way, and as events unfold in the dynamic project environment, project plans will need to be adjusted accordingly. Vehicles for accomplishing this process should be established at the outset. Plans must be kept current. Out-of-date plans hurt project credibility. Besides, the project team needs an accurate roadmap if it is to arrive successfully at its desired destination.

Project Scheduling

Creating a viable schedule and adjusting it as necessary throughout the life of the project is an important part of the project manager's responsibilities.

The scheduling cycle (see Figure 17.7.1) begins with the work breakdown structuring process, which subdivides the project's work in to a set of manageable and easily understood tasks. The next step is to estimate the time, man hours, and other resources required to complete each task. Then the relationships between tasks must be documented. Many can be performed in parallel while some must necessarily come before others. In any case, each task's interdependence with every other task must be determined. The project manager is now ready to create the actual schedule, using one of the methods discussed in the next section. Once the schedule is formulated, it must be checked against available resources for feasibility. If the schedule contains periods of time during the project when more resources are called for than will be available, the scheduler must iterate back to an earlier part of the scheduling cycle and make adjustments so that the eventual schedule will be compatible with available resources. Once this is accomplished, the project manager should make a final check to ensure that the proposed schedule will in fact satisfy the project constraints (budget, time, and performance criteria). If constraints are met, the schedule can be finalized.

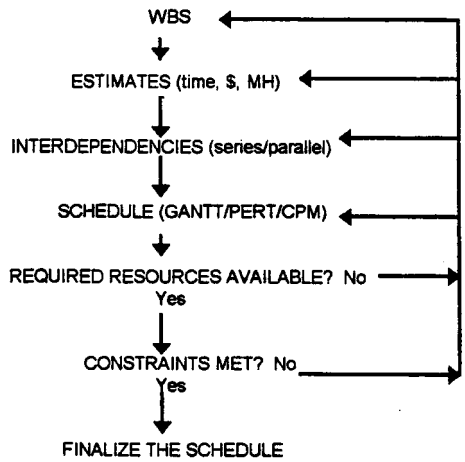


FIGURE 17.7.1 Project scheduling and estimating cycle.

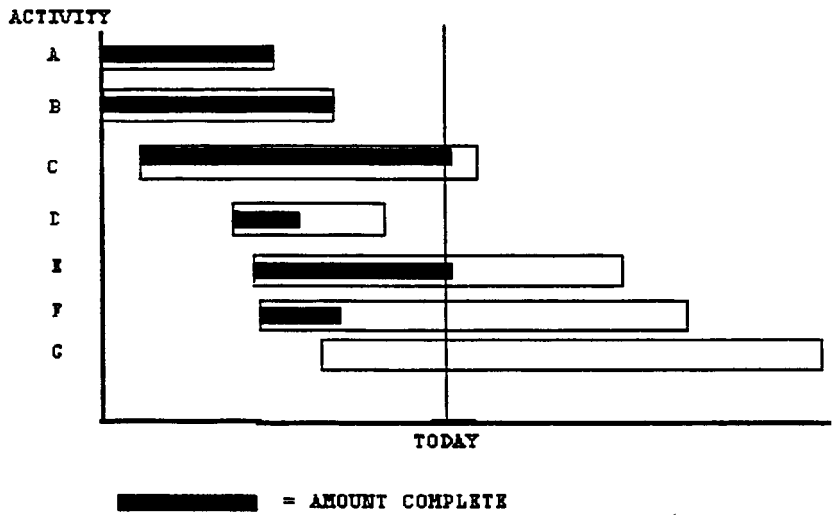


FIGURE 17.7.2 Gantt chart example.

Scheduling is a continuous process during the life of a project. In the dynamic project environment, changes are always occurring which necessitate replanning and rescheduling. Mechanisms that keep the project plan and schedule valid and current should be put in place at the outset of the project.

Scheduling Methods. The Gantt chart or bar chart (see Figure 17.7.2) was one of the first techniques used by project managers to schedule and control projects, and it is still widely used. It is simple, graphic, easy to construct and change, and convenient for displaying progress. Project task times are usually shown as horizontal bars that begin and end at the scheduled activity start and end times, respectively. However, Gantt charts do not do a good job of displaying task dependencies and interrelationships. Furthermore, the critical path, the set of sequential activities which define the ultimate length of the project, is not obvious (Rosenau, 1992). So, while the Gantt chart is a useful tool for quick overviews and broad statusing of projects, it needs to be augmented by a more comprehensive scheduling tool in all but the most elementary project management situations.

Critical path method (CPM) and program evaluation and review technique (PERT) are two critical path systems developed to address the shortcomings of the Gantt chart, and enable project managers to

do a better job scheduling complex projects. They both show project task durations, and indicate the earliest task start times as the Gantt chart does, but they also show the *precedence* between activities. This allows the project manager to determine the most critical activities and to more easily modify and determine the status of the schedule.

CPM assumes activity durations are known with a reasonable degree of certainty, and are thus useful in industries like construction that have a long history of performing the various tasks to draw from in formulating accurate project activity estimates (e.g., dig basement, pour footer, etc.). On the other hand, PERT is designed to be used in applications like R&D and aerospace wherein many activities are being performed for the first time and there is no history to draw upon in formulating activity duration estimates. PERT employs statistical techniques to assist the scheduler to devise reasonable activity duration estimates and an estimate for the overall project duration.

Figure 17.7.3 depicts a simple CPM network drawn in the activity on node (AON) format. The nodes or circles denote activity. The connecting arrows show precedence between activities. There are various possible routes through the network. The longest route defines the predicted total project duration. The activities on this longest route, or critical path, are called critical activities. Any delay in the execution of these critical activities translates into a delay in overall project completion. Other non-critical activities have a certain amount of leeway (or slack) associated with them which allows the project manager some flexibility in scheduling resources. Thus the project manager should pay special attention to critical activities in making tradeoff decisions, as these are the ones that can potentially extend the project duration. However, it should be noted that non-critical activities can potentially become critical if they experience delays greater than their original slack, so the project manager cannot become complacent about noncritical activities either.

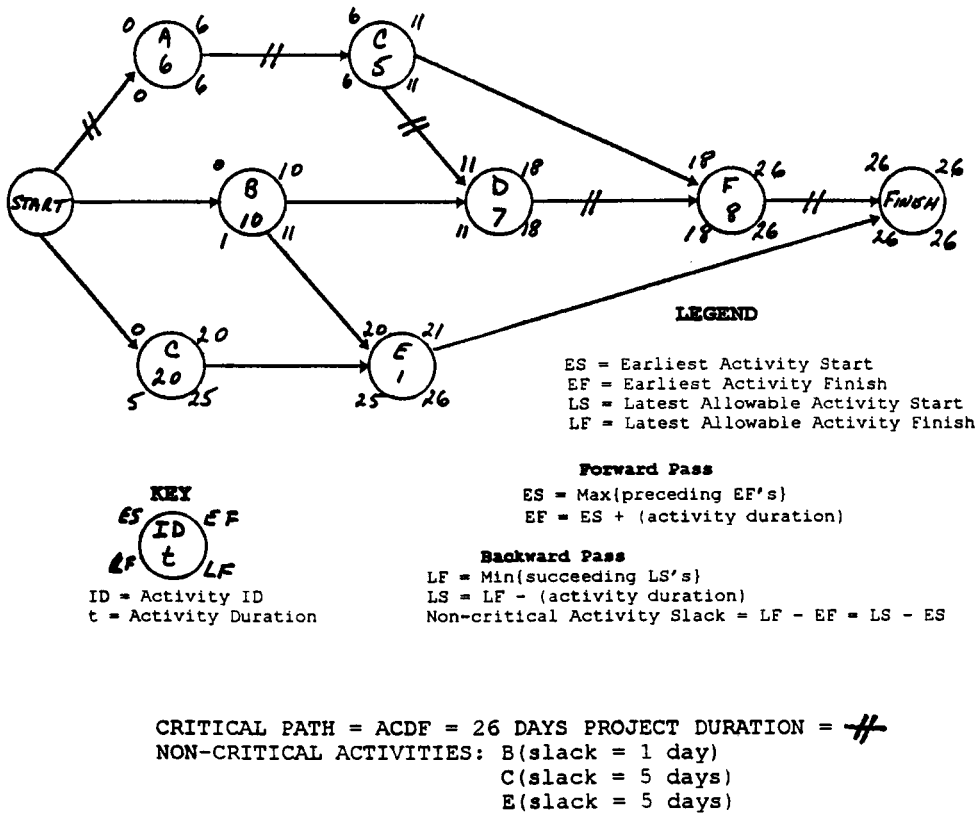


FIGURE 17.7.3 Example of AON CPM network.

In solving the network, forward and backward passes are made according to the formulas given in Figure 17.7.3. Together these two passes determine the earliest and latest starts and finishes for each activity. Also determined are the project's duration, and the critical path (path that contains activities without slack), which represents the predicted project duration. In the example, activities A, C, D, and F have no slack and are thus critical. The path ACEF is the critical path and defines the project duration, 26 days. Once created, the network can be continually updated and used during the life of the project to assist the project manager to understand current project status and make informed tradeoff decisions.

It is not feasible to solve any but the most simple PERT/CPM networks by hand. Fortunately, software is widely available for modern personal computers which is capable of constructing and solving these networks for small and medium sized projects. Very large project networks require the power of mainframe computers. The Dreger (1992) text presents comprehensive examples using typical personal computer project management software in realistic project situations.

A more detailed exposition of PERT/CPM techniques can be found in Moder et al., 1983.

Staffing and Organizing

In staffing and organizing a project, management must bring together a diverse group of people, equipment, information, and other resources that, when properly integrated, is capable of achieving the project objectives. Each element of the resulting system is critical to overall project success.

Selecting the Project Manager. The project manager is the most critical selection. A person should be selected who:

- Serves full time on the project. With few exceptions, appointing a part-time project manager is not practical for projects of any significance. Naming a part-time project manager signals the organization that the project is a low priority. This practice is not fair to the individual or to the project, and usually produces less than optimum results.
- Has a proven track record. Most project situations are poor training grounds for project managers with little or no experience or training. A project manager's credibility with the organization is vital.
- Functions as a manager, not a doer. If the project manager spends all of his or her time performing technical tasks, then, in effect, there is no project manager. With no one in the driver's seat, budgets, schedules, customer relations, and other critical project management components are left unattended.
- Possesses good interpersonal skills. A project manager may be called upon at any time to be a communicator, mediator, negotiator, motivator, coach, counselor, and leader.

Project Team Staffing. The project manager may have little or no control over the composition of the project team, as the functional managers of the contributing departments are often the ones who decide which individuals serve on the project. Hence, project managers must be able to mold a potentially broad mix of talents, functional backgrounds, and professional experience into an integrated team. This is one reason why team building skills are vitally important. Given a choice, the project manager would opt for project team members who are assigned full-time to the project, experienced in project work and their discipline, are team players, and relate well with customer representatives when direct interchange is required.

Organizing for Project Management. Projects are conducted under an infinite variety of organizational structures. Project managers usually have little input in this area and must learn to make the best of the organizational structure in which they find their project operating. However, project managers should bear in mind that, while organizational structures may serve either to facilitate or impede project progress, they are usually not a principal factor in determining overall project success or failure. Other factors like thorough planning and building a cohesive, motivated project team are more predictive of project

TABLE 17.7.1 Advantages and Disadvantages of the Three Main Organizational Forms from a Project Management Perspective

Organizational Structure	Advantages	Disadvantages
Functional/Hierarchical/ Classical	* Close control of budget and people. * Well-defined authority and communications channels. * Well-understood career paths. * Best for mass production. * Nurtures functional areas; assures up-to-date technology. * Provides security for individuals (in functional groups)	* No one person truly in charge of the entire project. * Long lead times, ineffective integration, very slow internal communications. * Ineffective communications with the customer. * Difficult to establish status of projects. * Very slow to respond to change. * Strongest functional group tends to dominate decision making. * Poor conflict resolution, great reliance on hierarchical referral.
Pure Project/Product/ Divisional	* One individual in authority to speak for the project. * Quick response to changing environment. * Personnel allocated completely to the project * Good communications. * Less need for hierarchical referral to resolve conflict. * Most likely to complete the project on time.	* Less efficient use of resources. * Does not nurture functional disciplines. * Best when company is project-oriented. * Uncertainty for personnel when project ends (where do they go?).
Matrix	* More efficient: project can share personnel and more easily accommodate changing requirements. * Good communications. * Personnel security (people have homes to go to after project is finished). * Technical disciplines are supported/nurtured. * Quick to respond to change. * One person speaks for the project.	* High potential for conflict. * Project personnel must work for two bosses. * Heavy penalty in communications requirements and other administrative * overhead costs. * Best for project-oriented organization. * Competition for best resources. Upper management may use matrix as requiring them to give up some authority.

success. Besides, as Kerzner (1998) so aptly points out, any organizational form will work if the people want it to.

Nevertheless, project managers should be cognizant of the strengths and weaknesses of the organizational structure under which they are operating so they can make the best of their situation. Table 17.7.1 gives some of the advantages and disadvantages of the three principal organizational structures and their many variants, from a project management perspective.

Regardless of the organizational form used, it is a good idea to create a steering team to oversee and guide important projects. A steering team of senior managers demonstrates management support of the project, helps with integration and hard problems, provides oversight and guidance, makes major dollar decisions, and acts as an interface with upper management.

Team Building

Today's projects characteristically require the efforts of many people from multiple disciplines whose efforts must be effectively integrated to meet project objectives. Further, project managers must cope with dynamic environments, high degrees of organizational complexity, and ever-increasing competitive pressures. Experienced project managers know that the only way to effectively deal with these challenges

is to mold their project personnel into true project *teams* or groups of people with complementary skills who are committed to a common purpose, set of performance goals, and approach for which they hold themselves mutually accountable (Katzenbach and Smith, 1993). Teams feature enhanced efficiency, increased motivation, self-regulation, synergistic output, flexibility, and heightened confidence (Raudsepp, 1984), all of which are vital to project success in the present competitive environment. Thus, team building is an essential project management skill.

Robert P. Hagen has advanced six key elements of most successful team building plans which, if implemented by project managers, would greatly enhance the state of team building on project teams (Table 17.7.2).

TABLE 17.7.2 Six Key Team Building Elements

1. In all actions, demonstrate respect and consideration for all employees as valued members of the team.
   Are employees encouraged by example and admonition to respect each other? Do they know enough about each other's job to appreciate the contributions others are making? Does a general atmosphere of consideration exist?
2. Identify individual job responsibilities and performance standards and see that they are known.
   Are individual discussions held to ensure that each employee knows their job's standards and responsibilities? Does each team member understand how their portion of the project is important to overall project success?
3. Work to secure good communications with employees as individuals and as a team.
   Do team members feel their inputs and suggestions are valued? Do they receive regular feedback on how they are doing? Is advance warning of changes conveyed whenever possible along with reasons why changes are necessary? Are regular exchange meetings held? Are team members included in decision-making?
4. Establish individual and group goals, preferably in coordination with those concerned.
   Are *individual* goals established for each team member? Is consideration given to each individual's opportunities for professional development? Are *group* goals established and communicated to the team? Is a goal established that encourages growth in team development factors like team planning, conflict resolution, and problem solving?
5. Reward teamwork and team building efforts.
   Who issues rewards? Project managers? Functional managers? Are rewards mostly based on extrinsic job factors like pay, bonuses, and working conditions, or are they based on intrinsic job factors like accomplishment, recognition, responsibility, and growth? Does management recognize the difference between rewarding an individual as a member of a team, and rewarding an individual as an individual? Are individuals singled out and rewarded for their performance on the project team? Do team members have input into what and how rewards are given and to whom?
6. Practice and encourage loyalty to the team.
   Does the project manager defend team members against unfair criticism? Is there a climate of trust on the project team? Is effective leadership practiced by project managers?

Project Control

Planning puts the team in a position to launch a project. However, rarely does everything go according to plan. As soon as the project begins, deviations start to occur. Thus, once the project is launched, conforming to the plan becomes a principal function of project management.

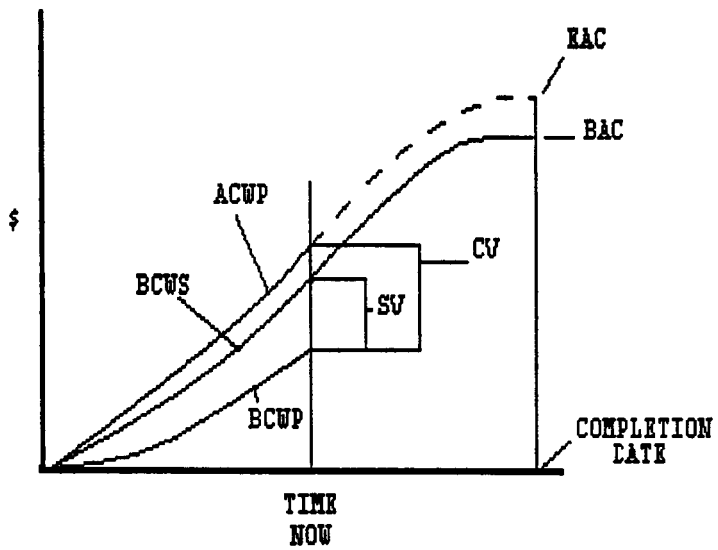
It is up to the project control system to keep the project manager apprised of how all components of the project are progressing, and highlight significant problems and deviations from plan, so that corrective action can be taken. Without a control system, the project manager has little influence over the project, and it will meander to completion in some form or fashion.

An effective project control system combines a cost accounting function and a cost control function with appropriate mechanisms for monitoring progress against schedule and technical performance targets. The cost accounting function accumulates actual costs, ensures costs are properly allocated, and verifies work is carried out and billed correctly. The cost control function provides the information to support cost analysis, prediction, and reporting. Costs are maintained within budget by taking corrective action wherever predicted costs exceed the budget.

Among the most critical requirements of a project control system is providing a capability to compare budgeted costs to actual costs, and then conduct variance analyses which lead to corrective action. The earned value approach to variance analysis is widely used. It compares the budgeted cost of work performed (BCWP), or earned value, with the actual cost of work performed (ACWP) to determine the cost variance (CV). It compares the BCWP with the budgeted cost of work scheduled (BCWS), or planned earned value, to determine the schedule variance (SV).

ACWP and BCWS are fairly straightforward to obtain. On the other hand, it may be difficult to obtain an accurate representation of BCWP. One approach is to use standard dollar expenditures for the project, so that x% of the costs are booked for x% of the time. Another is the 50/50 rule, wherein half the budget is recorded when an activity is scheduled to begin and the other half when the activity is scheduled to complete. Other alternative approaches may depend on percentage of man hours expended, parts installed, programs written/tested, etc. There are many other possible approaches; however the object is to somehow measure the percentage complete for project activities (or *earned value*). On some projects the time, trouble, and expense to obtain an accurate BCWP may not be cost effective.

Figure 17.7.4 gives a graphic presentation of the earned value approach and summarizes some quantitative indices useful to project managers conducting variance analysis. Once identified, each critical variance must be investigated to determine the cause and the appropriate corrective action. Such an investigation should begin at the lowest organizational level by the supervisor involved. It then progresses upward to the project manager, who reviews variance causes and approves corrective actions. Normally the project manager will review these results with upper-level management prior to preparing the contractually required reports to the customer.



BCWP = BUDGETED COST OF WORK PERFORMED = EARNED VALUE
BCWS = BUDGETED COST OF WORK SCHEDULED = PLANNED EARNED VALUE
ACWP = ACTUAL COST OF WORK PERFORMED
CV = COST VARIANCE = BCWP - ACWP
SV = SCHEDULE VARIANCE = BCWP -BCWS
CV% = COST VARIANCE PERCENTAGE = CV/BCWP
SV% = SCHEDULE VARIANCE PERCENTAGE = SV/BCWS
EAC = ESTIMATE AT COMPLETION = (ACWP/BCWP)*(TOTAL BUDGET)
(NEGATIVE IS UNFAVORABLE)

FIGURE 17.7.4 Variance analysis indices and graphics.

Estimating and Contracting

Project Estimating. Estimating time and resource requirements on projects is frustrating because the project manager knows in advance that the probability of being exactly correct is slim. The key, then, is to not be afraid to be wrong, but rather simply strive to be as accurate as possible, given what is known at the time.

A simple three-step approach will serve as a starting point for estimating:

1. Top-Down Estimate — Make an initial estimate at an overview level…an order-of-magnitude estimate using the broad, big picture view.
2. Detailed Bottom-Up Estimate — Follow the top-down estimate with a detailed estimate built from the bottom up. Start with the lowest tasks in the work breakdown structure and work up by task, by department, and so on, arriving at a total figure.
3. Compare the two estimates.

In effect, each estimate provides a reality check for the other. If the two are in relative agreement, it is likely that a good estimate has been developed. However, if the two figures are very different, more analysis is indicated. Go back and examine the individual work packages to discover where costs may differ, and to what extent there have been incorrect assumptions made as to the scope and content of work called for. Once the areas of principal differences between the two estimates have been determined, decisions can be taken as to which is correct.

Contingency planning is an important part of estimating. It is usually not appropriate to place a blanket contingency figure (e.g., 10%) across all tasks. Rather, contingency budgets should be set deliberately considering each activity individually, being careful not to build contingency on top of contingency.

In developing detailed estimates, time and dollars to cope with the following issues should not be overlooked (Rosenau, 1992).

- Interfacing with customers, upper management, and other departments
- Obtaining customer furnished items
- Getting approvals
- Working at remote/difficult locations
- Training people
- Making mistakes
- Obtaining security clearances
- Placing major subcontractors and purchase orders
- Replacing sick and vacationing people

Contracts. There are many types of contractual agreements used on projects. However, most can be categorized as variations upon two basic types of contracts: fixed-price and cost-reimbursable. Under a fixed-price contract the customer agrees to pay a specific amount upon satisfactory completion of the project. A cost-reimbursable contract is one in which the customer agrees to reimburse the contractor organization for costs actually incurred in performing the project.

The fixed price contract presents the lowest financial risk to the customer because the maximum financial liability is specified. This form presents the highest risk to the contractor organization, though it also opens the door for increased profits if the contracted costs can be underrun. On the other hand, under a cost-reimbursable contract, the customer is obligated to reimburse the contractor for all costs incurred, and thus bears considerable risk. The usual arrangement under cost-reimbursable contracts is to add a fixed fee or an incentive fee to the costs to arrive at the total compensation figure due the contractor. Cost-reimbursable contracts are traditionally used when the nature of the project dictates that accurate pricing cannot be determined.

Market conditions also influence the type of contract used. When work is scarce, customer organizations often insist upon fixed-price contracts. Conversely, when business is good, customers are not able to insist upon fixed-price contracts, and more cost-reimbursable type contracts are used.

Summary

Project management is among the most challenging undertakings in the field of management today. Project managers routinely face formidable performance requirements, firm budget constraints, and tight schedules. They must organize people from diverse technical backgrounds into a cohesive, motivated team, and then lead that team as it plans and executes its complex tasks in an ever-changing environment.

While the most successful technical people are often promoted to project management, it takes more than a solid technical background to make an effective project manager. To qualify an individual for project management, sound training in the fundamentals is essential, along with on-the-job experience as a project team member and/or assistant project manager or project engineer.

The above constitutes but the tip of the iceberg in terms of required knowledge for project managers, but it will perhaps give the uninitiated a starting point. The excellent references listed below will serve to fill in many of the gaps for the interested reader.

Defining Terms

Activity on Node (AON): A scheduling network in which the circles (or nodes) represent activities, while the connecting lines denote the sequence the activities must follow (see Figure 17.8.3).
Actual Cost of Work Performed (ACWP): Actual dollars expended on a given project activity.
Budgeted Cost of Work Performed (BCWP): Earned value of completed work.
Budgeted Cost of Work Scheduled (BCWS): Planned earned value.
Gantt Chart: The Gantt chart, named after Henry L. Gantt, is a bar chart useful in project management to depict planned and actual progress of project activities on a horizontal time scale (see Figure 17.8.2).
Network: A combination of arcs and nodes linked together to represent project activity durations and precedences. Used in conjunction with a solution technique like CPM or PERT, estimates of project completion times, critical activities, and resource requirements can be determined.
PERT/CPM: PERT stands for program evaluation and review technique, while CPM stands for critical path method. Both are methods for solving a project activity network. CPM is used on projects where activity time estimates are well established, like on construction jobs. PERT is used on projects where times to complete activities are not known with a great deal of certainty, such as in aerospace and R&D applications. Almost always used in conjunction with computer software, these techniques yield estimates of overall project completion times, listings of the most critical activities, and resource requirements by time frame.
Statement of Work (SOW): Formal statement delineating specifically what the project will do for the customer. As a minimum it should contain performance, budget and schedule objectives.
Work Breakdown Structure (WBS): A logical division of the work involved in the project into a hierarchy of component parts. It forms the basis for subsequent project budgeting, scheduling, and controlling.

References

Dreger, J.B. 1992. *Project Management: Effective Scheduling*, Van Nostrand Reinhold, New York.
Hagen, R.P. 1985. Team Building, *Management*, First Quarter.
Katzenbach, J.R. and Smith, D.K. 1993. The discipline of teams, *Harvard Business Review*, Mar/Apr.
Kerzner, H. 1998. *Project Management: A Systems Approach to Planning, Scheduling, and Controlling*, 6th ed. New York: Van Nostrand Reinhold.
Meredith, J.R. and Mantel, Jr., S.M. 1995. *Project Management, A Managerial Approach*, 3rd ed. New York: John Wiley & Sons.

Moder, J.J., Phillips, C.R., and Davis, E.W. 1983. *Project Management With CPM, PERT, and Precedence Diagramming*, 3rd ed. New York: Van Nostrand Reinhold.

Raudsepp, E. 1984. Effective teamwork, *Manage.*, April.

Rosenau, M.D., Jr. *Successful Project Management*, 2nd ed. 1992. New York: Van Nostrand Reinhold.

Shtub, A., Bard, J., and Globerson, S. 1994. *Project Management: Engineering, Technology, Implementation.* Englewood Cliffs, N.J.: Prentice-Hall.

Taylor, L.W.; et. al. “Communications and Information Systems”
Mechanical Engineering Handbook
Ed. Frank Kreith
Boca Raton: CRC Press LLC, 1999

Communications and Information Systems

Lloyd W. Taylor
DIGEX, Inc.

Daniel F. DiFonzo
Planar Communications Corp.

A. Brinton Cooper III, PhD
Bel Air, Maryland

Dhammika Kurumbalapitiya
Harvey Mudd College

S. Ratnajeevan H. Hoole
Harvey Mudd College

18.1 Introduction

This chapter provides a broad introduction to and reference resource for the field of communications and information systems.

The first section covers the areas of computer networks and their underlying technologies. These technologies include electrical, optical, wireless, and satellite communications channels, as well as the protocols (such as TCP/IP) used to transfer information over these channels. The purpose of this section is to provide a high-level understanding of the infrastructure which underlies all modern electronic communications.

The second section introduces Communications and Information Theory. This section provides the mathematical background necessary to better understand the technologies used in electronic communications. Issues such as noise, compression, and error correction are explained. Of necessity, this section is somewhat more theoretical than the others.

The third section concludes the chapter with information on two applications of computers and networking. The first, *Accessing the Internet*, introduces the Internet, provides a summary of the software tools used to access information on the Internet, and discusses methods of finding information using a World Wide Web browser. The second application, *Data Acquisition*, describes the components of a data acquisition system and shows how they follow the model of a computer network.

18.2 Network Components and Systems

There are a variety of network architectures, standards, and transmission methods used to interconnect computers and communications systems. This section provides the information necessary to untangle the internetworking web. It begins by surveying the electrical and optical communications standards and architectures commonly used in modern networks, moves on to discussions of optical fibers and satellite communications, and concludes with a discussion of computer communications standards.

Electrical and Optical Communications

Lloyd W. Taylor

The field of networking is generally broken into two domains: the **local area network** and the **wide area network**. The distinction between the two tends to be rather fuzzy. In general, *local area* refers to networks within the same building, and *wide area* refers to networks that interconnect buildings, cities, or countries.

Local area networks are generally faster than wide area networks. A common rule of thumb used by network engineers is that speed and distance are inversely related; that is, the longer the distance the network must travel, the lower the speed that will be used. This is not an electrical or physical limit, but rather a financial one. Longer links and higher speeds simply cost more.

Cabling Architectures

All networks are based on underlying cabling architectures. These architectures define how the nodes are interconnected and may also limit the choices of network electrical standards. For example, ethernet requires a bus architecture, but can be implemented in a star architecture with the proper network electrical equipment. Ethernet cannot be implemented over a ring architecture.

In a *bus architecture* (Figure 18.2.1), all members of the network share a common cable for their network signaling. The bus is a shared resource. All traffic can be read by every node.

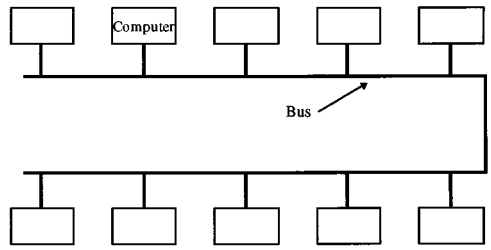


FIGURE 18.2.1 Bus architecture.

For a bus-oriented network (such as ethernet) to work, there must be a mechanism for sharing access to the bus. In the ethernet world, the standard for this mechanism is carrier-sense multiple access/collision detection (CSMA/CD). When a computer wishes to transmit on the bus, it first listens to see if the bus is busy. If not, the computer begins to transmit while simultaneously listening to its own message. If the message is heard back as sent, then the transmission has been successful. If there are errors, it is likely that another computer decided to transmit at the same time, resulting in a collision. When a collision occurs, all currently transmitting computers stop transmitting for a random amount of time and then begin the process again by listening to see if the bus is busy.

Bus architectures are useful where no one node requires a large portion of the shared bandwidth. If one or two nodes are constantly using the majority of the available bandwidth, performance for all connected nodes can become seriously affected.

Bus architectures also are relatively insecure. It is quite simple to run a program to monitor all traffic on a bus and capture sensitive information such as user IDs and passwords. Because all systems must have access to all transmissions (to determine whether or not the bus is busy), there is no simple way around this limitation.

In a *ring architecture* (Figure 18.2.2), each node has two connections, an inbound connection and an outbound connection. Network traffic passes through each node until it reaches its destination.

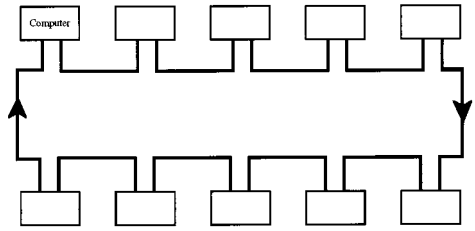


FIGURE 18.2.2 Ring architecture.

For a ring architecture to work, there must be some method of determining which node has the right to transmit at any given time. A common method to do this is by passing a **token** around the ring. When a node wishes to transmit, it captures the token and sends on its message in the token's place. When the message reaches the intended destination, the recipient node marks it "read" and sends it on around the ring. When the originating node receives back its own message, it removes it from the ring and replaces the token on the ring. Note that the node that has just transmitted will not have another chance to transmit until all other nodes on the ring have had a chance to transmit. In other words, until the token makes it all the way around the ring, the first node may not transmit again. This has the effect of making a ring architecture fairer than a bus architecture, as no one node can co-opt the entire available bandwidth.

Ring architectures suffer from the same security problems as bus networks: all traffic flows through all nodes. In addition, ring architectures are more difficult to install and maintain. If any one node becomes disconnected from the ring, the entire ring can fail, as there will be no path for the data. Some rings are implemented redundantly, with two separate traffic paths rotating in opposite rotations. Should any one node fail, the nodes on either side of the failure will wrap the ring, sending data back around the second ring to reach the computer on the other side of the failed node. In this way, no single-point failure can disrupt the ring.

In a *star architecture* (Figure 18.2.3), each node has a separate connection to a central point. Each of these connections is to an **active hub**, which is a networking device designed to manage and switch network traffic.

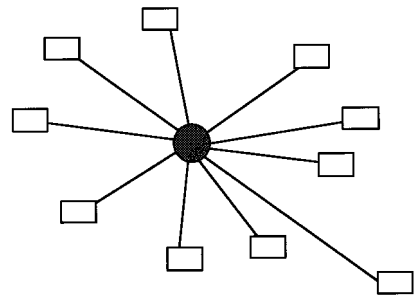


FIGURE 18.2.3 Star architecture.

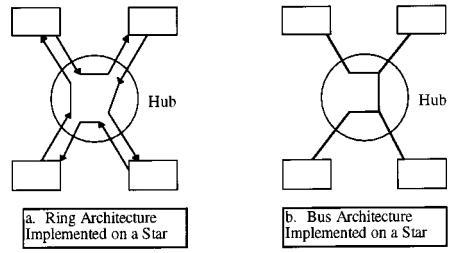


FIGURE 18.2.4 Star implementation of other architectures.

A star architecture can be used to implement both bus and ring architectures. For a ring architecture (Figure 18.2.4a), the central hub simply sends data out one wire to the computer and receives it back on another wire. For a bus architecture (Figure 18.2.4b), the hub electrically or logically interconnects all wires together.

For implementations of both architectures, the hub can monitor the bandwidth being used by each attached node, and in some implementations can actually limit the total bandwidth used by each. Many hubs can also "censor" the data going to each node, sending unchanged data only to the destination system. All other nodes get scrambled data, making it impossible for them to capture sensitive information.

Most larger networks are now implemented as stars, because of the increased flexibility and manageability of a star architecture. The same cabling system can be used to implement bus and ring architectures, as well as point-to-point connections such as telephone lines. In addition, the costs of operation are lower, as all moves and changes can be made in a central location, rather than requiring that new cable be pulled for each new computer or terminal.

Complex systems are larger networks generally using a combination of architectures. For campus-wide networks, a common architecture is the "ring of stars" (Figure 18.2.5a). In this architecture, a star architecture is implemented in each building, and the buildings are interconnected in a ring, usually using fiber optic cables. The key advantage of this configuration is that no single-point failure will disrupt the network. If the ring architecture has been implemented as a redundant ring, it will self-heal and route around the damaged section.

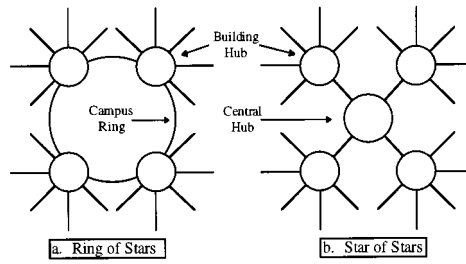


FIGURE 18.2.5 Complex architectures.

Another common architecture is the "star of stars" (Figure 18.2.5b). This architecture is implemented identically in each building, but each building is interconnected to a single central location in a star configuration. The key advantage of this architecture is that it can support very high bandwidth networks. The key disadvantage is that the loss of the central hub will disrupt the entire network.

Local Area Networks

Copper Standards. Local networks are still most commonly implemented using copper wire, rather than fiber optics, because of the significantly higher cost of fiber. As the demand for higher bandwidth

connections to the desktop increases and the cost of fiber decreases, we can expect to see more fiber to the desktop.

There are three common categories of copper cable used in networking: unshielded twisted pair (**UTP**), shielded twisted pair (**STP**), and coaxial (**Coax**). Each of these has its unique advantages and disadvantages.

All copper cabling is available in two general grades of insulation. Normal (nonplenum) insulation is usually made from PVC (polyvinyl chloride) and may be used anywhere other than in air handling spaces. The more expensive plenum cable (Teflon®-based insulation) is required in air handling spaces because of its fire-resistant characteristics.

Unshielded twisted pair cable is the most commonly used network cable today. It comes in a variety of qualities, called categories. These categories (Table 18.2.1) define the electrical characteristics of the cable and specify the maximum data rate that they support.

TABLE 18.2.1 UTP Cable Categories

Category	# Pairs	Max. Bandwidth	Uses
1	Any	1 Mbps	Telephone
2	2–25	4 Mbps	Token Ring — 4 Mbps
3	1–6	10 Mbps	Ethernet
4	1–6	20 Mbps	Token Ring — 16 Mbps
5	1–6	100 Mbps	100BaseT, CDDI

UTP cable is always installed in a star architecture. It requires active electronics in the wiring closets to operate the connections to each computer.

As the category number gets higher, the manufacturing standards for such things as the number of twists per meter, variations in impedance, insulation consistency, etc. become stricter. This is because the cables must be of uniformly high quality to handle the higher frequencies required for higher data rates.

For higher category cabling, installers must be specially trained. The rules for installation of high-bandwidth cable limit the number of twists that can be unwrapped for termination, the minimum bend radius, and the maximum force that can be applied while pulling the cable through walls or conduits.

Once the cable is installed and terminated, it must be tested and certified to the appropriate level. Improperly handled or installed Category 5 cabling may perform at only a Category 3 level, through no fault of the cable itself. Testing includes time-domain reflectometer measurements, noise and crosstalk measurements, and impedance measurements.

UTP cable is specified for use in two common network electrical standards, **10BaseT** ethernet and **100BaseT** ethernet. IBM Token Ring can be run over UTP cable if an appropriate impedance matching transformer (usually called a *balun*) is used at both ends of the UTP run. In an increasing number of systems, these matching transformers are being integrated into the network interface cards and network hubs themselves.

Shielded twisted pair (STP) cable is identical to UTP cable with the addition of a shield around each twisted pair and usually another shield around the entire cable bundle. This shielding reduces susceptibility to electrical noise (such as that caused by heavy electrical machinery), provides a more predictable impedance, but also reduces the efficiency of the cable. A typical connection with STP cable can only run one third the distance of an equivalent UTP cable. This is because the capacitance of the STP is much higher than that of the UTP, resulting in three times the attenuation at a given frequency.

STP cable is commonly used in IBM Token Ring network installations. IBM-Standard STP cable is available in two versions: type 1 cable contains two pairs of 22-gauge wire, each with an individual shield. It is a heavy, stiff cable that is difficult to handle and pull. Type 9 cable is a lighter, more flexible version that is made with 26-gauge wire. The lighter gauge wires result in a higher impedance, limiting the length of a cable run to two thirds that of a type 1 cable.

IBM has defined a number of other cabling standards that include fiber optics, UTP, and mixed cable types within a single sheath.

Coaxial cable was the original cable used for ethernet. It has largely been superseded by UTP, but is still in use in older installations.

Coaxial cable can be used either in a bus or a star architecture. In a bus architecture, a heavy coaxial cable (often called *thickwire*) is run through the hallway, above the drop ceiling. Each end is terminated with a 50-Ω resistor to minimize electrical reflections. The coax is drilled wherever a connection to a computer is to be made, and an active tap is installed (Figure 18.2.6). The connection between the tap and the computer is via a 15-pin drop cable that carries signal and power to the active tap.

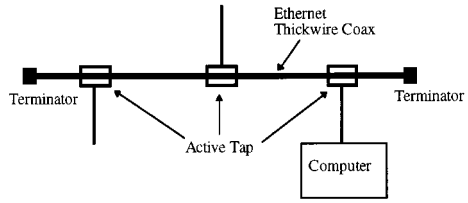


FIGURE 18.2.6 Ethernet thickwire coax network.

In a star architecture, a much lighter coaxial cable (often called "thinwire") is used. A standard BNC connector is installed on each end, and the cable is plugged directly into the network interface card in the computer and into the network hub in the wiring closet. Thinwire networks can be "daisy-chained," allowing several computers to share a single run of coaxial cable.

Fiber Standards. Fiber optics are playing an increasingly important role in computer networks. The most common use at present is in backbone and wide area networks. As costs decrease, we can expect to see fiber to the desktop become the norm.

How Fiber Optics Work. Fiber-optic cables are essentially "light pipes." They are made up of two coaxial layers of glass, the inner called the core and the outer called the cladding (Figure 18.2.7). Each of the components has a different index of refraction, resulting in a reflective surface at their interface.

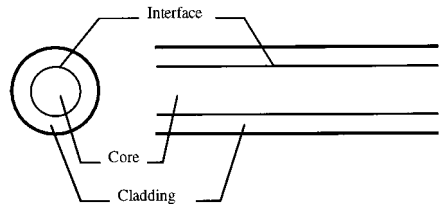


FIGURE 18.2.7 Optical fiber construction.

A modulated electrical signal drives a lightsource such as a laser diode, which injects a collimated beam of light into one end of the optical fiber. The lightbeam is reflected back and forth along the inner fiber and is coupled with an optical detector at the other end (Figure 18.2.8). The optical detector converts the lightbeam back to an electrical signal.

While optical fibers can carry signals much greater distances than can electrical conductors operating at the same frequency, they do suffer from attenuation just as do electrical conductors. The mirror surface at the interface between the core and the cladding is not perfect, and photons can escape through this interface and be lost in the cladding. Also, not all lightbeam frequencies propagate at the same rate through the fiber, resulting in degradation of the signal over long distances.

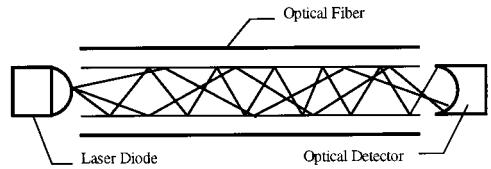


FIGURE 18.2.8 Optical transmission system.

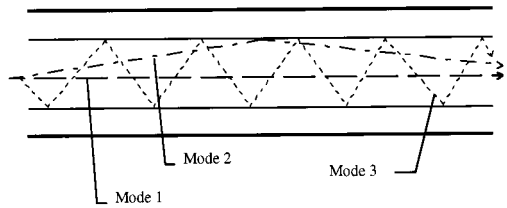


FIGURE 18.2.9 Multimode fiber.

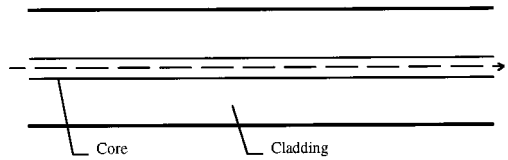


FIGURE 18.2.10 Single-mode fiber.

Types of Fiber-Optic Cables. There are two kinds of fiber-optic cables in common use. The first, called multimode fiber, is less expensive to make and use, but has a more limited bandwidth. The second, called single-mode fiber, is more expensive to make and use, but can support very high bandwidths.

Multimode fiber is used for distances of up to 2 km. It typically has a core diameter of 62.5 μm and a cladding diameter of 125 μm. Because of the large diameter of the core, there are many paths, or modes, between the two ends of the fiber. As demonstrated in Figure 18.2.9, a given photon may travel down the center of the core (mode 1), may reflect a few times at a shallow angle to the interface (mode 2), or may strike the interface at a sharp angle (mode 3) and reflect many times as it travels to the far end of the fiber.

These many paths result in a high attenuation of the lightbeam as it travels down the fiber. This attenuation is caused by the large number of photons lost through the interface (for a given probability of reflection, the more reflections a photon must make, the higher the chance that the photon will pass through the interface rather than being reflected by it), as well as other attenuation mechanisms such as phase distortion.

Single-mode fiber is designed for use in higher bandwidth and longer distance applications. It is made up of a much smaller core (typically 10 μm) in a somewhat smaller cladding (typically 100 μm) than is used in multimode fiber, resulting in a greatly reduced number of modes (Figure 18.2.10). As a result, single-mode fiber has a much lower attenuation than an equivalent multimode fiber.

Single-mode fiber is much more difficult to splice and terminate than is multimode fiber, as there is very little margin for error in aligning the core. A 2-μm alignment offset in a multimode fiber is only about 3% of the total diameter, whereas it is 20% of the diameter of a single-mode fiber. Thus, a minor misalignment in single-mode fiber splice will result in a major increase in attenuation.

Typical Fiber-Optic-Based Networks. One ubiquitous use of fiber in corporate networks is for fiber distributed data interface (FDDI) backbone networks. FDDI networks transfer data at 100 Mbps using a token-ring architecture. Rings of up to 200 km are supported, with a maximum distance between adjacent nodes of 2 km.

Wide Area Networks

Once out of the local network area, the characteristics of networks change. In general, connections between widely distributed locations are not handled by the installation of cable by the company itself. The company will generally turn to a telecom provider, such as the local telephone company, to provide the necessary connections.

The telephone company does not generally sell cable connections. They sell bandwidth. Thus, the standards for wide area networks are based on fixed bandwidth allocations over telephone company facilities. The telephone company may user copper or fiber, at their discretion, to provide the requested service. The telephone company's networks are usually implemented as rings or stars, as are local networks.

There are three broad classes of digital service available. The first, traditionally copper based, is the DS-n service. The second, also copper based, is basic rate Integrated Services Digital Network (ISDN). The third, optical fiber based, is the OC-n service.

DS-n. The DS series of services provide low- to medium-speed data connections. Table 18.2.2 summarizes the commonly available services and their customary uses.

TABLE 18.2.2 DS-n Class Services

Service Name	Data Rate	Common Use
DS-0	56/64 Kbps	Voice, low speed data
DS-1	1.544 Mbps	Multiple voice lines, data
DS-2	6.2 Mbps	Data
DS-3	44.736 Mbps	Data

Pricing for these services is heavily dependent on the total mileage of the link. The links are usually leased for a minimum period (typically 1, 2, or 5 years). These leased data lines can support all network protocols.

Leased lines are point-to-point. That is, they run from one location to another. A single line cannot be configured as a ring, a star, or a bus. To implement these architectures, several leased lines must be ordered, and connected together by the customer into the desired configuration.

Basic Rate ISDN. An increasingly important wide area networking technology is the Integrated Services Digital Network (ISDN). The most commonly implemented version of this service is the basic rate service, which provides three data channels over a single pair of telephone cables. Two channels, known as the bearer (or "B") channels, each provide a 64-Kbps dialup digital connection. The third channel, known as the data (or "D") channel, is used for control information between the telephone switch and the ISDN terminal.

In practice, an ISDN line is installed at a business or home location. The line is connected to a piece of equipment known as an ISDN terminal adapter. This device allows a variety of other devices to make use of the various channels provided by the ISDN line. One typical device has a connection for a voice telephone and an ethernet, and can be used to place normal telephone calls simultaneously with a digital dialup connection to a company's ethernet.

A higher-speed version of ISDN, known as primary rate ISDN, supports 24 B channels over a DS-1 leased line. One of these channels is typically used for control information, leaving 23 channels available for data transmission.

OC-n. As demand for higher-speed connections continued to increase, standards for optical networks were established to meet these demands. The OC-n (Optical Carrier) standards (Table 18.2.3) provide data rates starting at 55 Mbps and going up to multiple gigabits per second. OC-based networks use the SONET (Synchronous Optical Network) signaling standards for transferring data.

TABLE 18.2.3 OC-n Data Rates

OC-n	Data Rate (Mbps)
OC-1	51.84
OC-3	155.52
OC-12	622.08
OC-48	2488.32
OC-192	9953.28

OC/SONET networks are always implemented as rings. They can be configured to be fault-tolerant, so that no single break in the cable (or failed network node) will cause connections to fail.

Logical connections are made across a SONET ring by the use of *virtual circuits*. These virtual circuits can either be permanent (PVC) or switched (SVC). A PVC is configured manually into the control electronics of the originating and terminating SONET control equipment. It will persist indefinitely until it is manually terminated. An SVC is established dynamically upon a request by a computer connected to the control equipment. The SVC persists until the requesting computer informs the control equipment that the circuit is no longer required. Conceptually, PVCs are like leased lines (permanent), and SVCs are like dial-up telephone lines (temporary).

Asynchronous transfer mode (ATM) is an increasingly important protocol that is commonly implemented over SONET networks. ATM makes use of fixed-sized packets, called cells, to transfer data at very high speeds. Each of these cells is 53 bytes in size, with 5 bytes for addressing and cell management and 48 bytes for data. Because the cells are fixed in size and format, it is possible to build network switches that are very fast, as the processing and routing of the cell can be entirely done in hardware.

To set up a virtual circuit between nodes, the originating node sends a request to its local ATM switch. In this request, the originating node specifies the bandwidth required for the link, the length of time the link is needed, and the quality of service required. The originating node will check its available network resources to see if it can grant the service requested. If it can, it will pass on the request to the next switch in the network which will repeat the check for available resources. This process continues until the destination node is reached and consents to the link request. An acknowledgment is relayed back to the originating node confirming the availability of the requested service. As this acknowledgment is returned through each switch, the switch commits the required resources to the virtual circuit.

A key strength of ATM is that it can carry any type of information over the same link, rather than requiring separate links for voice, data, and video as is common today. This ability will likely reduce the overall cost of networking as separate voice, data, and video networks will not be required in the future.

Defining Terms

100 The standard for running 10 Mb/sec ethernet over unshielded twisted pair network cabling.
100BaseT: The standard for running 100 Mb/sec ethernet over unshielded twisted pair network cabling.
Active hub: A device that interconnects multiple network links.
Coax: Coaxial cable.
Local area network: A network that is within the local area, generally within an office area or building.
STP: Shielded twisted pair cable.
Token: A data packet that circulates around a token-ring network indicating that the network is available for data transmission.
UTP: Unshielded twisted pair cable.
Wide area network: A network that is outside the local area, generally between buildings, between cities, or between countries.

References

Acampora, A.S. 1994. *An Introduction to Broadband Networks*. Plenum Press, New York.
Bates, R.J. 1992. *Introduction to T1/T3 Networking*. Artech House.
Clark, M.P. 1991. *Networks and Telecommunications: Design and Operation*. John Wiley & Sons, New York.
Conard, J.W. (Ed.) 1991. *Handbook of Communications Systems Management*. Auerbach, Boca Raton, FL.
Davidson, R.P. 1994. *Broadband Networking ABCs for Managers*. John Wiley & Sons, New York.
Kosiur, D.R. 1995. *How Local Area Networks Work*. Prentice-Hall, Englewood Cliffs, NJ.
McElroy, M.W. 1993. *The Corporate Cabling Guide*. Artech House.
McNamara, J.E. 1988. *Technical Aspects of Data Communication*. Digital Press.
Minoli, D. 1993. *Enterprise Networking*. Artech House.

Wireless Networks

Introduction

With increasing mobility of the workforce comes the need to provide network access to computers located in automobiles, in briefcases, on shipboard, and even in aircraft. Wireless networks provide the required connectivity.

This section covers the basic radio frequency (RF) link types used in wireless networking, compares wireless networks to traditional wired networks, and discusses key trends in the field. The reference section provides pointers to sources of additional information.

RF Technologies

All mobile wireless networks make use of a radio frequency carrier*. This carrier may be fixed frequency (like an AM radio station signal) or spread spectrum (where the signal is spread over a wide frequency band). Section 18.4 provides additional information on radio frequency communications, including modulation methods.

Point to Point. A point-to-point communications system relays the signal from the mobile user to a base station (Figure 18.2.11). The radio link is a single hop, that is, the signal is transferred directly from the mobile user to the base station without going through intermediate reception and retransmission steps. A point-to-point system generally requires a line of sight between the transmitter and the receiver. Because the curvature of the earth limits the line of sight, this type of system has a typical maximum range of tens of miles.

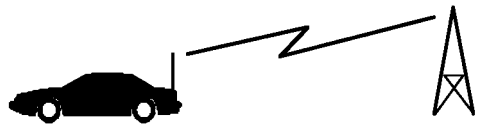


FIGURE 18.2.11 Point-to-point link.

Several point-to-point links can be put in series to relay a signal. Microwave relay stations are typically placed approximately every 30 mi to relay a signal from point to point over long distances. Each relay point receives the signal and retransmits it to the next station in line.

* Some local area wireless networks use infrared light or laser carriers. These are not used in mobile applications because of their limited range.

Cellular. To overcome the limitations of point-to-point systems for mobile telephony, a cellular radio system can be used. This system requires a large number of antennas, each at the center of a "cell", as shown in Figure 18.2.12. As a ground-based vehicle approaches the edge of the coverage range for a particular antenna, the cellular control system communicates with the mobile data unit or telephone and assigns it to the next antenna in the direction in which the vehicle is heading. At a coordinated time, the in-progress conversation is transparently (to the user) switched to the newly assigned antenna without interruption.

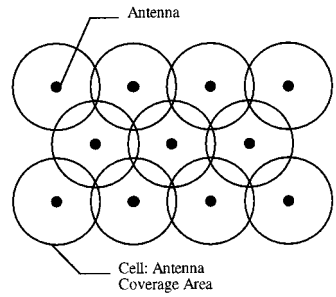


FIGURE 18.2.12 Coverage pattern for multiple cellular antennas.

In congested downtown areas, cell antennas may be only a few hundred meters apart and use lower-powered transmitters to reduce the size of the cell. This provides more channels for use by more simultaneous conversations, but requires more handoffs as the mobile unit moves through the city. In suburban and rural areas, the cell antennas may be several thousand meters apart, with higher-powered transmitters to provide a large cell. This minimizes the number of cell antennas that must be installed, reducing the cost of the system.

A personal communications system (PCS) is a version of cellular technology that uses smaller "microcells" with lower power transmitters. This second-generation cellular system requires many more cell antennas (every few hundred meters), making it most effective in urban areas. PCS telephones are typically smaller and lighter than first-generation cellular telephones, because of their lower power and higher frequency operation.

Satellite — Low Earth Orbit. Cellular and PCS systems require a very large number of antennas installed in a grid over a large area to provide coverage for mobile units. This limits the rate at which a system with complete coverage for a given area can be installed. Each cell antenna installation requires permission from the cell-site property owner, building permits, power, telecommunications connections, and regulatory approval before it can be built and made operational. In addition, cellular systems differ from country to country, requiring different mobile units for access in each country.

One approach to addressing these problems is to stand the problem on its head. Rather than having a large set of fixed antennas which hand off the mobile unit as the mobile unit moves, use a number of satellite transceivers in low earth orbit (LEO). Establish the orbits of these transceivers so that there is at least one satellite in view of any place on the face of the earth at a given time. Then, as a given satellite begins to move out of range of the mobile unit, it hands off the call to another satellite that is just moving into range.

This type of system can easily provide data and voice connectivity to any location on (or slightly above) the face of the Earth using a common mobile transceiver. For more information, see the subsection Satellite Applications.

Comparing Wireless to Wired Networks

It is clearly posible to provide data and voice access to mobile users with any of the above technologies. Yet there are significant limitations to these systems when compared to a hardwired network connection. As always, there are trade-offs that must be made.

Bandwidth. A directly wired network connection can easily provide 100 Mbps. The systems discussed in the previous section are typically limited to tens of Kbps, four to five orders of magnitude below what is commonly available to directly wired users.

This bandwidth limitation has profound implications for mobile users. For example, complex graphics and images require several minutes to download, rather than the several seconds that are required for a directly wired user. Any function that requires the transfer of large amounts of data will be greatly slowed by a wireless network link.

The limitation is not likely to ever go away. There is a limited amount of RF spectrum available for mobile computing, while there is an essentially unlimited amount of network spectrum available for directly wired users (just add more fibers or wires!). While improvements in wireless bandwidth availability will come in time, it simply will never compare with the bandwidth available to a desktop computer user.

Security. Wireless transmissions can be easily intercepted. The signal is broadcast through the atmosphere and can be received by anyone within range who has a properly tuned receiver. Thus, it is necessary to use encryption to protect the transmitted data.

Encryption adds complexity and cost to any system. Securely exchanging a cryptographic key for a call can be difficult to do in a way that is difficult to compromise. Every system that needs to interoperate must agree on what cryptographic algorithm will be used and how it will be keyed.

Costs. As can be inferred, the costs of establishing a mobile wireless network are very high. These costs must be recovered from the users of the system within its projected useful life. As an example, compare the cost of a typical local wired telephone call with the cost of a cellular telephone call. In many areas, the local wired call is offered at a fixed rate for an unlimited call length, while the cellular call is charged by the minute.

Conclusions

Wireless networking has clear usefulness where it is necessary to have access to data while mobile. Applications of wireless networking in such areas as police work or emergency services have already proven their value.

The costs and limitations of wireless networking must be carefully considered before embarking on a major initiative. With the technology in this area developing rapidly, in a wide variety of incompatible directions, caution must be exercised.

References

Bates, R.J. 1994. *Wireless Networked Communication.* McGraw-Hill, New York.
Breed, G. 1994. *Wireless Communications Handbook.* Cardiff.
Calhoun, G. 1992. *Wireless Access and the Local Telephone Network*. Artech House.
Davis, P.T. and McGuffin, C.R. 1995. *Wireless Local Area Networks*. McGraw-Hill, New York
Lee, W.C.Y. 1995. *Mobile Cellular Telecommunications*. McGraw-Hill, New York
Nemzow, M.A.W. 1995. *Implementing Wireless Networks*. McGraw-Hill, New York

Satellite Communications

Daniel F. DiFonzo

Introduction

The impact of satellites on world communications since commercial operations began in the mid-1960s is such that we now take for granted many services that were not available a few decades ago: worldwide TV, reliable communications with ships and aircraft, wide area data networks, communications to remote areas, direct TV broadcast to homes, position determination, and Earth observation (weather and map-

ping). Future satellite-based global personal communications to hand-held portable telephones may usher in yet another new era.

Satellites function as line-of-sight microwave relays in orbits high above the Earth which can "see" large areas of the Earth's surface. This unique feature ensures the continued growth of satellites even as fiber-optic cables capture a larger market share of high-density point-to-point traffic. Satellites provide cost-effective access for areas with low (thin-route) communications traffic, because Earth terminals can be installed in locations where the high investment cost of terrestrial facilities might not be warranted. Satellites are particularly well suited to wide area coverage for broadcasting, mobile communications, and point-to-multipoint communications.

Satellite Applications

Figure 18.2.13 depicts several kinds of satellite links and orbits. The geostationary Earth orbit (GEO) is in the equatorial plane at an altitude of 36,000 km with a period of one sidereal day (23h 56m 4.09s). GEO satellites appear to be almost stationary from the ground (subject to small perturbations) and the Earth antennas pointing to these satellites may need only limited or no tracking capability. The orbits for which the highest altitude (apogee) is at or greater than GEO are sometimes referred to as high Earth orbits (HEO). Low Earth orbits (LEO) typically range from a few hundred kilometers to about 1000 km, and medium Earth orbits (MEO) are at intermediate altitudes.

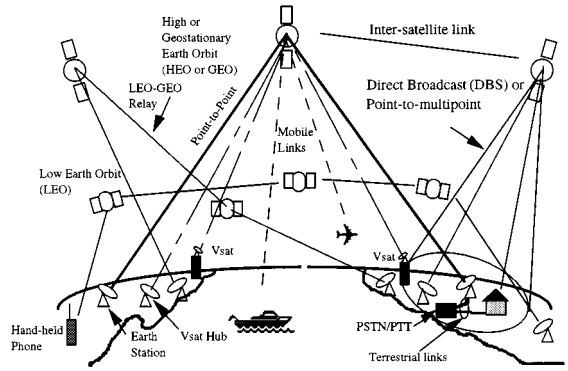


FIGURE 18.2.13 Satellite links and orbits.

Initially, satellites were used primarily for point-to-point traffic in the GEO fixed satellite service (FSS), e.g., for telephony across the oceans and for point-to-multipoint TV distribution to cable *head end* stations. Large Earth station antennas with high-gain narrow beams and high uplink powers were needed to compensate for limited satellite power. This type of system, exemplified by the early global network of the International Telecommunications Satellite Consortium (Intelsat), used "Standard A" Earth antennas with 30-m diameters. Since the start of Intelsat, many other satellite organizations and consortia have been formed around the world to provide international, regional, and domestic services (Rees, 1990).

As satellites have grown in power and sophistication, the average size of the Earth terminals has been reduced. High-gain satellite antennas and relatively high-power satellite transmitters have led to *very small aperture* Earth *terminals* (VSAT) with diameters of less than 2 m and modest powers of less than 10 W (Gagliardi, 1991). As depicted in Figure 18.2.13, VSAT terminals may be placed atop urban office buildings, permitting private networks of hundreds or thousands of terminals which bypass terrestrial lines. VSATs are usually incorporated into *star* networks where the small terminals communicate through the satellite with a larger *Hub* terminal. The Hub retransmits through the satellite to another small terminal. Therefore, VSAT-to-VSAT links require two *hops* with attendant time delays. With high-gain satellite antennas and relatively narrowband digital signals (e.g., compressed voice at ≤8 kbps), direct single-hop *mesh* interconnections of VSATs may be used.

Satellite Functions

The traditional function of a satellite is that of a "bent pipe" quasilinear repeater in space. As shown in Figure 18.2.13, *uplink* signals from Earth terminals directed at the satellite are received by the satellite's antennas, amplified, translated to a different *downlink* frequency band, channelized into *transponder channels*, further amplified to relatively high power, and retransmitted toward the Earth. Transponder channels are generally rather broad (e.g., bandwidth of 36 MHz) and each may contain many individual or user channels.

Multiple access techniques, to be discussed later, allow many users to share a satellite's resources of bandwidth and power and to avoid interfering with each other and with other satellite or terrestrial systems. Multiple access systems segregate users by frequency, space, time, polarization, and signaling code orthogonality.

Analog and digital modulations are both in widespread use. While frequency modulation (FM) has been prevalent, recent advances in digital voice and video compression will lead to the widespread use of digital modulation methods such as quartenary phase shift keying (QPSK) and quartenary amplitude modulation (QAM). Figure 18.2.14 depicts the functional diagram appropriate to a satellite using frequency division multiple access (FDMA) and reusing available frequencies by means of multiple antenna beams. Interference can result if the sidelobes of one beam receive or transmit substantial energy in the direction of the other beam.

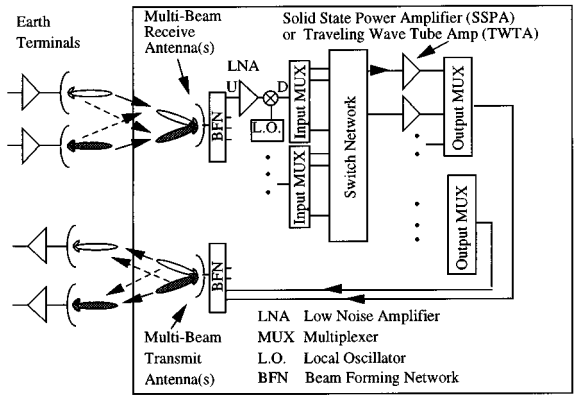


FIGURE 18.2.14 Satellite system block diagram.

Newer satellite architectures, such the NASA Advanced Communications Technology (ACTS) and Motorola's *Iridium* system, may use *regenerative repeaters*, which process the uplink signals by demodulating them to baseband. These baseband signals, which may be for individual users or may represent *frequency division multiplexed* (FDM) or *time division multiplexed* (TDM) signals from many users, are routed to downlink channels, modulated onto one or more radio frequency (RF) carriers, and transmitted to Earth.

High-power *direct broadcast satellites* (DBS) operating at Ku-band (around 12 GHz) deliver TV directly to home receivers having antennas less than 1 m in size. Such systems using analog FM are operational in Japan and Europe. In the United States, DBS with digital modulation and compressed video will provide more than four NTSC TV channels per 24-MHz transponder channel. For the United States, where each DBS orbital location is allocated 32 transponder channels of 24 MHz each, more than 128 conventional TV channels can be provided from a single DBS orbital location. DBS is seen as an attractive medium for delivery of high-definition TV (HDTV) to a large number of homes.

Mobile satellite services (MSS) operating at L-band around 1.6 GHz have revolutionized communications with ships and, more recently, with aircraft which would normally be out of reliable communi-

cations range of terrestrial radio signals. The International Maritime Satellite Consortium (Inmarsat) operates the dominant system of this type.

Links between LEO satellites (or the NASA shuttle) and GEO satellites are used for data relay, e.g., via the NASA Tracking and Data Relay Satellite System (TDRSS). Some systems will use intersatellite links (ISL) to improve the interconnectivity of a wide-area network. ISL systems would typically operate at frequencies above 20 GHz or even use optical links.

An exciting new development is the prospective use of L-band frequencies with a large number (12 to 66) of LEO satellites for personal communications systems (PCS) directly with small hand-held portable telephones anywhere in the world.

Access and Modulation

Satellites act as central relay nodes which are visible to a large number of users who must efficiently use the limited power and bandwidth resources. A brief summary of issues specific to satellite systems is given below.

Frequency division multiple access (FDMA) has been the most prevalent access for satellite systems until recently. Individual users assigned a particular frequency band may communicate at any time. Satellite filters subdivide a broad frequency band into a number of *transponder channels*, e.g., the 500-MHz uplink FSS band from 5.925 to 6.425 GHz may be divided into 12 transponder channels of 36 MHz bandwidth plus guard bands. This limits the interference among adjacent channels in the corresponding downlink band of 3.7 to 4.2 GHz.

FDMA implies that several individual carriers coexist in the transmit amplifiers. In order to operate the amplifiers in a quasilinear region relative to their saturated output power to limit intermodulation products, the amplifiers must be operated in a backed-off condition. For example, in order to limit third-order intermodulation power for two carriers in a conventional TWT (traveling wave tube) amplifier to ≈–20 dBc, its input power must be reduced (*input* ***backoff***) by about 10 dB relative to the power that would drive it to saturation. The output power of the carriers is reduced by about 4 to 5 dB (*output backoff*). Amplifiers with fixed bias levels will consume power even if no carrier is present. Therefore, dc-to-RF efficiency degrades as the operating point is backed off. For amplifiers with many carriers, the intermodulation products have a noise-like spectrum and the *noise power ratio* is a better measure of multicarrier performance.

Time division multiple access (TDMA) users share a common frequency band and are each assigned a unique time slot for their digital transmissions. At any instant there is only one carrier in the transmit amplifier, requiring little or no backoff from saturation. The dc-RF efficiency is high. A drawback is the system complexity required to synchronize widely dispersed users in order to avoid intersymbol interference caused by more than one user signal appearing in a given time slot. Also, the total transmission rate in a TDMA satellite channel must be essentially the sum of the users' rates, including overhead bits such as for framing, synchronization and clock recovery, and source coding. At the present state of the art, Earth terminal hardware costs may be higher than for FDMA. Nevertheless, TDMA systems are gaining acceptance for some applications.

Code division multiple access (CDMA) modulates each carrier with a unique pseudorandom code, usually by means of either a direct sequence or frequency hopping spread spectrum modulation. As the CDMA users occupy the same frequency band at the same time, the aggregate signal in the satellite amplifier is noise-like. Individual signals are extracted at the receiver by correlation processes. CDMA tolerates noise-like interference but does not tolerate large deviations from average loading conditions. One or more very strong carriers could violate the noise-like interference condition and generate strong intermodulation signals.

User access is via assignments of a frequency, time slot, or code. Fixed assigned channels allow a user unlimited access. However, this may result in poor utilization efficiency for the satellite resources and may imply higher user costs (analogous to a leased terrestrial line). Other assignment schemes include *demand assigned multiple access* (DAMA) and *random access* (e.g., for the Aloha concept). DAMA systems require the user to first send a channel request over a common control channel. The

network controller (at another earth station) seeks an empty channel and instructs the sending and receiving units to tune to it (either in frequency or time slot). A link is maintained for the call duration and then released to the system for other users to request. Random access is economical for lightly used burst traffic such as data. It relies on random time of arrival of data packets and protocols are in place for repeat requests in the event of collisions (Gagliardi, 1991).

In practice, combinations of multiplexing and access techniques may be used. A broad band may be channelized or *frequency division multiplexed* (FDM) and FDMA may be used in each subband, e.g., FDM/FDMA.

The traditional satellite modulation format has been FM. However, recent trends indicate that digital modulations such as M-ary PSK and QAM will become more prevalent for nearly all applications including voice, data, and TV. The efficiencies afforded by digital modulations arise partly because they allow signal processing for bandwidth compression. Compressed digital TV transmission allows a significant improvement in capacity compared with FM.

Trends

Satellite communications have approached a mature stage of development, and their competitiveness for point-to-point voice traffic, compared with fiber, has been questioned. However, as mentioned, satellites will continue to exploit their unique wide view of the Earth for such applications as broadcast and personal communications. The satellite industry's maturity also presents another challenge. To date, satellite construction has resembled a craft industry with extensive custom design, long lead times, long test programs, and high cost. Satellites will benefit from modern "lean production" and "design-to-cost" concepts that could lead to systems having lower cost per unit of capacity and higher reliability. Technology advances that are being pursued include development of lightweight "lightsats" for economical provision of services at low cost, more sophisticated on-board processing to improve interconnectivity, intersatellite links, improved components such as batteries, and even such speculative concepts as providing satellite power from the ground via high-power laser beams using adaptive optics (Landis and Westerlund, 1992).

Defining Terms

Attitude: The angular orientation of a satellite in its orbit, characterized by roll (R), pitch (P), and yaw (Y). The roll axis points in the direction of flight, the yaw axis points toward the earth's center, and pitch axis is perpendicular to the orbit plane such that $R \times \mathrm{P} \rightarrow Y$b

Backoff: Amplifiers are not linear devices when operated near saturation. To reduce intermodulation products for multiple carriers, the drive signal is reduced or backed off. Input backoff is the dB difference between the input power required for saturation and that employed. Output backoff refers to the reduction in output power relative to saturation.

Bus: The satellite bus is the ensemble of all the subsystems that support the antennas and payload electronics. It includes subsystems for electrical power, attitude control, thermal control, TT&C, and structures.

Frequency reuse: A way to increase the effective bandwidth of a satellite system when available spectrum is limited. Dual polarizations and multiple beams pointing to different Earth regions may utilize the same frequencies as long as, for example, the gain of one beam or polarization in the directions of the other beams or polarization (and vice versa) is low enough. Isolations of 27 to 35 dB are typical for reuse systems.

Polarization isolation: Frequency reuse allocates the same bands to several independent satellite transponder channels. The only way these signals can be kept separate is to isolate the antenna response for one reuse channel in the direction or polarization of another. The beam isolation is the coupling factor for each interfering path (ideally 0 or $-\infty$ dB).

References

Gagliardi, R.M. 1991. *Satellite Communications.* Van Nostrand Reinhold, New York.

Griffin, M.D. and French, J.R. 1991. *Space Vehicle Design.* American Institute of Aeronautics and Astronautics, Washington, D.C.

Long, M. 1990. *The 1990 World Satellite Annual.* MLE, Winter Beach, FL.

Morgan, W.L. and Gordon, G.D. 1989. *Communications Satellite Handbook.* John Wiley & Sons, New York.

Rees, D. 1990. *Satellite Communications: The First Quarter Century of Service.* John Wiley & Sons, New York.

Richaria, M. 1995. *Satellite Communications Systems: Design Principles*, McGraw-Hill, New York.

Roddy, D. 1995. *Satellite Communications.* Prentice-Hall, Englewood Cliffs, NJ.

Wertz, J.R. and Larson, W.J., eds. 1991. *Space Mission Analysis and Design.* Kluwer Academic Publishers, Dordrecht, The Netherlands.

Further Information

For a brief history of satellite communications see *Satellite Communications: The First Quarter Century of Service*, by D. Rees (Wiley, 1990). Propagation issues are summarized in *Propagation Effects Handbook for Satellite Systems Design*, 1983, NASA Reference Publication 1082(03), November 1990; descriptions of the proposed LEO personal communications systems are in the FCC filings for *Iridium* (Motorola), *Globalstar* (SS/Loral), *Odyssey* (TRW), *Ellipso* (Ellipsat), and *Aries* (Constellation Communications), 1991 and 1992.

Computer Communications

Lloyd W. Taylor

The previous three sections have discussed various communications links used in digital interconnections. This section discusses common **protocols** used to handle network addressing and delivery over these communication links.

The OSI Network Model

The International Standards Organization (**ISO**) developed a model of network communications called the Open Standards Interconnect (**OSI model**). This is an abstract model that does not necessarily imply any particular implementation. In fact, it is most commonly used as a reference model into which various network protocols are mapped.

The OSI model (Table 18.2.4) defines seven layers. The topmost layer defines the interface between the user-interface software and the network, while the bottommost layer defines the electrical characteristics of the network itself. Not all protocols implement all seven layers, as will be seen later in this section.

TABLE 18.2.4 The OSI Network Model

Layer Name	Function
Application	Interface between user applications and network
Presentation	Interface between network applications and network
Session	Destination definition
Transport	Internet addressing
Network	Switching, routing, management
Data link	Network access control, local addressing
Physical	Electrical or optical cable, signaling, and transmission standards

The seven layers have the following functions:

The *application layer* is the layer that the user or the user's application interfaces with. This layer provides such capabilities as file transfer, electronic mail transfer, and videoteleconferencing transfer.

The *presentation layer* defines the format of the information to be transferred. It handles the mapping between the network representation of data and the local representation of data. This layer is typically where network encryption and compression will occur.

The *session layer* is responsible for setting up connections between processes on separate computers. It connects the processes, manages the connection, and terminates the connection when the data transfer is complete.

The *transport layer* guarantees error-free communication for the connection. It handles error checking and retransmission of lost packets, and manages the traffic flow between the end points of the connection.

The *network layer* manages the creation, maintenance, and termination of connections between computers. It handles the routing of the data across the intervening networks. It interfaces between the logical abstractions of the higher layers with the technical specifics of the lower.

The *data link layer* handles data at the packet level. It enables the receiving computer to check that these blocks of data have been received reliably.

The *physical layer* implements the standards for sending bits across a particular type of network. It includes standards for the wire or fiber itself, standards for the signal characteristics, and standards for the connection of the network to the computer.

These layers work together to ensure reliable communications between processes on different computer systems. Note that each computer must implement the same set of protocols to be able to interoperate. Even if both systems use protocols which are "OSI Standard," there is no guarantee that they will talk to each other.

Common Network Protocols

A typical corporate network has many different protocols in use at any one time. This is a legacy of the days when every vendor implemented his own network standards, which resulted in serious interoperability problems.

In the recent past, there has been a significant move toward the use of **TCP/IP** as a standard protocol for all network communications. Most vendors today offer TCP/IP as an option, and many have adopted it as a standard. If all systems on a corporate network "speak the same language," it is much easier to get them to talk to each other. It is also much simpler to configure and operate the computers on that network, as they need speak only one language, rather than several, to interconnect with all other computers.

The three protocols selected for discussion in this section are among the most common. They are provided here as a set of examples of how various protocols are implemented, to provide an introduction to the complexities of internetworking.

TCP/IP. The Transmission Control Protocol/Internet Protocol (TCP/IP) was developed by two engineers in 1974. They built on the work of a Ph.D. student at Harvard, who had developed the idea for ethernet as his doctoral thesis. By 1982, these ideas had been further developed by others and were published as standards by the U.S. Department of Defense.

TCP/IP is a very simple set of protocols. It implements only those pieces that are necessary for reliable and efficient network communications and leaves the more complex tasks (such as network routing) to other applications. As a result, TCP/IP is relatively simple to implement and therefore has been implemented on every computer operating system. It has become the *lingua franca* of networking.

The mapping of TCP/IP into the OSI model is shown in Table 18.2.5. Note that TCP/IP does not implement all layers of the OSI model explicitly, but combines several layers into a single implementation piece.

TABLE 18.2.5 Mapping of TCP/IP into the OSI Model

OSI Layer Name	TCP/IP Implementation
Application	File transfer protocol (**FTP**)
Presentation	Simple mail transport protocol (SMTP)
Session	Virtual terminal protocol (Telnet)
Transport	Transmission control protocol (TCP)
	User datagram protocol (UDP)
	Internet control message protocol (ICMP)
Network	Internet protocol (IP)
	Address resolution protocol (ARP)
Data link	Ethernet (coax, twisted pair), token ring,
Physical	FDDI, ATM, X.25, SONET, LocalTalk, ...

The protocols listed perform the following functions:

File transfer protocol implements the messaging necessary to transfer files from one computer to another. It handles such things as authentication, directory listings, file selection, and datastream format. Most systems implement an FTP command that allows the user to interact with the FTP system through the use of either text commands or a graphical user interface.

Simple mail transport protocol implements a standard for the transfer of electronic mail messages between computer systems. It handles such things as system identification, addressing, the separation of the body of the message from the headers of the message, and the termination of the transfer. A typical computer will implement a mail user interface that completely isolates the user from this protocol.

Virtual terminal protocol implements a standard for the negotiation and connection of a terminal emulator to a remote computer. **Telnet** negotiates such things as data representation standards (ASCII, EBCDIC, Latin1), datastream format (8-bit or 7-bit), and control characters to be used for interrupt and flow control. Typical computers implement a user interface that provides interpretation of terminal control codes sent by the remote computer (e.g., “clear screen”), as well as controls for establishing and terminating a terminal emulation session.

Transmission control protocol underlies the above protocols and provides a reliable, full-duplex, datastream service between two computers. It establishes a virtual circuit (or connection) between the two systems, then sends and receives data across that connection until the communicating processes request that the circuit be disestablished. TCP depends on IP for sending and receiving the packets that are transferred across the virtual circuit.

User datagram protocol is used by processes that do not require guaranteed delivery of data. No virtual circuit is established. The sending process simply addresses the packet to a particular process on the remote computer and sends it across the network. It is up to the processes at either end to ensure that the packet is received properly, as the UDP protocol makes no guarantees.

Internet control message protocol is used to send control messages concerning network functions to TCP/IP clients. Instructions such as “reduce your transmission rate” and “your requested destination is unreachable” are transferred at this level.

Internet protocol defines the standard datagram that transports information across the network. It does not guarantee delivery, but expects higher-level protocols (such as TCP) to handle that function. It handles packet-level flow control and error checking.

Address resolution protocol handles the mapping of a logical IP address to a physical network address. It enables the local computer to determine where the packet must be sent for the next step of its journey across the network.

IPX. The Internet packet exchange (IPX) protocol is commonly used in Novell NetWare filesharing systems. It is based on the XNS suite of protocols designed by Xerox. IPX is a *lightweight* protocol. It is designed to be very efficient (easy to process) and to provide maximum performance on a local network. The downside of lightweight protocols is that they tend not to work as well in large internets. Because they are less complex, they lack the necessary components to work well in complex network environments.

The differences between IPX and TCP/IP can be seen by comparing Tables 18.2.6 and 18.2.5. Notice the significantly fewer number of layers in the IPX set of protocols. This simplification is the source of the efficiency of IPX.

TABLE 18.2.6 Mapping of IPX into the OSI Model

OSI Layer Name	IPX Implementation
Application Presentation Session	Netware core protocol (NCP)
Transport	Sequenced packet exchange (**SPX**)
Network	Internet packet exchange (IPX)
Data link Physical	Ethernet (coax, twisted pair), token ring, FDDI, ATM, X.25, SONET, LocalTalk, …

The protocols listed perform the following functions:

Netware core protocol implements directory services, connection services, security, and similar functions. It is essentially the operating system for NetWare servers and, as such, implements far more functionality than the TCP/IP protocol, where the top of the OSI stack is underneath the operating system. Core services can be added to NetWare servers by adding netware loadable modules (NLM), which can be used to add things such as data base engines and gateways to TCP/IP networks.

Sequenced packet exchange builds on the services offered by IPX. SPX adds packet acknowledgment to the functions in IPX, so that the sending computer can know for certain that the packet was received at the far end. Note that the use of SPX is not required in NetWare networks.

Internetwork packet exchange implements an efficient packet transfer service. The packet contains only address information, data, and a error-checking code, resulting in a packet with little overhead.

It should be noted that current NetWare networks can be implemented using TCP/IP protocols. This is appropriate where the network is large (or expected to grow), or where there is already a requirement for TCP/IP, such as in networks connected to the Internet.

AppleTalk. **AppleTalk** was designed by Apple Computer to provide a simple-to-install, simple-to-manage network. It has an extensive set of protocols to provide such services as automatic address selection (for workstations and servers), printer sharing, routing, and security. Anyone who has installed an AppleTalk network has benefited from its simple “plug and play” characteristics.

This simplicity for the user comes at the cost of a significant complexity in the protocol. Since AppleTalk must discover on its own what most other protocols require the network administrator to manually configure, it follows that there must be many additional functions, resulting in a necessarily more complex protocol suite.

The major protocols used in AppleTalk are listed in Table 18.2.7. Notice the large number of protocols at the session and transport layers. These additional protocols are used to self-configure the AppleTalk network.

TABLE 18.2.7 Mapping of AppleTalk into the OSI Model

OSI Layer Name	AppleTalk Implementation
Application	AppleShare
Presentation	AppleTalk filing protocol (AFP), PostScript
Session	AppleTalk session protocol (ASP), AppleTalk data stream protocol (ADSP), zone information protocol (ZIP), printer access protocol (PAP)
Transport	AppleTalk transaction protocol (ATP), name binding protocol (NBP), AppleTalk echo protocol (AEP), routing table maintenance protocol (RTMP)
Network	Datagram delivery protocol (DDP)
Data link Physical	Ethernet (coax, twisted pair), token ring, FDDI, ATM, X.25, SONET, LocalTalk, …

The protocols listed perform the following functions:

AppleShare provides file sharing services.

AppleTalk filing protocol handles the communications between the user's computer and the AppleShare fileserver. It allows users to share files and applications, handle security, and ensure that each user has a current view of the shared file service.

PostScript defines the format of documents printed from Macintosh applications. It is a page description language, developed by Adobe.

AppleTalk session protocol manages the creation, operation, and destruction of sessions between computers. It organizes the sequencing of function requests (who gets access to what service in what order).

AppleTalk data stream protocol is a connection-oriented protocol (like TCP) that reliably transfers a stream of bytes between two computers.

Zone information protocol handles the discovery and mapping of the entire AppleTalk network. It finds all connected networks and builds a table that is used by AppleTalk to find other computers, and to route packets appropriately.

Printer access protocol sets up and manages connections between the user's computer and printers or print servers.

AppleTalk transaction protocol is used to pass requests and responses reliably between computers. It ensures that the receiving computer accurately gets the request and also ensures that the answer returns reliably to the requester.

Name binding protocol translates between the name of a computer and its address.

AppleTalk echo protocol responds to a request by a remote computer by simply acknowledging that an echo request was received. This function is used to be sure that a remote computer is active.

Routing table maintenance protocol is used by AppleTalk routers to discover and keep track of the configuration of the AppleTalk internet. It is used to maintain routing tables or internal network maps.

Datagram delivery protocol is the core of the AppleTalk protocol suite. It moves packets of data from one computer to another, but does not guarantee reliable delivery.

Defining Terms

AppleTalk: A set of protocols developed by Apple computer to interconnect Macintosh computers and peripherals.
FTP: File transfer protocol.
IP: Internet protocol.
IPX: Internetwork packet exchange.
ISO: International Standards Organization.
Layer: One part of a protocol stack.
OSI model: The model of network interfaces as a seven-layer entity.
Protocol: A standard way of doing things.
SPX: Sequenced packet exchange.
TCP: Transmission control protocol.
Telnet: Virtual terminal protocol.

References

Comer, D. 1988. *Internetworking with TCP/IP.* Prentice-Hall, Englewood Cliffs, NJ.
Cypser, R.J. 1991. *Communications for Cooperating Systems: OST, SNA, and TCP/IP.* Addison-Wesley, Reading, MA.
Sheldon, T. 1993. *Novell NetWare 4*. McGraw-Hill, New York.
Sidhu, G.S. et al. 1990. *Inside AppleTalk*. Apple Computer.
Tittle, E. and Robbins, M. 1994. *Network Design Essentials*. Academic Press, New York.
Wilder, F. 1993. *A Guide to the TCP/IP Protocol Suite*. Amacom, New York.

18.3 Communications and Information Theory

A. Brinton Cooper, III

Communications Theory

Communications science and technology have been experiencing unprecedented growth and impact in the last decade of the 20th century. The reasons are twofold. Advances in making electronic devices smaller and in reducing their requirement for electric power permit the use of complex communication processing, coding, and decoding functions that until recently languished in the research literature. On the other hand, spurred by the demands of society for more and better communication functions, scholars and their students continue to bring forth more efficient algorithms for using communication **channels** and to advance networking from the traditional techniques used by the "telephone company" to a plethora of methods for networks of portable and mobile terminals.

Yet, these advances could not have occurred without a firm theoretical basis. Reliance on intuitive and untested ideas in communications often leads down the wrong path. Shannon's idea of error-free communications at nonzero rates through noisy channels and Nyquist's notion that all the information in a continuous band-limited **signal** is contained in a string of **samples** that are taken at a uniform, finite rate ran counter to the collective intuition of practicing communications engineers of 1948 and 1928, respectively.

In what follows we introduce communications theory and information theory, believing that an introduction to the principles upon which a discipline is based provides a firm foundation for exploring the technology which it underlies.

Structure and Functions of a Communication System

The purpose of a communications system is to convey information. In order that the information be transmitted efficiently and economically and be received reliably and with fidelity, several important signal processing operations are performed prior to sending the information over the (usually noisy) channel. Complementary functions are performed at the receiver.

The center of focus of a communications system (Figure 18.3.1) is the channel. A channel is a medium of communication by which information from a **source** is conveyed to a destination (or sink). The source can be a microphone, a sensor, a video camera, a compact disc player — anything that produces an electrical signal which represents information. The destination is the recipient of the source signal. It can be a piece of magnetic tape, a computer file, a meter, or a loudspeaker. For the moment, we consider "information" to have its intuitive meaning. Information is placed on the channel (or sent over the channel) by a transmitter and obtained from the channel by a receiver. The physical nature of most communication channels renders them unsuitable for conveying information in the form produced by the source. That is, the signal produced by a channel transmitter is quite different from that produced by a source. The modulator bridges this gap by modifying one or more parameters of the transmitter output signal in accordance with the output of the source. The corresponding demodulator in the receiver recovers the information signal from the received channel waveform.

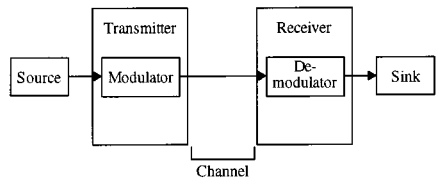


FIGURE 18.3.1 A simple communication system.

Consider the example shown in Figure 18.3.2. The source is an optical sensor that produces a continuous electrical waveform representing a picture or image of something. This waveform is processed in order to make it more suitable for transmission over a binary communication channel. A source encoder converts the analog waveform into a string of binary digits (analog to digital conversion) and removes redundant elements **(compression)** that can be reconstructed at the destination. If the information is sensitive, there may be a stage of encryption in order to prevent eavesdroppers from viewing the image. A channel encoder adds to this digit string some extra bits which are used by a channel decoder in the receiver to correct patterns of errors that can be caused by **noise**, interference, or **distortion** in the channel. A transmitter converts the digital signal to one that can be carried over the channel. At the receiving end, each of these functions is reversed, in sequence, to reproduce the original optical picture or image.

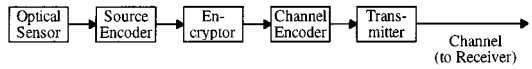


FIGURE 18.3.2 Example transmission system.

Signals

The contents of a communication are carried by a signal, typically a time-varying voltage, current, or electromagnetic wave. Information is actually conveyed in the time variations. Important distinctions among types of signals permit the development and use of methods and tools for analyzing and processing signals.

Characteristics of Signals. A signal $s(t)$ is said to be periodic if and only if $s(t + T) = s(t)$, $-\infty < t < \infty$, where T is said to be the period of $s(t)$. Any signal which is not periodic is said to be aperiodic. Familiar periodic signals include the sinusoid $\sin(2\pi ft + \phi)$, which has period $t = 1/f$. This is an important signal model in communication theory.

A signal which is a completely specified function of time at every instant of time is said to be deterministic. By contrast, the precise values of a random signal cannot be predicted in advance. Noise in a radio communication system often is modeled as a random signal. In many cases of interest, the output of a noisy communication channel can be modeled as the sum of a deterministic and a random part, $r(t) = s(t) + n(t)$, where $r(t)$ is the received waveform, $s(t)$ is the transmitted signal, and $n(t)$ is the noise induced by the channel.

A signal can be a continuous function of a continuous-time random variable, such as the sinusoid used above. On the other hand, in many systems, signals are sensed and measured only at discrete values of the time variable. This practice is the foundation of **digital communications**. Further, the signal $s(t)$ may not be a continuous function of time, but rather it may assume values from a finite set only. Such a signal is said to be a quantized or a discrete-valued signal.

Signal Representations. Analyzing the effects on communication signals of processing circuits, propagation paths, and noise and interference requires accessible mathematical representations of those signals. What follows introduces widely used and powerful representations for deterministic signals, both periodic and aperiodic. More complete treatments can be found in many excellent texts. Several are mentioned at the end of this section.

Suppose the signal $s(t)$ represents the waveform produced by someone singing. A microphone can be placed to capture the sound for display on an oscilloscope, thus permitting the viewing of $s(t)$ as a function of time. Now we know, for example, that men and women sound fundamentally different when they sing. Women are said to have voices that are “higher” than those of men. Yet, examining on the oscilloscope the waveforms produced by male and female singers may fail to reveal these differences to all but a trained observer. However, they would be captured easily by a graphical display of the frequency **spectrum**, or simply the spectrum of the singer’s voice.

The usual spectral representation is given by the Fourier transform $S(f)$ of the signal:

$$S(f) = \int_{-\infty}^{\infty} s(t)\exp(j2\pi ft)\,dt \tag{18.3.1}$$

We also say that $S(f)$ is the frequency domain representation of signal $s(t)$ and that it specifies the spectral composition of the signal. Such a tool makes it possible to determine easily the effects on a signal of filters, noisy and **bandwidth**-limited communication channels, antennas, and other devices through which it may pass.

Bandwidth. The breadth of spectral occupancy of a signal is called its bandwidth. This term makes specific a notion of the 'width' of the signal. As an important parameter of the communication channel, bandwidth measures the range of frequencies over which the channel passes energy with relatively little attenuation. (In this regard, a channel behaves as does any filter, a device which permits one or more bands of frequencies to pass relatively unattenuated while deeply attenuating all other frequencies.)

A signal whose spectrum is limited to a range of frequencies near the origin is said to be a low pass signal (Figure 18.3.3), while one whose spectrum is centered about some frequency away from the origin is called a band pass signal (Figure 18.3.4). Channels (and filters) have similar designations.

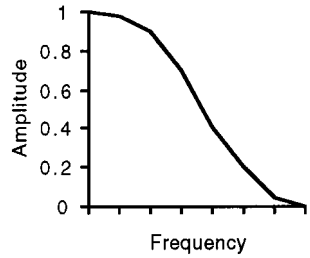


FIGURE 18.3.3 Low-pass filter.

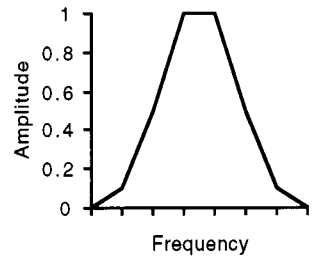


FIGURE 18.3.4 Band-pass filter.

It is not possible for physical channels and real signals to assume perfectly rectangular functional forms since neither mathematical discontinuities nor infinite slopes can occur in physical signals. Therefore, the definition of bandwidth must specify the level of attenuation at which spectral width is measured. Two definitions are commonly used. The half power or three-**decibel** (dB) bandwidth is defined by the maximum and minimum frequencies in the signal where its amplitude has dropped to $1/\sqrt{2}$ of its peak value. The definition comes from the decibel, the logarithmic expression of the ratio of power loss or gain:

$$\text{Power ratio (decibels)} = \log 10 \frac{P_{\text{out}}}{P_{\text{in}}}$$

Notice that –3 dB corresponds to a power ratio of one half. A major virtue of the 3-dB bandwidth is the ease with which it can be measured in the laboratory or in the field.

Noise in Communication Systems. Whenever a complex communication signal is transmitted through a channel, the channel output is not likely to be an exact replica of what was transmitted. If the channel is linear and time-invariant*, two types of signal modification can occur: (1) distortion, caused by a nonuniform attenuation of the signal spectrum or by a nonconstant time delay of all parts of the signal spectrum, and (2) noise. Sources of noise include thermal vibrations in the receiving circuits, nearby rotating machinery or automotive ignition systems, and natural processes in the propagation path. Most commonly experienced in communication systems is white Gaussian noise, so called because its spectrum is constant over all frequencies and because its amplitude follows a Gaussian probability law. Noise is, of course, a random phenomenon. Otherwise, the communications receiver could simply subtract the (known) noise waveform from the channel output, leaving a noise-free signal that is proportional to what was transmitted.

Some Communication Functions

Modulation. The spectra of signals such as speech, video, or computer data are concentrated at the low frequencies and are examples of **baseband** signals. The transmission of such signals by wire or cable for short distances, perhaps on the order of the dimensions of a room or office, is not difficult. An example is the wire or cable connection of loudspeakers to the amplifier of a home entertainment system. However, speech or music transmitted over 10, 50, or 100 mi of cable would experience so much attenuation that whatever is received would be quite useless. By contrast, **bandpass** signals, having spectra concentrated at much higher frequencies, are useful for such applications because they can be transmitted by the propagation of electromagnetic energy from an antenna (e.g., by radio). Radio transmission requires an antenna having physical dimensions which are approximately the wavelength of the signals to be transmitted. Typical wavelengths for mobile radio signals are approximately 1.0 cm to 1.0 m, while the wavelength of a 3000-Hz audio signal is 10^5 m, quite an impractical size for a transmitting antenna. Either baseband or bandpass signals can be transmitted via radio if suitable modulation of a sinusoidal **carrier** is used.

Modulation is the process of combining an information-bearing baseband signal $s(t)$ with a bandpass signal so that the combination is suitable for transmission over a specified communication medium. For analog modulation processes, the bandpass signal is usually a simple sinusoid, $A\sin(2\pi ft + \phi)$, the frequency of which is known as the carrier frequency f, the amplitude A, and the phase angle ϕ. A modulator modifies one or more of these parameters in accordance with the time variation of $s(t)$ as discussed below. Pulse or digital modulation requires first that $s(t)$ be digitized. The resulting string of binary digits **modulates** a parameter of the carrier in accordance with the sequence of its values.

Demodulation/Detection. The communications receiver must recover $s(t)$ from the received modulated signal for presentation to the destination. This demodulation process is more complicated than modulation because the signal has most likely been corrupted by noise in the transmission. In an analog receiver, the demodulator performs an estimation process by which it tries to determine the time-varying value of some parameter of the received carrier, e.g., its amplitude or frequency. The detector is concerned with whether or not a signal is present in noise. In the detection problem, the receiver is trying to decide which of a finite number (in this case, two) of signals was actually sent. Again, this is not a trivial problem, as the received signal is accompanied by channel noise.

Processing of the Information. Certain operations can be performed on $s(t)$ to prepare it for more efficient transmission. These operations include compression, a form of source coding which can reduce the bandwidth occupied by the signal and make the modulation, transmission, and demodulation processes more efficient. In a digital communication system, the binary representation of $s(t)$, can be further

* A linear, time-invariant channel transforms a signal $s(t)$ into $As(t+T) + n(t)$ where A is constant with time.

encoded in order to protect against channel errors. These operations are described in some detail under the subsection on "Information Theory."

Communications Techniques

Analog Modulation

Amplitude Modulation and Its Variants. Amplitude modulation (AM) continues to be widely used in everyday communications including standard broadcast radio, the Citizens' Radio Service, international shortwave broadcasting, and certain segments of the Amateur Radio Service. Its popularity stems from the relatively simple transmitters and receivers required.

Let $s(t)$ be the baseband signal to be transmitted. Then the amplitude-modulated carrier is $v(t) = A[1 + Ks(t)]\cos 2\pi f_c t$, where the carrier amplitude is A, the baseband signal amplitude is K, and the carrier frequency is f_c. We have omitted the random phase angle ϕ since it is of no significance in AM. Now, consider the modulating signal $s(t) = \cos 2\pi f_m t$.

$$v(t) = A\left(1 + K\cos 2\pi f_m t\right)\cos 2\pi f_c t$$

Trigonometric expansion gives

$$v(t) = A\cos 2\pi f_c t + \frac{KA}{2}\cos 2\pi\left(f_c + f_m\right)t + \frac{KA}{2}\cos\left(f_c - f_m\right)t \tag{18.3.2}$$

The Fourier transform of $v(t)$ is

$$\begin{aligned} V(f) = {} & \frac{A}{2}\left[\delta\left(f - f_c\right) + \delta\left(f + f_c\right)\right] + \frac{KA}{4}\left[\delta\left(f - f_c - f_m\right) + \delta\left(f + f_c + f_m\right)\right] \\ & + \frac{KA}{4}\left[\delta\left(f - f_c + f_m\right) + \delta\left(f + f_c - f_m\right)\right] \end{aligned} \tag{18.3.3}$$

where $\delta(t)$, the dirac delta function, is defined by

$$\delta(t) = 0, \qquad t \neq 0, \quad \text{and} \quad \int_{-\infty}^{\infty} \delta(t)\, dt = 1$$

This function provides a convenient way to represent sampling and time-shifting operations.

Equation (18.3.2) shows that the transmitted waveform is the sum of three terms: a carrier term and two sidebands, with each of the latter carrying the information-bearing modulation. For $K \leq 1$, the power in the carrier term is proportional to A^2, while the power in each sideband is proportional to $A^2/4$. Hence, two thirds of the power in the transmitted signal carries no useful information. It is merely "along for the ride." Further, each sideband carries the same information, so the signal occupies twice the bandwidth that is necessary to represent the modulation faithfully. Finally, on noisy channels, it is important to measure the ratio of signal power to noise power at the receiver output, $(SNR)_O$. Let $(SNR)_I$ be the signal-to-noise ratio at the receiver input. The quotient of these two ratios, a "figure of merit" for modulation schemes,

$$\frac{(SNR)_o}{(SNR)_I} = \frac{K^2}{2 + K^2}$$

is never greater than one third for AM. Compare this with other methods given below.

Each of the following three variants on AM is designed to provide savings in power and bandwidth.

Double sideband-suppressed carrier (DSB-SC) suppresses the transmission of the carrier itself, thus obtaining more efficient use of the power transmitted. However, it requires the same channel bandwidth as ordinary AM. The modulation process is represented as follows:

$$v(t) = As(t)\cos 2\pi f_c t \tag{18.3.4}$$

The Fourier transform of the DSB-SC signal is

$$V(f) = \frac{1}{2}A\left[S(f - f_c) + S(f + f_c)\right] \tag{18.3.5}$$

Now notice that all components of the signal carry the information being conveyed. No part of the signal is along for a "free ride". In the receiver, the demodulator first multiplies the received signal by $A\cos 2\pi f_c t$:

$$\begin{aligned} v_0(t) &= AA's(t)\cos(2\pi f_c t + \phi)\cos(2\pi f_c t) \\ &= (AA'/2)\cos(4\pi f_c t + \phi) + (AA'/2)s(t)\cos\phi \end{aligned} \tag{18.3.6}$$

Notice that the first term can be removed by a simple lowpass filter.* This leaves a term which is proportional to the information $s(t)$. However, as the value of ϕ (the random phase difference between the carrier and the locally generated copy of the carrier) approaches $\pi/2$, the amplitude of the coefficient of $s(t)$ will be small, and the output signal may be too weak to be useful. Even worse, ϕ may vary randomly with time, thus distorting the information $s(t)$. In practice, a phase tracking system is commonly used to keep the local oscillator "in phase" with the received signal. When such tracking is used, the receiver is said to perform coherent detection. The figure of merit for the DSB-SC is $(SNR)_O/(SNR)_I = 1$. This improvement over AM is a consequence of not having a carrier-only component in the spectrum.

Vestigial sideband (VSB) modulation is used in commercial television to convey video information. One entire sideband and a very small portion of the other are transmitted. Although carrier may or may not be sent, depending upon the application, the low cost of envelope detectors has dictated that commercial TV use VSB with a small carrier component. This is helpful because television video signals have large bandwidths and carry significant amounts of information in the low frequencies.

One can think of generating *single sideband* (SSB) modulation by generating an ordinary AM signal, then filtering out the carrier and one of the sidebands, so that the transmitted signal is either the upper sideband (USB modulation) or the lower sideband (LSB modulation) only. The Fourier transform shows that this is equivalent to a linear translation of the modulating signal $s(t)$ from baseband to frequencies near f_c. Since the transmitted signal is essentially a replica of the information-bearing signal $s(t)$, SSB is a very efficient modulation technique, placing all the transmitted power into transmitting the information and doing so with a minimal use of bandwidth. As with DSB-SC, the SSB receiver does not degrade the signal-to-noise ratio of the received signal in noise.

Frequency Modulation. The use of frequency modulation (FM) in standard radio broadcast communications has surpassed that of AM due in part to the higher fidelity afforded by the larger channel bandwidths and in part to the inherent resistance of FM receivers to propagation disturbances, electrical storms, and human-induced interference as well as the lack of many atmospheric propagation disturbances in the FM broadcast band.** FM is also used in a variety of mobile and public safety applications, in certain segments of the Amateur Radio Service and in military communications.

* A **low-pass filter** attenuates all frequencies above its **cut-off frequency** while permitting frequencies at or below that frequency to pass unhindered. In the example of DSB-SC demodulation, the cut-off frequency should be greater than f_C but less than $4f_C$.

Let $s(t)$ be the information-bearing baseband signal. The frequency modulated signal is

$$v(t) = A\cos\left[2\pi f_c t + 2\pi K\int_0^t s(\tau)\,d\tau\right] \tag{18.3.7}$$

Observe that the envelope of $v(t)$ is constant, independent of the message signal, $s(t)$.

As we did for AM, let us exhibit important properties of FM by studying a carrier that is frequency modulated with a single sinusoid. Let $s(t) = A_m\cos 2\pi f_m t$. Substituting into Equation (18.3.7), differentiating and dividing by 2π give the frequency at any instant of time as

$$\begin{aligned} f_i(t) &= f_c + KA_m\cos 2\pi f_m t \\ &= f_c + \Delta f\cos 2\pi f_m t \end{aligned}$$

where $\Delta f = KA_m$.

The quantity Δf is known as the frequency deviation and indicates the largest difference between the actual, instantaneous frequency of the FM signal and the carrier frequency f_c. Note that, while Δf is proportional to the amplitude of $s(t)$, it is independent of the frequency of $s(t)$. The derivative of the instantaneous frequency gives the phase angle of the FM signal as a function of time:

$$\phi_i(t) = 2\pi f_c t + \frac{\Delta f}{f_m}\sin 2\pi f_m t$$

The quantity $\beta = \Delta f/fm$ is called the modulation index of the FM signal. It is the maximum difference between the instantaneous value of the time-varying phase of the signal and the phase of the unmodulated carrier f_c.

The modulation index β will be used to distinguish between two types of FM systems. Write the FM signal as a function of time:

$$v(t) = A_m\cos\left(2\pi f_c t + \beta\sin 2\pi f_m t\right) \tag{18.3.8}$$

Expanding Equation 18.3.8 using the trigonometric formula for the cosine of the sum of two variables gives:

$$v(t) = A\cos 2\pi f_c t\cos\left(\beta\sin 2\pi f_m t\right) - A\sin 2\pi f_c t\sin\left(\beta\sin 2\pi f_m t\right) \tag{18.3.9}$$

When β is much smaller than one, this quickly simplifies to

$$v(t) = A\cos 2\pi f_c t - A\beta\sin 2\pi f_c t\sin 2\pi f_m t \tag{18.3.10}$$

** The resistance of FM broadcast signals to nighttime fading and other propagation anomalies is actually due to the use by FM of carrier frequencies between 88 and 108 MHz (in the VHF band) where signals travel through the atmosphere in "direct line of sight" from transmitter to receiver. Contrast this with the signals between 0.540 and 1.600 MHz (used by standard AM broadcast radio) which travel through the ionosphere where they are reflected back to earth, often hundreds or thousands of miles from the transmitter. Even so, FM is far less vulnerable to additive noise and interference (such as lightning) because it carries information in the argument of a sinusoid. Further, most FM receivers enhance this effect by passing the received signal through a **limiter** circuit which prevents the amplitude of the received signal from varying above a set value.

which represents the sum of two signals at frequency f_c, one of which has constant amplitude and the other of which has an amplitude which is proportional to the modulation. Thus, the narrowband FM signal seems to exhibit some amplitude modulation and will not have a constant amplitude. After trigonometric expansion, Equation (18.3.10) shows

$$v(t) = A\cos 2\pi f_c t + 0.5A\beta\cos 2\pi\left(f_c + f_m\right)t - 0.5A\beta\cos 2\pi\left(f_c - f_m\right)t \qquad (18.3.11)$$

Equation (18.3.11), which represents a narrowband FM signal, has the appearance of the AM signal shown in Equation (18.3.1) except for a change of sign in the last term. Thus, one might conclude that a narrowband FM signal has the same bandwidth as an AM signal. This approximation can be shown to hold so long as β is less than one.

On the other hand, when β is larger than one, the small angle approximations used in narrowband FM do not offer a valid representation of the signal. Careful analysis of the mathematics shows that the wideband FM signal with sinusoidal modulation has the following characteristics:

- Its spectrum consists of a carrier at frequency f_c and sidebands at all integer multiples of f_m above and below f_c.
- The amplitude of the carrier component of the spectrum is a function of the modulation index β. This occurs because the envelope of $v(t)$ is constant for an FM signal, so any power in the sidebands must be taken from power in the unmodulated carrier. Although the infinite number of sidebands suggests that the transmission bandwidth of the wideband FM signal must also be infinite, it is found that, at frequencies sufficiently far from the carrier, the sideband amplitudes are of insignificant magnitude to cause noticeable distortion in the demodulated signal. It has been found that the effective transmission bandwidth of an FM signal with sinusoidal modulation is $W_T = 2\Delta f + 2f_m = 2\Delta f\,(1 + 1/\beta)$.

A more accurate estimate of the required transmission bandwidth can be made by choosing to retain all sidebands whose amplitudes exceed some specified fraction of the carrier amplitude. Determination of the transmission bandwidth required to support this criterion then becomes a numerical exercise.

Digital Communications. Most modern communications advances are occurring in the rapidly expanding field of digital communications. In digital communication systems, all information is represented as strings of symbols that take values from a set of finite size. The most common is the binary representation, in which all information is represented as strings of binary (0,1) symbols. More generally, however, "digital" can imply a finite set of any size. In a quartenary system, for example, information is represented as strings of symbols that take values from an alphabet of four symbols. Industry is working on digital television, implementing what is essentially an analog function using digital techniques and circuits.

Why is this happening? In communication channels, digital signals are fundamentally more robust and resistant to corruption than are analog signals. Digital signals do not experience **intermodulation**; they do not exhibit **crosstalk**; if the amplitude is abruptly cut off by a faulty amplifier (clipping), no signal distortion results. When errors occur in a digital bitstream, error control coding techniques can be used to seek out and reverse the errors. Typically, there is no corresponding way to reverse the distortion of an analog signal. To communicate over very long distances, it is quite easy to regenerate digital pulses at sufficiently frequent intervals to avoid error or loss. The analog counterpart is a repeater that must exhibit a linear relationship between output and input signals over a wide range of input signal amplitudes. In manufacturing, the cost of digital hardware is lower than that of analog circuitry, and the reliability is higher. Digital bitstreams from a wide variety of sources can be transmitted over the same channels. They can even be intermingled and stored easily on magnetic media for later transmission, thus giving impetus to multimedia communications. Digital communications are easily encrypted to protect them from eavesdropping; this is quite difficult for analog signals and usually results in their being digitized when such protection is required. Finally, with the advent of modern, computer-controlled

telephone switching systems, the exclusive use of digital communications signals affords great ease in switching and multiplexing operations.

On the other hand, digital communications poses certain challenges. If analog information is to be transmitted via a digital channel, it must first be converted into a digital format. This requires two basic steps:

Sampling: Values of the analog signal are measured ("sampled") at equally spaced time intervals. Nothing that occurs between the sample values will affect the transmitted signal. In fact, according to Nyquist's sampling theorem, all the information contained in the original, continuous-time signal is contained in the samples, so long as they are taken at least *2B* times per second, where *B* is the highest frequency component in the signal.

Quantization: By itself, sampling is not adequate to prepare an analog signal for digital transmission. Within the limits of resolution of the sampling circuitry, each sample is a value of a continuous variable. In order to transmit the sample values in finite time, they must be represented by a finite string of symbols. Therefore, each sample is rounded or truncated to its "nearest neighbor" in the finite set; the process is known as quantization. For example, if the amplitude of the continuous signal remains within the interval $(-A_{max}, A_{max})$, the quantizer divides this interval into L levels and the quantized value is taken as the midpoint of the interval in which the continuous value falls. Each sample value, therefore, could be in error by plus or minus half the value of the sample interval. These sample values are expressed as binary numbers which are transmitted over the channel. In the receiver, they are recovered and a stepwise approximation to the original signal is built. If this stepwise approximation is completely correct, it still differs from the actual continuous signal because of the loss of amplitude information caused by the quantization process. While smoothing circuits can remove the discontinuities, they cannot assure recovery of the correct value of the analog signal. Thus, there is a residual quantization error that can be minimized in the design but never eliminated. For example, assume the use of uniform quantization (all quantization levels are the same size). If the average power in the original signal is *S*, and if $L = 2R$ where *R* is the number of bits used to encode each sample, then the signal-to-noise ratio at the quantizer output is

$$SNR_Q = \left(\frac{3S}{A_{max}^2}\right)2^{2R}$$

Additional functions such as compression and error control may be required prior to transmission over the channel. These are discussed in the subsection on "Information Theory."

Baseband Digital Modulation. Baseband digital modulation schemes are suitable for transmission of digital information over wire or cable, for example. A typical baseband system is shown in Figure 18.3.5.

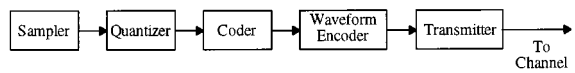


FIGURE 18.3.5 Typical digital baseband transmitter.

For each symbol to be transmitted, a waveform is chosen as its representation on the channel. Examples of waveform selection schemes include the following.

- Nonreturn-to-zero (NRZ) schemes represent a binary ONE as +V volts and ZERO as –V volts for some design value of V.
- In return-to-zero (RZ) systems, ONE is represented by +V volts and ZERO by a voltage of value zero.

- In phase-coded representations, a voltage transition occurs during each symbol interval, whether or not the symbol has changed value from the previous interval. Such a property is useful in magnetic recording systems and optical communications where arbitrarily long, constant voltages (which approximate direct current) cannot be physically transmitted. For example, in the popular Manchester **code**, ONE is represented by a positive pulse for one half the symbol period and a negative pulse for the other half, while ZERO uses the same scheme with the polarities reversed.

Bandpass Digital Modulation Schemes. The most general form of bandpass-modulated signal is $v(t) = A(t)\cos\phi(t)$. Note that the amplitude $A(t)$ or phase $\phi(t)$ of the sinusoid, or both, can vary with time and, hence, carry information via digital modulation. More specific to our uses, we can also write

$$\varphi(t) = 2\pi f_0(t) + \phi_0(t)$$

We now consider a few important examples of digital modulation. We assume that the phase of the carrier is used in the demodulation process, so that the receivers are coherent. In general, coherent modulation gives a lower error probability for a given signal-to-noise ratio than does noncoherent modulation. We consider channels in which additive, zero-mean, stationary, white Gaussian noise is added to the signal. Thus, the binary data recovered at the receiver are not guaranteed to be free of errors.

If a binary communication signal is received with power P (watts) at a speed of B bits per second, each binary symbol (or bit) contains $E_b = P/B$ joules of energy. We call E_b the signal energy per bit. A characteristic of white Gaussian noise is a uniform spectrum having constant noise power in every unit of bandwidth. If this power is N_0 watts per Hz of bandwidth, then a receiver of bandwidth W will intercept N_0W watts of channel noise power. In such cases, it is useful to consider the ratio E_b/N_0, a dimensionless signal-to-noise ratio often called the ratio of signal energy per bit-to-noise spectral density. Digital modulation schemes are compared by plotting the probability of error per bit, P_b, as a function of E_b/N_0.

Phase-shift-keying (PSK) carries digital information in discrete shifts of the carrier phase. The modulated waveform is

$$v(t) = \sqrt{\frac{2E}{T}}\cos\left(\omega_0 t + \frac{2\pi i}{M}\right) \qquad i = 1,\ldots,M, \qquad 0 \le t \le T$$

where E is the signal energy and T is the duration of the transmission of a symbol. When $M = 2$, this is called binary PSK (or simply PSK). The general case ($M \neq 2$) is often denoted MPSK. The average probability of error in the binary case is

$$P_e = \frac{1}{\sqrt{\pi}}\int_{\sqrt{E_b/N_0}}^{\infty}\exp\left(-x^2\right)dx$$

Of all the modulation techniques considered, PSK offers the minimum probability of error for fixed E_b/N_0.

Binary *frequency-shift-keying* (FSK) is one of the oldest digital bandpass modulation forms in existence. It was used extensively in early dial-up computer modems at speeds up to 1200 bits/sec, beyond which more sophisticated modulation methods are needed in order to signal faster in the same bandwidth.

Each discrete source symbol is represented by one of two frequencies, f_i:

$$v(t) = \sqrt{\frac{2A}{T}}\cos\left(2\pi f_i t + \phi\right) \qquad i = 1,2 \qquad 0 \le t \le T_b$$

where T_b = the length of a transmitted bit.

The arbitrary, constant phase term is represented by ϕ, and the general (nonbinary) case is usually denoted MFSK. The case of interest here is coherent FSK and its performance is given by

$$P_e = \frac{1}{\sqrt{\pi}} \int_{\sqrt{E_b/2N_0}}^{\infty} \exp\left(-x^2\right) dx$$

It is interesting to note that the signal-to-noise ratio for coherent binary FSK must be twice that for coherent binary PSK to give the same probability of error.

Amplitude-shift-keying (ASK) carries the information in the discrete-valued amplitude of the signal.

$$v(t) = \sqrt{\frac{2E_i(t)}{T}} \cos\left(\omega_0 t + \phi\right) \qquad i = 1,2$$

When one of the amplitude values is 0, this gives rise to "on-off" keying, surely the oldest form of digital modulation. It was commonly used in radiotelegraphy from the earliest days of radio and continues in use by amateur radio operators. In mid-1995, the U.S. Navy retired its remaining radiotelegraph operators because of the very slow rate of information exchange afforded by ASK.

Information Theory

Information is the commodity in which communication systems deal. Information is conveyed to the commuter listening to a news broadcast on the car radio during morning rush hour; it is the response to a customer's inquiry about his or her bank account; it is what the patient learns following a physical examination. In each case, the content of the information cannot be completely predicted. The information is "news"; it is something of a surprise.

Information theory provides mathematical bounds on the performance of communication systems. Specifically, it affords:

- A lower bound on the number of discrete symbols (or on the bandwidth of a continuous signal) necessary to represent a source without loss of information
- An upper bound on the rate at which information can be reliably transmitted over a noisy channel

It relies heavily on concepts of uncertainty, by which the nature of information is represented. Therefore, probabilistic concepts are needed.

Information theory provides a formal definition of information but does not provide a subjective meaning. Information is provided by a source and is transmitted over a channel where it is accepted by a destination or a sink. Sources (as well as channels and sinks) can be discrete or continuous depending upon the representation of the information.

Sources and Source Coding

A common model of an information **source** is something that periodically emits one symbol X from a known alphabet, $A = \{x_0, x_1, \ldots, x_{K-1}\}$, according to the probabilities:

$$p_j = P\left(X = x_j\right) \qquad j = 0,1,\ldots K-1$$

If successive source symbols are statistically independent, this is said to be a discrete memoryless source. From the event x_j, an observer gains an amount of information defined by $i(x_j) = \log_2(1/p_j)$, from which we conclude:

1. A certain event produces no information: $i(x_j) = 0$, $p_j = 1$.
2. The occurrence of an event cannot take away information that the observer already has: $0 \le p_j \le 1 \Rightarrow i(x_j) \ge 0$; that is, information is always positive.
3. Unlikely events produce more information than do likely events: $p_j < p_i \Rightarrow i(x_j) > i(x_i)$.
4. The contributions of successive symbols to information are additive: $i(x_i, x_j, x_k) = i(x_i) + i(x_j) + i(x_k)$.

Because the logarithm is taken to the base 2, we call the unit of information the *bit*, a contraction of *bi*nary digi*t*.

Example

Consider a fair coin (one for which the chances of a head and a tail are equal). We say that one *bit* of information is conveyed each time the coin is tossed.

Example

A modem which communicates to another modem across telephone lines by sending one of four audio tones every second is sending information at the rate of two bits per second.

The average value of the information per source symbol is an important quantity. It is given by

$$H(A) = E\left[i\left(x_j\right)\right] = \sum_{j=0}^{K-1} p_j i\left(x_j\right) = \sum_{j=0}^{K-1} p_j \log_2\left(\frac{1}{p_j}\right)$$

and is called the **entropy*** of the discrete memoryless source with source alphabet A. From this definition, it can be shown that $0 \le H(A) < \log_2 K$, where K is the size of alphabet A. In addition, $H(A) = 0$ if and only if $p_j = 1$ for some j and all other probabilities are zero. The upper bound on entropy is given by $H(A) < \log_2 K$ if and only if all symbols in the alphabet are equiprobable, that is, $p_j = 1/K$ for all j.

The output of the typical information source is rarely a sequence of equiprobable symbols, so its entropy is rarely maximum, and its transmission is not as efficient (in terms of bits of information per symbol) as is possible. In order to improve this efficiency source coding is employed. Typically, a **source code** assigns short binary sequences (or code words) to more probable source symbols and long binary sequences to less probable symbols**. Such a variable length code should also be uniquely decodable so that the original source sequence can be recovered without ambiguity from the encoded sequence. We confine our attention to binary sequences. Figure 18.3.6 shows a source coding scheme in which the output symbol s_j of the discrete memoryless source is encoded into a binary string $b_j = (b_0, b_1, \ldots, b_{l_j})$.

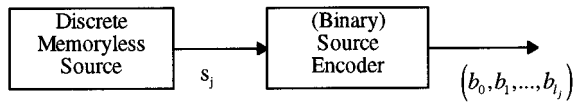


FIGURE 18.3.6 Source encoding.

If the string b_j has length l_j, then the average codeword length $\bar{L}$ is simply $\bar{L} = \Sigma_{j=0}^{K-1} p_j l_j$, and we say that the source code produces $\bar{L}$ (average) bits per source symbol. Now, let $\bar{L}_{\min}$ be the smallest possible value of $\bar{L}$. Then the coding efficiency of the source code is defined to be $\eta = \bar{L}_{\min} / \bar{L}$. The

* Mechanical engineers are likely to be familiar with the entropy ($k \ln\Omega$) from statistical mechanics. Here, k is Boltzmann's constant, Ω is the number of states in the system, and ln is the natural logarithm. This entropy measures quantitatively the randomness of the system and, therefore, is highly suggestive of information theoretic entropy. The relationship between the entropies has been the object of study by scientists and philosophers for years.

** The English language performs a heuristic form of source coding. Frequently used words such as "a", "is", "he", "it", "we", "I", "me", etc. are composed of a few letters while less-used words such as "subcutaneous" are assigned much longer strings.

fundamental bound on source codeword length is given by **Shannon's source coding theorem:** The average codeword length $\bar{L}$ for a discrete memoryless source of entropy $H(A)$ is lower-bounded by that entropy — $\bar{L} \geq H(A)$.

Lossless Compression. The discrete memoryless source provides a valid model for many physical sources, including sampled speech and the outputs of various environmental sensors. Typically, much of the information from such sources is redundant; that is, the same underlying information may be represented in the values of several output symbols. It is prudent to remove this redundancy in order to use less bandwidth and/or time for transmission. Source coding algorithms which remove such redundancy in such a manner that the original source data can be reconstructed exactly are said to perform lossless data compression. They work by assigning short sequences to the most probable source outputs and longer sequences to less probable source outputs.

A (variable length) source code in which no codeword is a prefix of another codeword is said to be a prefix code. Prefix codes are always uniquely decodable, and their average lengths obey the inequalities: $H(A) \leq \bar{L} \leq H(A) + 1$. Asymptotically, the efficiency of a prefix code approaches 1. Because of their large decoding complexities, however, we look for other classes of lossless compression algorithms. The Huffman code tries to assign to each source symbol a binary sequence, the length of which is roughly equal to the amount of information carried by that symbol. The algorithm is straightforward and transparently easy to decode. The average codeword length approaches the source entropy $H(A)$. However, the Huffman code requires knowledge of the probability distribution of source symbols, and this often is not available.

The Lempel-Ziv algorithm, by contrast, adapts to the statistics of the source as they are revealed in the source text. A binary source sequence is read left to right and parsed into segments, each of which is the shortest subsequence not encountered previously. Each of these sequences is encoded into a fixed-length binary code sequence. Thus, in contrast to Huffman codes and others, a variable number of source symbols is encoded into a fixed number of code symbols. **Lempel-Ziv** achieves, on average, a reduction of 55% in the number of symbols representing a source file; **Huffman coding** typically achieves 43%. This accounts for the enormous popularity of the former.

Rate-Distortion Theory. If the source output is a continuous variable x, the number of bits required for an exact representation of a sample value is infinite. If a finite number is used, the representation is not exact, but is often useful in many applications. The degree to which the representation $\hat{x}$ is inexact is called distortion and can be thought of as the cost associated with representing an exact source by an approximation. We write distortion as $D(x, \hat{x})$. A typical distortion function is the familiar squared-error criterion:

$$D(x, \hat{x}) = (x - \hat{x})^2$$

In a typical application, we think of a source as presenting a number n of samples per second. If each sample is, on average, represented by R bits, then we say that the source produces information at the rate of Rn bits per second. It is common, however, to normalize the source rate to the sample rate, in which case we say that we have a source of rate R (bits per sample). In principle, as R increases, we should be able to make distortion decrease. The mapping of source symbols (or sample values) to strings of bits is called a source code. Naturally, it is the aim of source code designers to minimize the distortion for a given rate. Let D be the expected distortion between two sequences of source values. We call (R,D) a rate distortion pair and say that a given (R,D) is achievable if there exists a source code having rate R and expected distortion D. The minimum rate R for which a source code exists having expected distortion D is called the **rate distortion function** $R(D)$. The average amount of uncertainty about a random variable X provided by observing another random variable Y is called the conditional entropy $H(X|Y)$. Thus, if the true source entropy is $H(X)$, then the average uncertainty provided by source coding

can be written $H(X|\hat{X})$. The average amount of information provided by one random variable X about another random variable $\hat{X}$ is called the average mutual information between the two and can be written:

$$I\left(X,\hat{X}\right) = H(X) - H\left(X|\hat{X}\right)$$

Then $R(D) = \min I(X,\hat{X})$. Shannon showed that it is possible to construct source codes that can achieve distortion D for any rate $R > R(D)$ and that it is impossible to construct source codes that can achieve arbitrary distortion for any rate below the rate distortion bound.

There is great interest in studying rate distortion bounds for various coding schemes. One of the oldest applications is in a speech compression application where 8000 samples per second are taken and quantized as eight-bit numbers to produce a 64-kb/sec uncompressed signal. Simply using the correlation between adjacent samples reduces the number of bits (hence the required channel bandwidth) by factors of two to four with little additional perceived distortion.

Of greater modern interest, however, are the efforts to compress images and video. While lossless compression methods can compress an image by a factor of three, methods employing (guaranteed loss) quantization can compress an image by a factor of 50 with what is claimed subjectively to be little loss in picture quality. A major outstanding problem is that the mean square error as a measure of distortion correlates very poorly with picture quality as judged subjectively by human viewers. A better, quantitative method of evaluating distortion in images and video is sorely needed. Alas, this is a difficult problem, and solutions do not seem immediately forthcoming.

Channels, Capacity, and Coding

On many channels, satisfactory communication is limited by noise. Whether traveling by coaxial cable or radio, as a signal gets farther from its source it grows weaker. In the cable, this attenuation is caused by the cable's nonzero resistance, which dissipates energy in the signal. In radio propagation, a variety of phenomena cause an apparent weakening of the signal, but in every case, radio waves spread out in space spherically just as circular waves spread out when a stone is dropped into a still pond. A receiving antenna can be thought of as a window of constant size which intercepts a fraction of the power in the signal. At distance R from the source, all of the signal power P (watts) passes through a sphere of area $4\pi R2$. An antenna of effective area A will, therefore, intercept $(P/4\pi R^2)A$ watts. Thus, radio signals attenuate as the square of the distance over which they travel. In either case, when the signal has been sufficiently attenuated, the noise resident in receiving equipment and the propagation medium can distort the signal or, in fact, cause bit errors.

Prior to 1948, communications engineers assumed that the constraints of noise would forever limit not only communications distance, or range, but signal quality as well. In that year, however, Claude Shannon published a remarkable theorem which continues to shape communications research and technology.

Shannon's Noisy Channel Coding Theorem. For a large and interesting class of communications channels, noise limits the rate at which information can be transmitted. In 1948, Shannon showed how to determine the channel capacity, the maximum rate at which information can be sent through the channel at an arbitrarily small error probability. He explained how noise induces uncertainty into the channel output. So, for example, if information from a source is fed to the channel at rate $H(X)$, it exits the channel with rate $H(X|Y)$, the amount of information that the receiver knows about source X when observing Y. Clearly, $H(X) \geq H(X|Y)$. So, the average amount of information passed through the channel is $I(X,Y) = H(X) - H(X|Y)$. The channel capacity then is defined as the maximum overall values of input distributions of $H(X) - H(X|Y)$; that is,

$$C = \max_{p(x)} I(X,Y)$$

Shannon's Channel Coding Theorem. There exist **channel codes** that permit communication at any rate less than capacity with as small an error probability as desired. It is not possible to achieve arbitrarily small error probability with any code at a rate greater than capacity. For example, for a channel of bandwidth W perturbed by additive white Gaussian noise having average power N, when the average received signal power is S, the capacity C of the channel is given by:

$$C = W \log_2\left(1 + \frac{S}{N}\right)$$

This example suggests that information can be encoded so that its transmission over a channel of finite bandwidth with finite signal-to-noise ratio can be received with a probability of error that is arbitrarily close to zero. Prior to this discovery, communication engineers believed that the only way to combat channel noise was to use very narrow bandwidths and send information at very low rates. They felt that errors would be inevitable in the received signal as long as the channel noise had finite power.

Although Shannon's noisy channel theorem is an impressive result, it does not show how to encode the transmitted signals in order to achieve capacity. This fact has provided employment opportunities for many coding theorists and designers for nearly a half century. Further, while most well-designed and correctly used error control codes can provide significant reductions in error rate, most do not provide an arbitrarily small error probability while sending information at capacity.

How Coding Improves Communications. While error control coding can provide a reduced error rate at the receiver output, its contribution to system performance is far more profound. Coding works by introducing *redundancy* into a stream of symbols being sent over the noisy channel. In order to transmit k binary symbols using an n-bit binary block code, n code symbols must be sent during the same amount of time that the uncoded system would send k information symbols. To support this increased transmission rate requires a larger channel bandwidth which admits proportionally more noise power into the receiver (assuming white noise) and which causes, therefore, a greater error rate in the received bit stream. Thus, the code not only has to correct errors at the original uncoded error rate, but also must actually correct errors at a higher error rate caused by the increased noise.

For example, suppose a source is transmitting noncoherent binary frequency-shift keying (FSK) at a rate of B bits per second using a transmitted power of P (watts) on a channel corrupted by white Gaussian noise. The *energy per bit* E_b is given by P/B, and the power spectral density of the noise is N_0 (watts/Hz). For binary FSK, the error probability is given by $P_e = {}^1/_2\, e^{-2E_b/N_0}$. If a code of rate R (bits/symbol) is used, then the energy per bit is reduced to $R\,E_b$, and the error probability is actually increased to $P_e = {}^1/_2\, e^{-2RE_b/N_0}$. Appropriate choice of error control code, however, will reduce the error probability not merely back to P_e, but rather to a significantly smaller number, P_d. Of course, it is always possible to achieve the same reduction in error probability merely by increasing the transmitter power from P to some larger value, P_H. The logarithmic ratio, $10\log P_H/P$, of these two values is called the *coding gain*. Thus, for a given modulation format and channel noise, an important feature afforded by coding is the use of less transmitter power than if coding is not used.

Error Control Codes. Perhaps the most widely known scheme for the control of channel errors is the single parity check that is appended to every fixed-length block of binary digits so that the number of ONEs in an augmented block is, for example, always even (or odd; the choice is arbitrary). Suppose, for example, that data are to be transmitted in four-bit blocks (b_0, b_1, b_2, b_3). A single parity check position $p4$ is set to 0 if the number of ONEs in the 4-tuple is even and to 1 if odd. (This is known as even parity. A similar rule for odd parity can be used as well.) A single error in a block of even parity will cause the number of ONEs to be odd. This odd parity condition can be detected, triggering an automatic retransmission of that portion of the data containing the block with the error condition. If the channel error rate is fairly low, this method is convenient and efficient.

However, if the channel error rate is not quite so low, more powerful error control is needed. Consider computing parity checks on subsets of the b_i. For example, using even parity, compute parity bits p_0 on

data bits b_1, b_2, and b_3; p_1 on b_0, b_2, and b_3; and p_2 on b_0, b_1, and b_2. Instead of transmitting 4 bits of information in five binary symbols (as above), we now transmit 4 bits of information using seven binary symbols. This is somewhat more inefficient, but if any single binary digit in the channel is corrupted, the location of that error in the codeword is uniquely determined by noting which parity checks fail.

One way to decode such a code is to recompute the parity checks p_0, p_1, and p_2. An error in b_0 (only) will cause p_1, and p_2 to fail. No other single error will cause p_1 and p_2, and any other parity check to fail. Similarly, a single error in b_1 will cause p_0 and p_2 to fail, etc.

This code is called a single error-correcting code because it can correct any single bit error in a block. Comparing this code with the single parity-check code that detects any odd number of errors, we see that the latter adds one redundant bit for every four data bits, while the single error-correcting code adds three redundant bits for every four data bits. We say that the single parity-check code has a rate $R = 4/5 = 0.8$ bits/symbol, while the single error-correcting code has $R = 4/7 = 0.57$ bits/symbol. Therefore, the price for error correction is, to a degree, decreased transmission rate.

The preceding single error-correcting code is a particular example (the Hamming code) of **linear block code** (LBC). The parity check computations used in encoding are succinctly expressed in its *generator matrix G:*

$$G = \begin{bmatrix} 1000011 \\ 0100101 \\ 0010111 \\ 0001110 \end{bmatrix}$$

Encoding of a 4-bit *information vector* $\boldsymbol{b} = (b_0, b_1, b_2, b_3)$ is performed by matrix multiplication:

$$\begin{aligned} v &= bG \\ &= \left(b_0, b_1, b_2, b_3, p_0, p_1, p_2\right) \end{aligned}$$

where the $\{p_j\}$ are as given above and v is the codeword produced by information vector.

Every LBC has a G-matrix having k rows and n columns and is said to be an (n,k) code; k is called the dimension of the code and n is its blocklength. For the foregoing example, $(n,k) = (7,4)$ and the code rate is

$$R = \frac{k}{n} = \frac{4}{7} = 0.57 \text{ bits/symbol}$$

The Hamming distance d between two words of an LBC is the number of positions in which they differ. The minimum Hamming distance $d_{\min}$ of an LBC is the smallest Hamming distance for all pairs of codewords. An LBC is guaranteed to correct $t = [(d_{\min} - 1)/2]$ or fewer errors in any word received from a noisy channel. Generally, codes with large values of $d_{\min}$ have small values of code rate R and vice versa, so the communication engineer faces a serious trade-off between attempted transmission speed (high rate) and error correction required to achieve a required fidelity. For example, the Varsharmov-Gilbert bound tells us that codes with rates greater than some number exist (if only we can find them):

$$\frac{k}{n} \geq 1 - \left\{ \frac{d_{\min}}{n} \log_2\left(\frac{d_{\min}}{n}\right) + \left(1 - \frac{d_{\min}}{n}\right) \log_2\left(1 - \frac{d_{\min}}{n}\right) \right\} = 1 - H\left(\frac{d_{\min}}{n}\right)$$

In the receiver, the error correction process is performed by a decoder, a digital (usually) circuit that implements an appropriate decoding algorithm. Of course, any LBC can be decoded by comparing the received word with every code word, choosing the codeword most likely to have been transmitted. In most cases, this maximum likelihood decoding procedure can be done only by looking up all the

codewords in a table, a process which is too complex and time-consuming for most practical communication situations.

Fortunately, mathematical decoding algorithms exist for most popular families of LBCs and can decode with much less complexity (but with a higher error rate) than maximum likelihood decoding.

Applications of Coding. Coding is found everywhere that digital communication is found. It has been said that if a digital communications system does not use coding, it is probably overdesigned. We find trellis codes in popular computer dial-up modems; Reed-Solomon codes in military communications equipment and in the compact disc player; convolutional codes in outer space. Error control codes will play an important part in the emerging digital cellular telephone systems and in any digital communication application where transmitter (or battery) power is at a premium.

Defining Terms

Bandpass: The characterization of a signal or a channel, the spectrum of which is centered around a frequency far from zero. Contrast with baseband.

Bandwidth: The nominal width of the spectrum of a signal or of the bandpass characteristic of an electronic filter or a communication channel.

Baseband: The characterization of a signal or a channel, the spectrum of which is concentrated at low frequencies, typically in the audio and video ranges.

Carrier: The nominal frequency at which a bandpass signal exists. For some formats, such as amplitude modulation, the carrier is the center frequency of the spectrum. It is the frequency remaining in the spectrum as the level of modulation is gradually reduced to zero.

Channel: A medium of transmission of a communication system. Typically, it refers to the physical medium and includes fiber-optic channels, voice-grade telephone channels, mobile wireless channels, and satellite communication channels.

Channel code: A code (q.v.) which is designed to control channel noise-induced errors in a transmission. Channel codes use redundant symbols to detect the presence of and to locate and correct errors.

Coherent demodulation: Any demodulation process in which the phase of the arriving signal is known, measured, or estimated with sufficient accuracy to improve the signal-to-noise ratio at the receiver output.

Code: A mapping from a set of symbols or messages into strings of symbols. Codes often, but not always, map into strings of binary symbols.

Compression: The technique of removing redundancy from a signal so that it can be transmitted in less time or use less bandwidth or so that it occupies a smaller space on storage media.

Crosstalk: Leakage of a signal being carried on one communication channel to another. It is an undesirable phenomenon that is often caused when the output of a channel is inadvertently coupled into another.

Decibel: A logarithmic expression of power ratio computed by multiplying the common logarithm of the ratio by 10.

Demodulate: In a radio receiver, to remove the information-bearing signal from the received signal.

Digital communications: The field of communications in which all information is represented as strings of symbols drawn from a set of finite size. Most commonly, the term implies binary communications, but, in fact, it refers equally well to nonbinary communications as well.

Distortion: A measure of the difference between two signals. The various forms of distortion usually arise from nonlinear phenomena.

Entropy: A measure of the average uncertainty of a random variable. For a random variable with probability distribution $p(x)$, the entropy $H(X)$ is defined as $-\sum_x p(x)\log p(x)$.

Huffman coding: A procedure that constructs the source code of minimum average length for a random variable.

Intermodulation: The unintentional modulation of one communication signal by another. It is an undesirable effect of having elements of two communication systems in too close proximity to one another.

Lempel-Ziv coding: A dictionary-based procedure for source coding that does not use the probability distribution of the source and is nonetheless asymptotically optimal.

Linear block code: A channel code with dimension k, blocklength n, and k by n generator matrix G which produces a code word when multiplied by an information block of length k.

Modulate: To combine source information with a bandpass signal at some carrier frequency.

Noise: A signal $n(t)$, the value of which at any time t is a random variable having a probability distribution that governs its values.

Quantization: A process by which the output of a continuous source is represented by one of a set of discrete values.

Rate-distortion function: The minimum rate at which a source can be described to within a given value of average distortion.

Sample: The value of a continuously varying function of time measured at a single instant.

Signal: a time-varying voltage, current, or electromagnetic wave that conveys or represents information.

Source: Anything that produces an electrical signal that represents information.

Source code: A code (q.v.) that compresses a source signal, reducing its data rate and attempting to maximize its entropy.

Spectrum: The representation of a time-varying signal in the frequency domain, usually obtained by taking the Fourier transform of the signal.

References

Cover, T.M. and Thomas, J.A. 1991. *Elements of Information Theory.* John Wiley & Sons, New York.

Gallager, R. 1968. *Information Theory and Reliable Communication.* John Wiley & Sons, New York.

Haykin, S. 1994. *Communication Systems,* 3rd ed., John Wiley & Sons, New York.

Michelson, A.M. and Levesque, A.H. 1985. *Error Control Techniques for Digital Communication.* John Wiley & Sons, New York.

Proakis, J.G. 1995. *Digital Communications,* 3rd ed. McGraw-Hill, New York.

Sklar, B. 1988. *Digital Communications: Fundamentals and Applications.* Prentice-Hall, Englewood Cliffs, NJ.

Sloane, N.J.A. and Wyner, A.D., Eds. 1993. *Claude Elwood Shannon: Collected Papers.* IEEE Press, New York.

Wicker, S.B. 1995. *Error Control Systems for Digital Communication and Storage.* Prentice-Hall, Englewood Cliffs, NJ.

Further Information

Information theory as presented here treats only the case where one user transmits one message over one channel at a time. During the past 15 years, a *multiuser information theory* has been emerging. This theory provides bounds on the information rates of more than one user transmitting over the same channel simultaneously. The beginnings of the theory are documented in the reference by Cover and Thomas. New results appear frequently in the *IEEE Transactions on Information Theory.*

Source and channel coding remain fertile areas of research and development. In addition to the *information theory* transactions, important results and some applications can be found in the *IEEE Transactions on Communications.* Special areas are given emphasis in the *IEEE Journal on Selected Areas in Communications*, nearly all the issues of which are “special” issues.

Another accessible text on digital communications is *Digital Communications*, by E. A. Lee and D. G. Messerschmitt (Kluwer, 1988).

18.4 Applications

Accessing the Internet

Lloyd W. Taylor

The Internet can provide access to a wide variety of information. Understanding what the Internet is, how it works, and what tools you can use will help in finding that fact or that person that is needed.

What Is the Internet?

The Internet is a voluntary, worldwide association of networks. There is no one "in charge" of the Internet, and anyone can belong.

There have been many attempts to define the Internet, all of which fall short in one way or another. Probably the most accurate definition is that the Internet is any interconnected network that uses the **TCP/IP** protocol suite.

A Brief History of the Internet

The Internet got its start as a project of the Defense Advanced Projects Research Agency (DARPA) in 1962. The plan was to create a military network that could survive nuclear attack by automatically rerouting traffic around network nodes that had been destroyed. The first nodes were installed in 1969. By 1971 there were 15 nodes spread across the United States.

In 1983, responsibility for the ARPANet (as it was then called) was split. The military portion of the Internet was transferred to the Department of Defense and merged with the Defense Data Network (DDN). The civilian portion continued to be operated as ARPANet. In 1986, the NSFNet was established, which superseded the ARPANet by 1990. By this time there were over 313,000 connected computers.

Use of the Internet continued to grow. In 1993, the NSF decided to get out of the business of running the Internet, and the entire system was privatized. In the meantime, a number of private Internet access providers (**IAP**s) had come to exist, largely eliminating the need for any government subsidy or operation of the Internet. By the beginning of 1994, there were over 2.2 million connected computers.

Today, there are dozens of IAPs and thousands of Internet service providers (**ISP**s) around the United States, and many more internationally. The deregulation of telephone companies in the United States has resulted in explosive growth, as the costs of long distance leased lines have dropped precipitously. Anyone with a few thousand dollars can set up an ISP by simply purchasing a Unix system and several modems and signing up for Internet access with an IAP.

The Internet works by **routing** TCP/IP packets from one router to another until the packet reaches its destination (Figure 18.4.1). Each router determines the best path to the next router by the use of routing tables it maintains, based on conversations with adjacent routers.

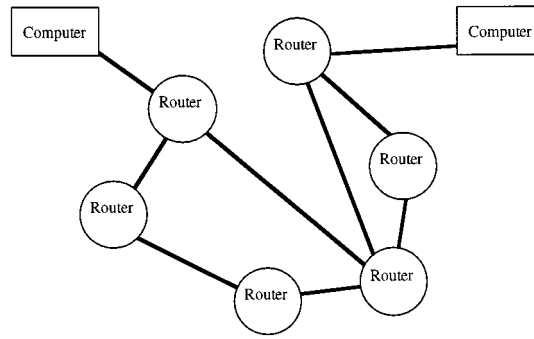


FIGURE 18.4.1 Router network.

Typically, a company will contract with an Internet access provider (IAP) for a fixed bandwidth connection to that IAP's nearest point of presence (POP). The IAP will usually interconnect its routers via high speed leased data lines. The IAP will install a leased line between its POP and a router located at the company's site, and then connect the router to the company's local network. Once the router is programmed and enabled, your company's network is part of the Internet.

Individuals can also gain access to the Internet for a monthly fee by contracting with an Internet service provider (ISP). The ISP will typically offer either a "shell account," which provides you with dialup access to a Unix system via a terminal emulator, or a "**SLIP/PPP** account," which allows you to connect your personal computer directly to the Internet via a dialup modem.

Internet Tools

There are several basic tools that can be used to access information across the Internet. Each has its particular uses and shortcomings. This section provides a brief introduction to each.

Telnet. Telnet is a tool that allows a computer to emulate a simple terminal. It provides a mechanism by which the user can connect to another computer as if that computer were physically attached to the user's workstation.

A typical telnet application will emulate one or more terminal types. The most commonly emulated terminal type is the Digital Equipment Corporation's VT100 terminal, with the IBM 3270 terminal a close second. The VT100 terminal is commonly used to communicate with Unix and VMS systems, while the IBM3270 is used to communicate with IBM mainframe systems based on the VM or MVS operating systems.

Telnet provides a text-only interface and is incapable of displaying graphics. It is a good choice when a simple, efficient connection to a remote computer is desired, or when a text-only interface is sufficient. If a graphical interface is required for an interactive session, the X-Window System is a better choice.

X-Window System. The X-Window System ("X") was developed as part of MIT's Project Athena. It is capable of full color graphics display and requires a mouse on the workstation for window and command selection.

The X-Window server runs on the user's workstation. It receives drawing commands from the remote computer and creates the requested text and graphics. Mouse clicks and keyboard input are sent to the remote computer, which responds by sending commands back to the workstation to fulfill the user's request.

In general, a window manager must be running either on the remote computer or on the user's workstation, but not both. The window manager is responsible for controlling the placement and movement of windows, for creating and destroying windows, and for starting applications on the remote computer at the request of the user.

X requires significant processing power on the local workstation and can demand significant network bandwidth for proper operation. In general, it should be run only on high-end workstations that are directly connected to an ethernet. Attempting to run X-Windows over a dialup line will result in very slow performance.

FTP. The file transfer protocol (FTP) is used to move files from one computer to another. It is capable of moving all types of files (text or binary; images, sounds, software, etc.).

FTP provides a collection of commands that allow the user to connect to remote systems, to manipulate files and directories on the remote system, and to move files back and forth between local and remote systems. These commands vary somewhat between implementations of FTP.

Many systems on the Internet offer *anonymous FTP.* This de facto standard allows remote users to connect to a public portion of the computer and to retrieve (and sometimes place) files. The standard way to use anonymous FTP is to connect to the remote computer, log in as *anonymous*, and use your email address as the password. You will generally be granted read-only access to a set of directories and files which you can then download and use.

Email. Electronic mail ("email") is traditionally the most commonly used Internet application. Communicating with colleagues and friends around the world essentially instantaneously is a powerful capability.

There are two parts to email. The user agent (UA) is the part that the user directly interacts with. This is the program on the workstation that provides the necessary commands and capabilities to create and send messages. The UA may include features such as spell checking, attachments, and address books. There are many different implementations of user agents for every kind of operating system. All of them use the same transport mechanism.

The second part of email is the transport mechanism, or message transfer agent (MTA). It is the responsibility of the MTA to receive the message from the UA, to parse the mailing instructions (e.g., To:, Cc:), and to send the message on to the intended recipients. On the Internet, all MTAs use the simple mail transport protocol (SMTP) to carry messages between systems.

At this writing, the ability to send attachments and binary files between systems remains problematic. There are several different standards for encapsulating attachments and graphics within SMTP messages. Unfortunately, these standards are not interoperable. The most commonly used encapsulation standards are MIME (multipurpose Internet mail extensions) and X.400. To successfully transfer attachments and graphics via email, you must ensure that the recipients of your message use a UA that uses a compatible encapsulation standard.

Worldwide Web. The Worldwide Web (WWW or "Web") first appeared on the Internet in 1992. It was developed as a way of sharing scientific information among researchers, but quickly was adopted by millions of people as *the* way to access information on the Internet. By 1995, half of all the traffic on the Internet was Web traffic.

A typical Web browser runs on the local workstation and passes requests for information to the remote server. The browser may use a variety of transport mechanisms to obtain the information (e.g., ftp, http — hypertext transport protocol) as directed by the server.

The unifying concept behind the power of the Web is the universal resource locator (URL). A URL is a unique address for a specific piece of information on the Web. It is made up of four parts: a protocol identifier, a host name, a path name, and an item name. Figure 18.4.2 shows a typical URL, where the protocol identifier is *http*, the host name is *www.asmenet.org*, the path name is */techaff/*, and the item name is *techprog.html*.

```
http://www.asmenet.org/techaff/techprog.html
```

FIGURE 18.4.2 A typical URL.

To retrieve this item, the Web browser will use the *http* protocol to connect to the *www.asmenet.org* server. Then it will change to the *techaff* directory and retrieve the *techprog.html* item. Because the item ends in *html*, the browser knows that this is a hypertext document and will parse and directly display the item. If it were a different type of item (Table 18.4.1), the browser would either directly display the item or call an external program to process it. In the case where there is no external program identified for the particular extension, the browser will typically offer the user the option of saving the item to disk or of canceling the download.

TABLE 18.4.1 Typical Item Extensions

Extension	File Type	Typical External Program
.gif	Compressed image	None required
.au	Audio file	SoundPlayer
.mpeg	Digital movie	MPEGPlay
.html	Hypertext	None required

Web browsers are capable of transferring information using all major transport mechanisms (e.g., ftp, gopher, http). Because of this, they largely eliminate the need to use these other tools. For those just getting started on the Internet, a good Web browser should be the first tool learned.

Finding Things on the Internet

The Internet is largely an anarchy. There is no central authority that provides an organizing structure to the information available. Anyone with an Internet connection can become an information provider by simply setting up a Webserver or an FTP server.

An inescapable consequence of this is that it can be very difficult to find specific information on the Internet. There is no rhyme or reason to where information is placed, and similar types of information may be located thousands of miles apart.

A number of organizations have taken upon themselves to try to provide some organization. They typically take one of two approaches: either a content-based approach or an automated indexing approach.

Content-based tools make use of discipline specialists to discover and index information manually. These discipline specialists comb the Internet for relevant information. When they find such information, they add a link to it from their content-based page. Thus, if you wish to find information of a specific type, you may start from a general definition of information type, and then further refine your query by clicking on the appropriate keywords. Table 18.4.2* lists a few content-based indexes current as of the date of publication of this book. A directory of content-based sites is maintained at

http://home.netscape.com/home/internet-directory.html

TABLE 18.4.2 Content-Based Information Indexes

Name	URL	Discipline
Yahoo Directory	http://www.yahoo.com/	All
McKinley Internet Directory	http://www.mckinley.com/	All
Lycos	http://www.lycos.com/	All
Virtual Tourist	http://www.vtourist.com/	Webservers sorted by geographic region

Content-based searching is most effective when you know the area in which you are interested and want to find a resource that is relevant to that area. It has a couple of drawbacks: first, if the content specialist has not yet indexed an information resource, there will be no way for you to find it; second, links *across* disciplines may not exist, limiting the breadth of the information you may be able to locate.

Automated indexing tools are based on software agents that visit every Web site on a periodic basis and index the content of those sites by keyword. A variety of indexing strategies are used, resulting in a wide range of usefulness for a particular purpose. A selection of indexing tool sites are listed in Table 18.4.3. A directory of these sites is available at

http://home.netscape.com/home/internet-search.html

TABLE 18.4.3 Index-Based Information Sites

Name	URL
InfoSeek Search	http://www.infoseek.com/
WebCrawler Search	http://webcrawler.com/
W3 Search Engine	http://cuiwww.unige.ch/meta-index.html
Altavista Search Engine	http://altavista.digital.com

* URLs and search engines come and go with surprising frequency. These specific addresses may well cease to operate during the lifetime of this book. Similar services are likely to be available - check with a local Internet specialist or reference librarian for assistance.

Automated indexing tools are useful when you are looking for information by keyword or keyphrase. Their major disadvantage is the volume of information that will be returned on simple queries. For example, a simple keyword search on "Mechanical Engineering" returned over 12,000 pointers to information.

It is therefore important to make your keywords as specific as possible. It is also important to read the guidelines for keywords on each of the servers you use. They each have their own specific search rules and syntax.

Internet Security Issues

The Internet is an inherently insecure environment. Anything that you send over the Internet can be read by someone with sufficient motivation. Because of this, it is critical to think twice about what you send. In general, the best rule of thumb is to ask yourself the question "Would I be upset if this message were printed on the front page of *The New York Times*?" If the answer is yes, you should seriously consider using encryption to protect your message. One popular freeware package that implements email encryption, and is available for all major operating systems, is PGP. Check with your system administrator to see if it is available on your system.

The Internet also provides a path for people outside your organization to access computers and information within. If your network staff has not installed an Internet firewall, your system is directly accessible to anyone. *You* are the only one who can ensure that your system is secure.

If you are using a multiuser workstation (like a Unix or VMS system), your system administrator is responsible for ensuring that known security holes have been closed. Your responsibility is to pick strong passwords (ones that do not appear in the dictionary and are hard to guess) so that someone else will not be able to compromise your account. You should change your password periodically (every 3 months or so) to help protect yourself.

Defining Terms

IAP: Internet access provider — a company that sells access to the Internet.
ISP: Internet service provider — a company that sells Internet services.
PPP: Point-to-point protocol — a standard for encapsulating multiple protocols (including TCP/IP) over dialup telephone lines.
Routing: Moving data packets from one router to the next until they reach their destination.
SLIP: Serial line internet protocol — a standard for encapsulating TCP/IP and transmitting it over dialup lines.
TCP/IP: Transmission control protocol/Internet protocol.

References

Crumlish, C. 1994. *A Guided Tour of the Internet*, Sybex, Inc.
Eddings, J. and Wattenmaker, P. 1994. *How the Internet Works.* Ziff-Davis.
Gaffin, A. 1994. *Everybody's Guide to the Internet.* Baker,
Kehoe, B.P. 1994. *Zen and the Art of the Internet.* West.
Otte, P. 1994. *The Information Superhighway: Beyond the Internet.* QUE.
Randall, N. 1994. *Teach Yourself the Internet: Around the World in 21 Days.* SAMS.
Salus, P.H. 1995. *Casting the Net: From Arpanet to Internet and Beyond.*
Tittle, E. and Robbins, M. 1994. *Internet Access Essentials.* Academic Press, New York.

Data Acquisition

Dhammika Kurumbalapitiya and S. Ratnajeevan H. Hoole

Data acquisition includes everything from gathering data, to transporting it, to storing it. The term *data acquisition* is described as the "phase of data handling that begins with sensing of variables and ends with a magnetic recording of raw data, may include a complete telemetering link" (McGraw-Hill, *Dictionary of Scientific and Technical Terms*, 2nd ed., 1978). Here the term *variables* refers to those physical quantities that are associated with a natural or artificial process. A data acquisition phase involves a real-time computing environment where the computer must be keyed to the time scale of the process. Figure 18.4.3 shows a block diagram of a data acquisition system which gives a simplified block diagram of a data acquisition system current in the early 1990s.

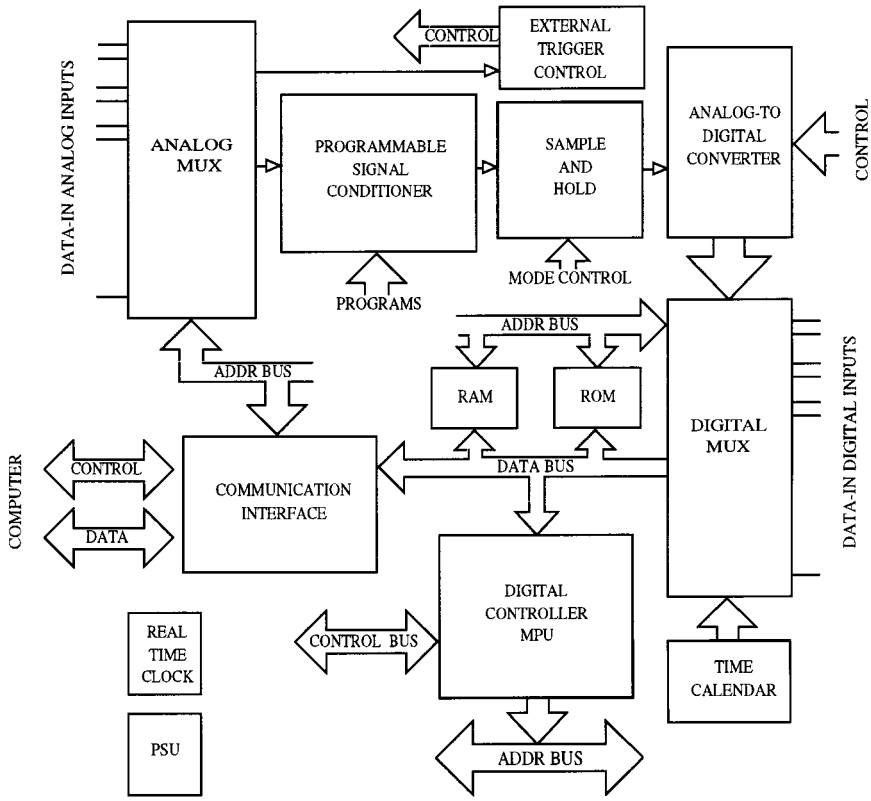


FIGURE 18.4.3 Block diagram of a data acquisition system.

The path the data travel through the system is called the data acquisition channel. Data are first captured and subsequently translated into usable signals using transducers. In this discussion, usable signals are assumed to be electrical voltages, either unipolar (that is, single ended, with a common ground so that we need just one lead wire to carry the signal) or bipolar (that is, common mode, with the signal carried by a wire pair, so that the reference of the rest of the system is not part of the output). These voltages can be either analog or digital, depending on the nature of the measurand (the quantity being captured). When there is more than one analog input, they are subsequently sent to an analog **multiplexer (MUX)**. Both the analog and the digital signals are then conditioned using signal conditioners. There are two additional steps for those conditioned analog signals. First they must be sampled and next converted to digital data. This conversion is done by **analog-to-digital converters (ADC)**.

Once the analog-to-digital conversion is done, the rest of the steps have to deal with digital data only. The calendar/clock block shown in Figure 18.4.3 is used to add the time-of-date information, an important parameter of a real-time processing environment, into the half-processed data. The digital processor performs the overall system control tasks using a software program, which is usually called system software. These control tasks also include display, printer, data recorder, and communication interface management. A well-regulated **power supply unit (PSU)** and a stable clock are essential components in many data acquisition systems. There are systems where massive amounts of data points are produced within a very short period of time, and they are equipped with *on-board memory* so that a considerable amount of data points can be stored locally. Data are transmitted to the host computer once the local storage has reached its full capacity. Historically, data acquisition evolved in modular form, until monolithic silicon came along and reduced the size of the modules.

The analysis and design of data acquisition systems are a discipline that has roots in the following subject areas: signal theory, transducers, analog signal processing, noise, sampling theory, quantizing and encoding theory, analog-to-digital conversion theory, analog and digital electronics, data communication, and systems engineering. Cost, accuracy, bit resolution, speed of operation, on-board memory, power consumption, stability of operation under various operating conditions, number of input channels and their ranges, on-board space, supply voltage requirements, compatibility with existing bus interfaces, and the types of data recording instruments involved are some of the prime factors that must be considered when designing or buying a data acquisition system. Data acquisition systems are involved in a wide range of applications, such as machine control, robot control, medical and analytical instrumentation, vibration analysis, spectral analysis, correlation analysis, transient analysis, digital audio and video, seismic analysis, test equipment, machine monitoring, and environmental monitoring.

The Analog and Digital Signal Interface

The data acquisition system must be designed to match the process being measured as well as the end-user requirements. The nature of the process is mainly characterized by its speed and number of measuring points, whereas the end-user requirement is mainly the flexibility in control. Certain processes require data acquisition with no interruption where computers are used in controlling. On the other hand, there are cases where the acquisition starts at a certain instance and continues for a definite period. In this case the acquisition cycle is repeated in a periodic manner, and it can be controlled manually or by software. Controllers access the process via the analog and digital interface submodules, which are sometimes called analog and digital front ends.

Many applications require information capturing from more than one channel. The use of the analog MUX in Figure 18.4.3 is to cater to multiple analog inputs. A detailed diagram of this input circuitry is shown in Figure 18.4.4 and the functional description is as follows. When the MUX is addressed to select an input, say $x_i(t)$, the same address will be decoded by the decoding logic to generate another address, which is used in addressing the programmable register. The programmable register contains further information regarding how to handle $x_i(t)$. The outcome of the register is then used in subsequent tuning of the signal conditioner. Complex programmable control tasks might include automatic gain selection for each channel, and hence the contents of this register are known as the channel gain list. The MUX address generator could be programmed in many ways, and one simple way is to scan the input channels in a cyclic fashion where the address can be generated by means of a binary counter. Microprocessors are also used in addressing MUXs in applications where complex channel selection tasks are involved. Multiplexers are available in integrated circuit form, though relay MUXs are widely used because they minimize errors due to cross talk and bias currents. Relay MUX modules are usually designed as plugged-in units and can be connected according to the requirements.

There are applications where the data acquisition cycle is triggered by the process itself. In this case an analog or digital trigger signal is sent to the unit by the process, and a separate external trigger interface circuitry is supplied. The internal controller assumes its duties once it has been triggered. It takes a finite time to settle the signal $x_i(t)$ through the MUX up to the signal conditioner once it is addressed. Therefore, it is possible to process $x_{i-1}(t)$ during the selection time of $x_i(t)$ for greater speeds.

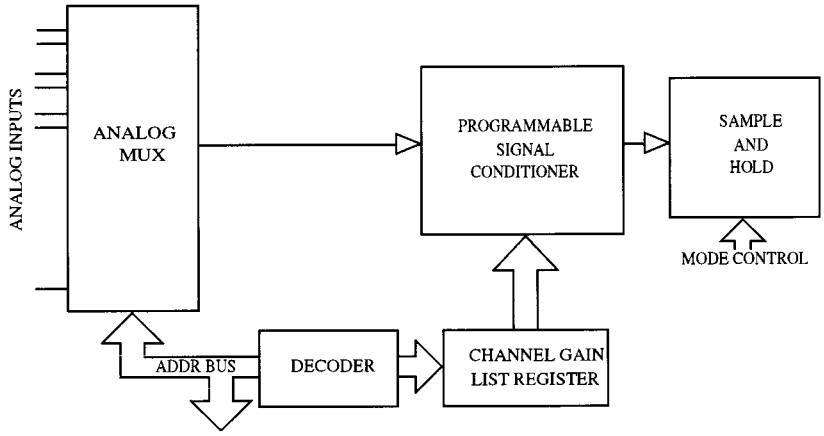


FIGURE 18.4.4 Input circuitry.

This function is known as pipelining and will be illustrated in the subsection 4, on "Analog Signal Conditioning."

In some data acquisition applications the data acquisition module is a plugged-in card in a computer, which is installed far away from the process. In such cases, transducers — the process sensing elements — are connected to the data acquisition module using transmission lines or a radio link. In the latter case a complete demodulating unit is required at the input. When transmission lines are used in the interconnection, care must be taken to minimize electromagnetic interference since transmission lines pick up noise easily. In the case of a single-ended transducer output configuration, a single wire is adequate for the signal transmission, but a common ground must be established between the two ends as given in Figure 18.4.5(a). For the transducers that have common mode outputs, a shielded twisted pair of wires will carry the signal. In this case, the shield, the transducer's encasing chassis, and the data acquisition module's reference may be connected to the same ground as shown in Figure 18.4.5(c). In high-speed applications the transmission line impedance should be matched with the output impedance of the transducer in order to prevent reflected traveling waves. If the transducer output is not strong enough to transmit for a long distance, it is best to amplify it before transmission.

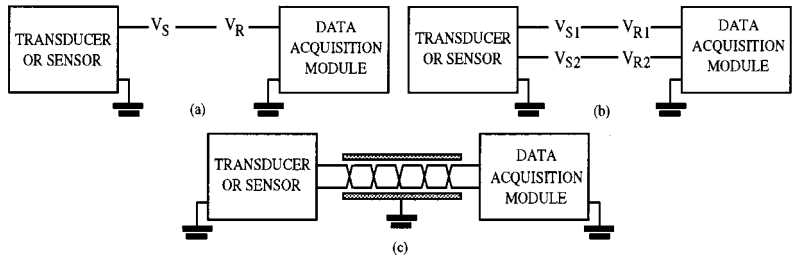


FIGURE 18.4.5 Sensor/acquisition module interconnections.

Transducers that produce digital outputs may be first connected to Schmitt trigger circuits for pulse shaping purposes, and this can be considered a form of digital signal conditioning. This becomes an essential requirement when such inputs are connected through long transmission lines where the line capacitance significantly affects the rising and falling edges of the incoming wave. Opto-isolators are sometimes used in coupling when the voltage levels of the two sides of the transducer and the input circuit of the data acquisition unit do not match each other. Special kinds of connectors are designed and widely used in interconnecting transmission lines and data acquisition equipment in order to screen

the signals from noise. Analog and digital signal grounds should be kept separate where possible to prevent digital signals from flowing in the analog ground circuit and including spurious analog signal noise.

Analog Signal Conditioning

The objective of an analog signal conditioner is to increase the quality of the transducer output to a desired level before analog-to-digital conversion. A signal conditioner mainly consists of a preamplifier, which is either an instrumentation amplifier or an operational amplifier, and/or a filter. Coupling more and more circuits to the data acquisition channel has to be done, taking great care that these signal conditioning circuits do not add more noise or unstable behavior to the data acquisition channel. General-purpose signal conditioner modules are commercially available for applications. Some details were given in the previous section about programmable signal conditioners and the discussion is continued here.

Figure 18.4.6 shows an instrumentation amplifier with programmable gain where the programs are stored in the channel-gain list. The reason for having such sophistication is to match transducer outputs with the maximum allowable input range of the ADC. This is very important in improving accuracy in cases where transducer output voltage ranges are much smaller than the full-scale input range of an ADC, as is usually the case. Indeed, this is equally true for signals that are larger than the full-scale range, and in such cases the amplifier functions as an attenuator. Furthermore, the instrumentation amplifier converts a bipolar voltage signal into a unipolar voltage with respect to the system ground. This action will reduce a major control task as far as the ADC is concerned; that is, the ADC is always sent unipolar voltages, and hence it is possible to maintain unchanged the mode control input which toggles the ADC between the unipolar and bipolar modes of an ADC.

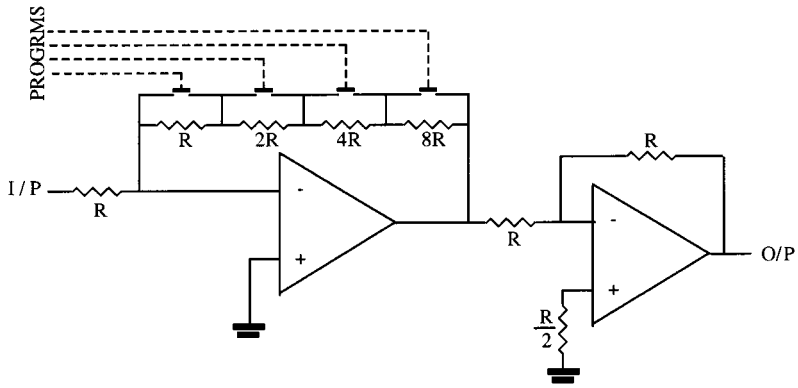


FIGURE 18.4.6 Programmable gain instrumentation amplifier.

Values of the **signal-to-noise ratio** (**SNR**)

$$\text{SNR} = \left[\frac{\text{RMS signal}}{\text{RMS noise}}\right]^2 \tag{18.4.1}$$

at the input and the output of the instrumentation amplifier are related to its **common-mode rejection ratio** (**CMRR**) given by

$$\text{CMRR} = \sqrt{\frac{\text{SNR}_{\text{output}}}{\text{SNR}_{\text{input}}}} \tag{18.4.2}$$

Hence, higher values of SNR_{output} indicate low noise power. Therefore, instrumentation amplifiers are designed to have very high CMRR figures. The existence of noise will result in an error in the ADC output. The allowable error is normally expressed as a fraction of the **least significant bit (LSB)** of the code such as ±(1/*X*)LSB. The amount of error voltage (V_{error}) corresponding to this figure can be found considering the bit resolution (*N*) and the ADC's maximum analog input voltage (V_{max}) as given in

$$V_{error} = \pm\left(\frac{V_{max}}{2^N - 1} \times \frac{1}{X}\right) \text{volt} \tag{18.4.3}$$

Other specifications of amplifiers include the temperature dependence of the input offset voltage (V_{offset}, μV/°C) and the current (V_{offset}, pA/°C) associated with the operational amplifiers in use. High slew rate (V/μs) amplifiers are recommended in high-speed applications. Generally, the higher the bandwidth, the better the performance.

Cascading a filter with the preamplifier will result in better performance by eliminating noise. Active filters are commonly used because of their compact design, but passive filters are still in use. The cutoff frequency, f_c, is one of the important performance indices of a filter that has to be designed to match the channel's requirements. The value f_c is a function of the preamplifier bandwidth, its output SNR, and the output SNR of the filter.

Sample-and-Hold and A/D Techniques in Data Acquisition

Sample-and-hold systems are primarily used to maintain a constant magnitude representing the input, across the input of the ADC throughout a precisely known period of time. Such systems are called **sample-and-hold amplifiers (SHA)**, and their characteristics are crucial to the overall system accuracy and reliability of digital data. The SHA is not an essential item in applications where the analog input does not vary more than ±(1/2)LSB of voltage. As the name indicates, an SHA operates in two different modes, which are digitally controlled. In the sampling mode it acts as an input voltage follower, where, once it is triggered into its hold mode, it should ideally retain the signal voltage level at the time of the trigger. When it is brought back into the sampling mode, it instantly assumes the voltage level at the input.

Figure 18.4.7 shows the simplified circuit diagram of a monolithic sampling-and-hold circuit and the associated switching waveforms. The differential amplifiers function as input and output buffers, and the capacitor acts as the storage mechanism. When the mode control switch is at its *on* position, the two buffers are connected in series and the capacitor follows the input with minimum time delay, if it is small. Now, if the mode control is switched *off*, the feedback loop is interrupted, and the capacitor ideally retains its terminal voltage until the next sampling signal occurs. Leakage and bias currents usually cause the capacitor to discharge and/or charge in the hold mode and the fluctuation of the hold voltage is called *droop*, which could be minimized by having a large capacitor. Therefore, the capacitance has to be selected such that the circuit performs well in both modes. Several time intervals are defined relative to the switching waveform of SHAs. The *acquisition time* (t_a) is the time taken by the device to reach its final value after the sample command has been given. The *setting time* (t_s) is the time taken to settle the output. The *aperture uncertainty* or *aperture jitter* (t_{us}) is the range of variation of the aperture time. It is important to note here that the sampling techniques have a well-formulated theoretical background.

ADCs perform a key function in the data acquisition process. The application of various ADC technologies in a data acquisition system depends mainly on the cost, bit resolution, and speed. Successive approximation types are more common at high resolution at moderate speeds (<1 MHz). This kind of ADC offers the best trade-offs among bit resolution, accuracy, speed, and cost. Flash converters, on the other hand, are best suited for high-speed applications. Integrating-type converters are suitable for high-resolution and -accuracy applications.

Many techniques have been developed in coupling sample-hold circuits and ADCs in data acquisition systems because no single ADC or sampling technology is able to satisfy the ever-increasing requirements of data acquisition applications. Figure 18.4.8 illustrates the various sampling and ADC configurations

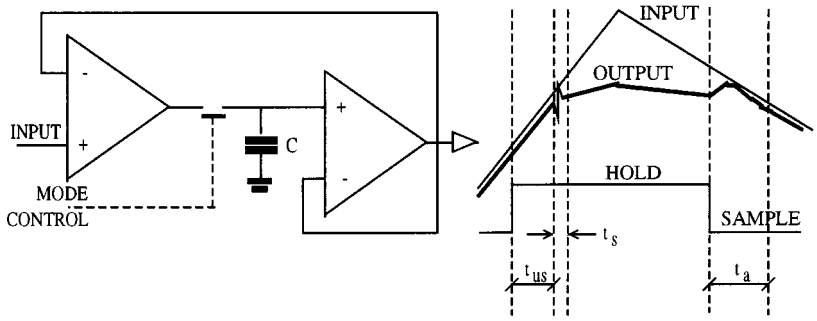


FIGURE 18.4.7 Monolithic sample-and-hold circuit.

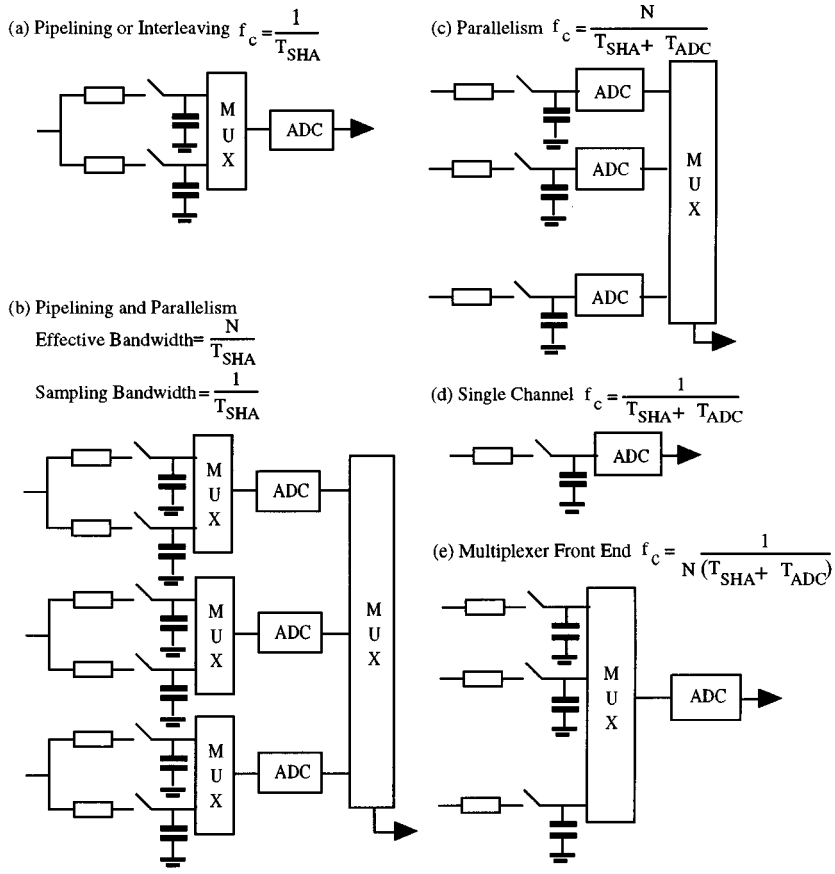


FIGURE 18.4.8 ADC and sampling configurations.

used in practice. It can be seen that the sampling frequencies are increased because of pipelining, parallelism, or concurrent architecture. The increase in the sampling frequency improves the bandwidth, improving, in turn, the SNR in the channel.

The Communication Interface of a Data Acquisition System

The communication interface is the module through which the acquired data are sent; as well, other control tasks are established between the data acquisition module and the host computer (Figure 18.4.3). There are basically two different ways of establishing a data link between the two. One way is to use

interrupts and the other is through **direct memory access (DMA)**. In the case of an interrupt-driven mode, an interrupt-request signal is sent to the computer. Upon receiving it, the computer will first finish the execution of the current instruction, suspend the next, and then send an interrupt-acknowledge signal asking the module to send data. The operation is asynchronous since the sender sends data when it wants to do so. Getting the computer ready to receive data is known as handshaking. In the case of a DMA transfer, the DMA controller is given the starting address of the memory location where the data have to be written. The DMA controller asks the computer to freeze its operations until it has finished writing data directly into the memory. The operation does not need any waiting time and therefore it is fast.

Data acquisition systems are usually designed to couple with existing computer systems, and many computer systems provide standard bus architecture, allowing users to connect various peripherals that are compatible with its bus. Data acquisition systems are computer peripherals that follow the above description. Since ADCs produce parallel data, many data acquisition systems provide outputs compatible with parallel instrument buses such as the IEEE-488 (HP-IB or GPIB) or the VMEbus. Personal computer-based data acquisition boards must have communication interfaces compatible with the computer bus in order to share resources. The RS-232 standard communication interfaces are widely used in serial data transfer. Communication interfaces for data acquisition systems are normally designed to satisfy the electrical, mechanical, and protocol standards of the interface bus. Electrical standards include power supply requirements, methods of supply, the data transfer rate (baud rate), the width of the address, and the line terminating impedance. Mechanical requirements are the type, size, and the pin assignments of the connectors. The data transfer protocol determines the procedure of data transfer between the two systems. A definition of the timing and input–output philosophy—whether the transfer is in synchronous, asynchronous, or quasi-synchronous mode and how errors are detected and handled — is an important factor to be considered.

Data Recording

It is important to provide storage media to cater to large streams of data being produced. Data acquisition systems use graph paper, paper tapes, magnetic tapes, magnetic floppy disks, hard disks, or any combination of these as their data recorders. Paper and magnetic tape storage schemes are known as sequential access storage, whereas disk storage is called direct access storage. Tapes are cost-effective media compared to disk drives and are still in wide use. In many laboratory situations it will be much more cost effective to network a number of systems to a single, high-capacity hard drive, which acts as a file server. This adoption of digital recording provides the ultimate in signal-to-noise ratio, accuracy of signal waveform, and freedom from tape transfer flutter. Data storage capacity, access time, transfer rate, and error rate are some of the performance indices that are associated with these devices.

Software Aspects

So far the discussion has been mainly on the hardware side of the data acquisition system. The other most important part is the software system associated with a data acquisition system, which can generally be divided into two — the system software and the user-interface program. Both must be designed properly in order to achieve the maximum use of the system. The system software is mainly written in assembly language with many lines of code, whereas the user interface is built using a high-level software development tool. One main part of system software is written to handle the input–output (I/O) operations. The use of assembly language results in the fast execution of I/O commands. The I/O software has to deal with how the basic input–output programming tasks such as interrupt and DMA handling are done. The other aspects of system software are to perform the internal control tasks such as providing trigger pulses for the ADC and SHA, addressing the input multiplexer, the accessing and editing of the channel-gain list, transferring data into the on-board memory, and the addition of the clock/calendar information into data. Multitasking software programs are best suited for many data acquisition systems because it may be necessary to read data from the data acquisition module and display and print it at the same time. Menu-driven user interfaces are common and have a variety of functions built into them.

Defining Terms

Analog-to-digital converter (ADC): A device that converts analog input voltage signals into digital form.

Common-mode rejection ratio (CMRR): A measure of quality of an amplifier with differential inputs and the ratio between the common-mode gain and the differential gain.

Direct memory access (DMA): The process of sending data from an external device into the computer memory with no involvement of the computer's central processing unit.

Least significant bit (LSB): The 2^0th bit in a digital word.

Multiplexer (MUX): A combinational logic device with many input channels and usually just one output. The function performed by the device is connecting one and only one input channel at a time to the output. The required input channel is selected by sending the channel address to the MUX.

Power supply unit (PSU): The unit that generates the necessary voltage levels required by a system.

Sample-and-hold amplifier (SHA): A unity gain amplifier with a mode control switch where the input of the amplifier is connected to a time-varying voltage signal. A trigger pulse at the mode control switch causes it to read the input at the instance of the trigger and maintain that value until the next trigger pulse.

Signal-to-noise ratio (SNR): The ratio between the signal power and the noise power at a point in the signal traveling path.

References

Feucht, D.L. 1990. *Handbook of Analog Circuit Design*. Academic Press, San Diego.

Fink, D.G. and Christiansen, D., Eds. 1989. *Electronic Engineers' Handbook*, 3rd ed. McGraw-Hill, New York.

Frizell, K.W. et. al. 1993. *Guidelines for PC-Based Data Acquisition Systems for Hydraulic Engineering.* American Society for Civil Engineering.

Holloway, P. 1990.Technology focus interview. *Electronic Eng.* December.

Rigby, W.H. and Dalby, T. 1994. *Computer Interfacing: A Practical Approach to Data Acquisition and Control*. Prentice-Hall, Englewood Cliffs, NJ.

Tatkow, M. and Turner, J. 1990. New techniques for high-speed data acquisition. *Electronic Eng.* September.

Ames, W.F.; et. al. “Mathematics”
Mechanical Engineering Handbook
Ed. Frank Kreith
Boca Raton: CRC Press LLC, 1999

Mathematics

William F. Ames
Georgia Institute of Technology

George Cain
Georgia Institute of Technology

Y. L. Tong
Georgia Institute of Technology

W. G. Steele
Mississippi State University

H. W. Coleman
University of Alabama

R. L. Kautz
National Institute of Standards and Technology

Dan M. Frangopol
University of Colorado

19.1 Tables

Greek Alphabet

Greek Letter	Greek Name	English Equivalent	Greek Letter	Greek Name	English Equivalent
A α	Alpha	a	N ν	Nu	n
B β	Beta	b	Ξ ξ	Xi	x
Γ γ	Gamma	g	O ο	Omicron	o
Δ δ	Delta	d	Π π	Pi	p
E ε	Epsilon	e	P ρ	Rho	r
Z ζ	Zeta	z	Σ σ ς	Sigma	s
H η	Eta	e	T τ	Tau	t
Θ θ ϑ	Theta	th	Y υ	Upsilon	u
I ι	Iota	i	Φ ϕ φ	Phi	ph
K κ	Kappa	k	X χ	Chi	ch
Λ λ	Lambda	l	Ψ ψ	Psi	ps
M μ	Mu	m	Ω ω	Omega	o

International System of Units (SI)

The International System of units (SI) was adopted by the 11th General Conference on Weights and Measures (CGPM) in 1960. It is a coherent system of units built from seven *SI base units*, one for each of the seven dimensionally independent base quantities: the meter, kilogram, second, ampere, kelvin, mole, and candela, for the dimensions length, mass, time, electric current, thermodynamic temperature, amount of substance, and luminous intensity, respectively. The definitions of the SI base units are given below. The *SI derived units* are expressed as products of powers of the base units, analogous to the corresponding relations between physical quantities but with numerical factors equal to unity.

In the International System there is only one SI unit for each physical quantity. This is either the appropriate SI base unit itself or the appropriate SI derived unit. However, any of the approved decimal prefixes, called *SI prefixes*, may be used to construct decimal multiples or submultiples of SI units.

It is recommended that only SI units be used in science and technology (with SI prefixes where appropriate). Where there are special reasons for making an exception to this rule, it is recommended always to define the units used in terms of SI units. This section is based on information supplied by IUPAC.

Definitions of SI Base Units

Meter: The meter is the length of path traveled by light in vacuum during a time interval of 1/299 792 458 of a second (17th CGPM, 1983).
Kilogram: The kilogram is the unit of mass; it is equal to the mass of the international prototype of the kilogram (3rd CGPM, 1901).
Second: The second is the duration of 9 192 631 770 periods of the radiation corresponding to the transition between the two hyperfine levels of the ground state of the cesium-133 atom (13th CGPM, 1967).
Ampere: The ampere is that constant current which, if maintained in two straight parallel conductors of infinite length, of negligible circular cross section, and placed 1 meter apart in vacuum, would produce between these conductors a force equal to 2×10^{-7} newton per meter of length (9th CGPM, 1958).
Kelvin: The kelvin, unit of thermodynamic temperature, is the fraction 1/273.16 of the thermodynamic temperature of the triple point of water (13th CGPM, 1967).
Mole: The mole is the amount of substance of a system which contains as many elementary entities as there are atoms in 0.012 kilogram of carbon-12. When the mole is used, the elementary entities must be specified and may be atoms, molecules, ions, electrons, or other particles, or specified groups of such particles (14th CGPM, 1971). Examples of the use of the mole:

- 1 mol of H_2 contains about 6.022×10^{23} H_2 molecules, or 12.044×10^{23} H atoms.
- 1 mol of HgCl has a mass of 236.04 g.
- 1 mol of Hg_2Cl_2 has a mass of 472.08 g.
- 1 mol of Hg_2^{2+} has a mass of 401.18 g and a charge of 192.97 kC.
- 1 mol of $Fe_{0.91}$ S has a mass of 82.88 g.
- 1 mol of e^- has a mass of 548.60 μg and a charge of –96.49 kC.
- 1 mol of photons whose frequency is 10^{14} Hz has energy of about 39.90 kJ.

Candela: The candela is the luminous intensity, in a given direction, of a source that emits monochromatic radiation of frequency 540×10^{12} Hz and that has a radiant intensity in that direction of (1/683) watt per steradian (16th CGPM, 1979).

Names and Symbols for the SI Base Units

Physical Quantity	Name of SI Unit	Symbol for SI Unit
Length	meter	m
Mass	kilogram	kg
Time	second	s
Electric current	ampere	A
Thermodynamic temperature	kelvin	K
Amount of substance	mole	mol
Luminous intensity	candela	cd

SI Derived Units with Special Names and Symbols

Physical Quantity	Name of SI Unit	Symbol for SI Unit	Expression in Terms of SI Base Units
Frequency[a]	hertz	Hz	s^{-1}
Force	newton	N	$m \cdot kg \cdot s^{-2}$
Pressure, stress	pascal	Pa	$N \cdot m^{-2} = m^{-1} \cdot kg \cdot s^{-2}$
Energy, work, heat	joule	J	$N \cdot m = m^2 \cdot kg \cdot s^{-2}$
Power, radiant flux	watt	W	$J \cdot s^{-1} = m^2 \cdot kg \cdot s^{-3}$
Electric charge	coulomb	C	$A \cdot s$

Physical Quantity	Name of SI Unit	Symbol for SI Unit	Expression in Terms of SI Base Units
Electric potential, electromotive force	volt	V	$J \cdot C^{-1} = m^2 \cdot kg \cdot s^{-3} \cdot A^{-1}$
Electric resistance	ohm	Ω	$V \cdot A^{-1} = m^2 \cdot kg \cdot s^{-3} \cdot A^{-2}$
Electric conductance	siemens	S	$\Omega^{-1} = m^{-2} \cdot kg^{-1} \cdot s^4 \cdot A^2$
Electric capacitance	farad	F	$C \cdot V^{-1} = m^{-2} \cdot kg^{-1} \cdot s^4 \cdot A^2$
Magnetic flux density	tesla	T	$V \cdot s \cdot m^{-2} = kg \cdot s^{-2} \cdot A^{-1}$
Magnetic flux	weber	Wb	$V \cdot s = m^2 \cdot kg \cdot s^{-2} \cdot A^{-1}$
Inductance	henry	H	$V \cdot A^{-1} \cdot s = m^2 \cdot kg \cdot s^{-2} \cdot A^{-2}$
Celsius temperature[b]	degree Celsius	°C	K
Luminous flux	lumen	lm	$cd \cdot sr$
Illuminance	lux	lx	$cd \cdot sr \cdot m^{-2}$
Activity (radioactive)	becquerel	Bq	s^{-1}
Absorbed dose (or radiation)	gray	Gy	$J \cdot kg^{-1} = m^2 \cdot s^{-2}$
Dose equivalent (dose equivalent index)	sievert	Sv	$J \cdot kg^{-1} = m^2 \cdot s^{-2}$
Plane angle	radian	rad	$1 = m \cdot m^{-1}$
Solid angle	steradian	sr	$1 = m^2 \cdot m^{-2}$

[a] For radial (circular) frequency and for angular velocity the unit rad s^{-1}, or simply s^{-1}, should be used, and this may not be simplified to Hz. The unit Hz should be used only for frequency in the sense of cycles per second.

[b] The Celsius temperature θ is defined by the equation

$$q/°C = T/K = 237.15$$

The SI unit of Celsius temperature interval is the degree Celsius, °C, which is equal to the kelvin, K. °C should be treated as a single symbol, with no space between the ° sign and the letter C. (The symbol °K, and the symbol °, should no longer be used.)

Units in Use Together with the SI

These units are not part of the SI, but it is recognized that they will continue to be used in appropriate contexts. SI prefixes may be attached to some of these units, such as milliliter, ml; millibar, mbar; mega-electronvolt, MeV; and kilotonne, kt.

Physical Quantity	Name of Unit	Symbol for Unit	Value in SI Units
Time	minute	min	60 s
Time	hour	h	3600 s
Time	day	d	86 400 s
Plane angle	degree	°	$(\pi/180)$ rad
Plane angle	minute	′	$(\pi/10\,800)$ rad
Plane angle	second	″	$(\pi/648\,000)$ rad
Length	angstrom[a]	Å	10^{-10} m
Area	barn	b	10^{-28} m^2
Volume	liter	l, L	$dm^3 = 10^{-3}$ m^3
Mass	tonne	t	$Mg = 10^3$ kg
Pressure	bar[a]	bar	10^5 Pa $= 10^5$ N $\cdot$ m^{-2}
Energy	electronvolt[b]	eV $(= e \times V)$	$\approx 1.60218 \times 10^{-19}$ J
Mass	unified atomic mass unit[b,c]	u $(= m_a(12C)/12)$	$\approx 1.66054 \times 10^{-27}$ kg

[a] The angstrom and the bar are approved by CIPM for "temporary use with SI units," until CIPM makes a further recommendation. However, they should not be introduced where they are not used at present.

[b] The values of these units in terms of the corresponding SI units are not exact, since they depend on the values of the physical constants e (for the electronvolt) and N_A (for the unified atomic mass unit), which are determined by experiment.

[c] The unified atomic mass unit is also sometimes called the dalton, with symbol Da, although the name and symbol have not been approved by CGPM.

Conversion Constants and Multipliers

Recommended Decimal Multiples and Submultiples

Multiple or Submultiple	Prefix	Symbol	Multiple or Submultiple	Prefix	Symbol
10^{18}	exa	E	10^{-1}	deci	d
10^{15}	peta	P	10^{-2}	centi	c
10^{12}	tera	T	10^{-3}	milli	m
10^{9}	giga	G	10^{-6}	micro	μ (Greek mu)
10^{6}	mega	M	10^{-9}	nano	n
10^{3}	kilo	k	10^{-12}	pico	p
10^{2}	hecto	h	10^{-15}	femto	f
10	deca	da	10^{-18}	atto	a

Conversion Factors — Metric to English

To Obtain	Multiply	By
Inches	Centimeters	0.393 700 787 4
Feet	Meters	3.280 839 895
Yards	Meters	1.093 613 298
Miles	Kilometers	0.621 371 192 2
Ounces	Grams	$3.527\,396\,195 \times 10^{-2}$
Pounds	Kilograms	2.204 622 622
Gallons (U.S. liquid)	Liters	0.264 172 052 4
Fluid ounces	Milliliters (cc)	$3.381\,402\,270 \times 10^{-2}$
Square inches	Square centimeters	0.155 000 310 0
Square feet	Square meters	10.763 910 42
Square yards	Square meters	1.195 990 046
Cubic inches	Milliliters (cc)	$6.102\,374\,409 \times 10^{-2}$
Cubic feet	Cubic meters	35.314 666 72
Cubic yards	Cubic meters	1.307 950 619

Conversion Factors — English to Metric

To Obtain	Multiply	By[a]
Microns	Mils	**25.4**
Centimeters	Inches	**2.54**
Meters	Feet	**0.3048**
Meters	Yards	**0.9144**
Kilometers	Miles	1.609 344
Grams	Ounces	28.349 523 13
Kilograms	Pounds	0.453 592 37
Liters	Gallons (U.S. liquid)	3.785 411 784
Millimeters (cc)	Fluid ounces	29.573 529 56
Square centimeters	Square inches	6.451 6
Square meters	Square feet	0.092 903 04
Square meters	Square yards	0.836 127 36
Milliliters (cc)	Cubic inches	16.387 064
Cubic meters	Cubic feet	$2.831\,684\,659 \times 10^{-2}$
Cubic meters	Cubic yards	0.764 554 858

[a] Boldface numbers are exact; others are given to ten significant figures where so indicated by the multiplier factor.

Conversion Factors — General

To Obtain	Multiply	By[a]
Atmospheres	Feet of water @ 4°C	2.950×10^{-2}
Atmospheres	Inches of mercury @ 0°C	3.342×10^{-2}
Atmospheres	Pounds per square inch	6.804×10^{-2}
Btu	Foot-pounds	1.285×10^{-3}
Btu	Joules	9.480×10^{-4}
Cubic feet	Cords	**128**
Degree (angle)	Radians	57.2958
Ergs	Foot-pounds	1.356×10^{-7}
Feet	Miles	**5280**
Feet of water @ 4°C	Atmospheres	33.90
Foot-pounds	Horsepower-hours	1.98×10^{6}
Foot-pounds	Kilowatt-hours	2.655×10^{6}
Foot-pounds per minute	Horsepower	3.3×10^{4}
Horsepower	Foot-pounds per second	1.818×10^{-3}
Inches of mercury @ 0°C	Pounds per square inch	2.036
Joules	Btu	1054.8
Joules	Foot-pounds	1.355 82
Kilowatts	Btu per minute	1.758×10^{-2}
Kilowatts	Foot-pounds per minute	2.26×10^{-5}
Kilowatts	Horsepower	0.745712
Knots	Miles per hour	0.868 976 24
Miles	Feet	1.894×10^{-4}
Nautical miles	Miles	0.868 976 24
Radians	Degrees	1.745×10^{-2}
Square feet	Acres	**43 560**
Watts	Btu per minute	17.5796

[a] Boldface numbers are exact; others are given to ten significant figures where so indicated by the multiplier factor.

Temperature Factors

$$°F = 9/5(°C) + 32$$

$$\text{Fahrenheit temperature} = 1.8(\text{temperature in kelvins}) - 459.67$$

$$°C = 5/9[(°F) - 32]$$

$$\text{Celsius temperature} = \text{temperature in kelvins} - 273.15$$

$$\text{Fahrenheit temperature} = 1.8(\text{Celsius temperature}) + 32$$

Conversion of Temperatures

From	To		From	To	
Fahrenheit	Celcius	$t_C = \frac{t_F - 32}{1.8}$	Celsius	Fahrenheit	$t_F = (t_c \times 1.8) + 32$
				Kelvin	$T_K = t_c + 273.15$
				Rankine	$T_R = (t_c + 273.15) \times 18$
	Kelvin	$T_k = \frac{t_F - 32}{1.8} + 273.15$	Kelvin	Celsius	$t_c = T_K - 273.15$
				Rankine	$T_R = T_k \times 1.8$
	Rankine	$T_R = t_F + 459.67$	Rankine	Fahrenheit	$t_F = T_R - 459.67$
				Kelvin	$T_K = \frac{T_R}{1.8}$

Physical Constants

General

Equatorial radius of the earth = 6378.388 km = 3963.34 miles (statute)
Polar radius of the earth = 6356.912 km = 3949.99 miles (statute)
1 degree of latitude at 40° = 69 miles
1 international nautical mile = 1.150 78 miles (statute) = 1852 m = 6076.115 ft
Mean density of the earth = 5.522 g/cm^3 = 344.7 lb/ft^3
Constant of gravitation $(6.673 \pm 0.003) \times 10^{-8} \cdot \text{cm}^3 \cdot \text{g}^{-1} \cdot \text{s}^{-2}$
Acceleration due to gravity at sea level, latitude 45° = 980.6194 cm/s^2 = 32.1726 ft/s^2
Length of seconds pendulum at sea level, latitude 45° = 99.3575 cm = 39.1171 in.
1 knot (international) = 101.269 ft/min = 1.6878 ft/s = 1.1508 miles (statute)/h
1 micron = 10^{-4} cm
1 angstrom = 10^{-8} cm
Mass of hydrogen atom = $(1.673\ 39 \pm 0.0031) \times 10^{-24}$ g
Density of mercury at 0°C = 13.5955 g/mL
Density of water at 3.98°C = 1.000 000 g/mL
Density, maximum, of water, at 3.98°C = 0.999 973 g/cm^3
Density of dry air at 0°C, 760 mm = 1.2929 g/L
Velocity of sound in dry air at 0°C = 331.36 m/s – 1087.1 ft/s
Velocity of light in vacuum = $(2.997\ 925 \pm 0.000\ 002) \times 10^{10}$ cm/s
Heat of fusion of water, 0°C = 79.71 cal/g
Heat of vaporization of water, 100°C = 539.55 cal/g
Electrochemical equivalent of silver 0.001 118 g/s international amp
Absolute wavelength of red cadmium light in air at 15°C, 760 mm pressure = 6438.4696 Å
Wavelength of orange-red line of krypton 86 = 6057.802 Å

π Constants

π = 3.14159 26535 89793 23846 26433 83279 50288 41971 69399 37511
$1/\pi$ = 0.31830 98861 83790 67153 77675 26745 02872 40689 19291 48091
π^2 = 9.8690 44010 89358 61883 44909 99876 15113 53136 99407 24079
$\log_e \pi$ = 1.14472 98858 49400 17414 34273 51353 05871 16472 94812 91531
$\log_{10} \pi$ = 0.49714 98726 94133 85435 12682 88290 89887 36516 78324 38044
$\log_{10} \sqrt{2\pi}$ = 0.39908 99341 79057 52478 25035 91507 69595 02099 34102 92128

Constants Involving *e*

e = 2.71828 18284 59045 23536 02874 71352 66249 77572 47093 69996
$1/e$ = 0.36787 94411 71442 32159 55237 70161 46086 74458 11131 03177
e^2 = 7.38905 60989 30650 22723 04274 60575 00781 31803 15570 55185
$M = \log_{10} e$ = 0.43429 44819 03251 82765 11289 18916 60508 22943 97005 80367
$1/M = \log_e 10$ = 2.30258 50929 94045 68401 79914 54684 36420 76011 01488 62877
$\log_{10} M$ = 9.63778 43113 00536 78912 29674 98645 – 10

Numerical Constants

$\sqrt{2}$ = 1.41421 35623 73095 04880 16887 24209 69807 85696 71875 37695
$\sqrt[3]{2}$ = 1.25992 10498 94873 16476 72106 07278 22835 05702 51464 70151
$\log_e 2$ = 0.69314 71805 59945 30941 72321 21458 17656 80755 00134 36026
$\log_{10} 2$ = 0.30102 99956 63981 19521 37388 94724 49302 67881 89881 46211
$\sqrt{3}$ = 1.73205 08075 68877 29352 74463 41505 87236 69428 05253 81039
$\sqrt[3]{3}$ = 1.44224 95703 07408 38232 16383 10780 10958 83918 69253 49935
$\log_e 3$ = 1.09861 22886 68109 69139 52452 36922 52570 46474 90557 82275
$\log_{10} 3$ = 0.47712 12547 19662 43729 50279 03255 11530 92001 28864 19070

Symbols and Terminology for Physical and Chemical Quantities

Name	Symbol	Definition	SI Unit
Classical Mechanics			
Mass	m		kg
Reduced mass	μ	$\mu = m_1 m_2/(m_1 + m_2)$	kg
Density, mass density	ρ	$\rho = m/V$	$kg \cdot m^{-3}$
Relative density	d	$d = \rho/\rho^{\theta}$	1
Surface density	ρ_A,ρ_S	$\rho_a = m/A$	$kg \cdot m^{-2}$
Specific volume	ν	$\nu = V/m = 1/\rho$	$m^3 \cdot kg^{-1}$
Momentum	p	$p = mv$	$kg \cdot m \cdot s^{-1}$
Angular momentum, action	L	$L = r \times p$	$J \cdot s$
Moment of inertia	I, J	$I = \Sigma\, m_i\, r_i^2$	$kg \cdot m^2$
Force	F	$F = d\,p/d\,t = m\,a$	N
Torque, moment of a force	T, (M)	$T = r \times F$	$N \cdot m$
Energy	E		J
Potential energy	E_p, V, Φ	$E_p = \int F \cdot ds$	J
Kinetic energy	E_k, T, K	$E_k = (1/2)\, mv^2$	J
Work	W, w	$W = \int F \cdot ds$	J
Hamilton function	H	$H(q,p) = T(q,p) + V(q)$	J
Lagrange function	L	$L(q,\dot{q}) = T(q,\dot{q}) - V(q)$	J
Pressure	p, P	$p = F/A$	Pa, $N \cdot m^{-2}$
Surface tension	γ, σ	$\gamma = dW/dA$	$N \cdot m^{-1}$, $J \cdot m^{-1}$
Weight	G (W, P)	$G = mg$	N
Gravitational constant	G	$F = Gm_1m_2/r^2$	$N \cdot m^2 \cdot kg^{-2}$
Normal stress	σ	$\sigma = F/A$	Pa
Shear stress	τ	$\tau = F/A$	Pa
Linear strain, relative elongation	ε, e	$\varepsilon = \Delta l/l$	1
Modulus of elasticity, Young's modulus	E	$E = \sigma/\varepsilon$	Pa
Shear strain	γ	$\gamma = \Delta x/d$	1
Shear modulus	G	$G = \tau/\gamma$	Pa
Volume strain, bulk strain	θ	$\theta = \Delta V/V_0$	1
Bulk modulus, compression modulus	K	$K = V_0(dp/dV)$	Pa
Viscosity, dynamic viscosity	η, μ	$\tau_{x,z} = \eta(dv_x/dz)$	$Pa \cdot s$
Fluidity	ϕ	$\phi = l/\eta$	$m \cdot kg^{-1} \cdot s$
Kinematic viscosity	ν	$\nu = \eta/\rho$	$m^2 \cdot s^{-1}$
Friction coefficient	μ, (f)	$F_{frict} = \mu F_{norm}$	1
Power	P	$P = dW/dt$	W
Sound energy flux	P, P_a	$P = dE/dt$	W
Acoustic factors			
Reflection factor	ρ	$\rho = P_r/P_0$	1
Acoustic absorption factor	α_a, (α)	$\alpha_a = 1 - \rho$	1
Transmission factor	τ	$\tau = P_t/P_0$	1
Dissipation factor	δ	$\delta = \alpha_a - \tau$	1

Name	Symbol	Definition	SI Unit
Classical Mechanics			
Electricity and Magnetism			
Quantity of electricity, electric range	Q		C
Charge density	ρ	$\rho = Q/V$	$C \cdot m^{-3}$
Surface charge density	σ	$\sigma = Q/A$	$C \cdot m^{-2}$
Electric potential	V, ϕ	$V = dW/dQ$	$V, J \cdot C^{-1}$
Electric potential difference	U, ΔV, $\Delta\phi$	$U = V_2 - V_1$	V
Electromotive force	E	$E = \int(F/Q) \cdot ds$	V
Electric field strength	E	$E = F/Q = -\text{grad } V$	$V \cdot m^{-1}$
Electric flux	ψ	$\psi = \int D \cdot dA$	C
Electric displacement	D	$D = \varepsilon E$	$C \cdot m^{-2}$
Capacitance	C	$C = Q/U$	$F, C \cdot V^{-1}$
Permittivity	ε	$D = \varepsilon E$	$F \cdot m^{-1}$
Permittivity of vacuum	ε_0	$\varepsilon_0 = \mu_0^{-1} c_0^{-2}$	$F \cdot m^{-1}$
Relative permittivity	ε_r	$\varepsilon_r = \varepsilon/\varepsilon_0$	1
Dielectric polarization (dipole moment per volume)	P	$P = D - \varepsilon_0 E$	$C \cdot m^{-2}$
Electric susceptibility	χ_e	$\chi_e = \varepsilon_\tau - 1$	1
Electric dipole moment	p, μ	$P = Qr$	$C \cdot m$
Electric current	I	$I = dQ/dt$	A
Electric current density	j, J	$I = \int j \cdot dA$	$A \cdot m^{-2}$
Magnetic flux density, magnetic induction	B	$F = Qv \times B$	T
Magnetic flux	Φ	$\Phi = \int B \cdot dA$	Wb
Magnetic field strength	H	$B = \mu H$	$A \cdot M^{-1}$
Permeability	μ	$B = \mu H$	$N \cdot A^{-2}$, $H \cdot m^{-1}$
Permeability of vacuum	μ_0		$H \cdot m^{-1}$
Relative permeability	μ_t	$\mu_\tau = \mu/\mu_0$	1
Magnetization (magnetic dipole moment per volume)	M	$M = B/\mu_0 - H$	$A \cdot m^{-1}$
Magnetic susceptibility	χ, κ, (χ_m)	$\chi = \mu_\tau - 1$	1
Molar magnetic susceptibility	χ_m	$\chi_m = V_m \chi$	$m^3 \cdot mol^{-1}$
Magnetic dipole moment	m, μ	$E_p = -m \cdot B$	$A \cdot m^2$, $J \cdot T^{-1}$
Electrical resistance	R	$R = U/I$	Ω
Conductance	G	$G = 1/R$	S
Loss angle	δ	$\delta = (\pi/2) + \phi_I - \phi_U$	l, rad
Reactance	X	$X = (U/I) \sin \delta$	Ω
Impedance (complex impedance)	Z	$Z = R + iX$	Ω
Admittance (complex admittance)	Y	$Y = 1/Z$	S
Susceptance	B	$Y = G + iB$	S
Resistivity	ρ	$\rho = E/j$	$\Omega \cdot m$
Conductivity	κ, γ, σ	$\kappa = 1/\rho$	$S \cdot m^{-1}$
Self-inductance	L	$E = -L(dI/dt)$	H
Mutual inductance	M, L_{12}	$E_1 = L_{12}(dI_2/dt)$	H
Magnetic vector potential	A	$B = \nabla \times A$	$Wb \cdot m^{-1}$
Poynting vector	S	$S = E \times H$	$W \cdot m^{-2}$
Electromagnetic Radiation			
Wavelength	λ		m
Speed of light			
In vacuum	c_0		$m \cdot s^{-1}$
In a medium	c	$c = c_0/n$	$m \cdot s^{-1}$
Wavenumber in vacuum	$\tilde{\nu}$	$\tilde{\nu} = \nu/c_0 = 1/n\lambda$	m^{-1}
Wavenumber (in a medium)	σ	$\sigma = 1/\lambda$	m^{-1}
Frequency	ν	$\nu = c/\lambda$	Hz
Circular frequency, pulsatance	ω	$\omega = 2\pi\nu$	s^{-1}, $rad \cdot s^{-1}$
Refractive index	n	$n = c_0/c$	1
Planck constant	h		$J \cdot s$

Name	Symbol	Definition	SI Unit
		Classical Mechanics	
Planck constant/2π	$\bar{h}$	$\bar{h} = h/2\pi$	J · s
Radiant energy	Q, W		J
Radiant energy density	ρ, w	$\rho = Q/V$	J · m^{-3}
Spectral radiant energy density			
In terms of frequency	ρ_ν, w_ν	$\rho_\nu = d\rho/d\nu$	J · m^{-3} · Hz^{-1}
In terms of wavenumber	$\rho_{\tilde{\nu}}$ $w_{\tilde{\nu}}$	$\rho_\nu = d\rho / d\tilde{\nu}$	J · m^{-2}
In terms of wavelength	ρ_λ, w_λ	$\rho_\lambda = d\rho/d\lambda$	J · m^{-4}
Einstein transition probabilities			
Spontaneous emission	A_{nm}	$dN_n/dt = -A_{nm}N_n$	s^{-2}
Stimulated emission	B_{nm}	$dN_n/dt = -\rho_\nu\,(\tilde{\nu}_{nm}) \times B_{nm}N_n$	s · kg^{-1}
Stimulated absorption	B_{nm}	$dN_n/dt = \rho_{\tilde{\nu}}(\tilde{\nu}_{nm})\, B_{nm}N_n$	s · kg^{-1}
Radiant power, radiant energy per time	Φ, P	$\Phi = dQ/dt$	W
Radiant intensity	I	$I = d\Phi/d\Omega$	W · sr^{-1}
Radiant excitance (emitted radiant flux)	M	$M = d\Phi/dA_{\text{source}}$	W · m^{-2}
Irradiance (radiant flux received)	E, (I)	$E = d\Phi/dA$	W · m^{-2}
Emittance	ε	$\varepsilon = M/M_{\text{bb}}$	1
Stefan-Boltzmann constant	σ	$M_{\text{bb}} = \sigma T^4$	W · m^{-2} · K^{-4}
First radiation constant	c_1	$c_1 = 2\pi hc_0^2$	W · m^{-2}
Second radiation constant	c_2	$c_2 = hc_0/k$	K · m
Transmittance, transmission factor	τ, T	$\tau = \Phi_{\text{tr}}/\Phi_0$	1
Absorptance, absorption factor	α	$\alpha = \Phi_{\text{abs}}/\Phi_0$	1
Reflectance, reflection factor	ρ	$\rho = \Phi_{\text{refl}}/\Phi_0$	1
(Decadic) absorbance	A	$A = \lg(1 - \alpha_i)$	1
Napierian absorbance	B	$B = \ln(1 - \alpha_i)$	1
Absorption coefficient			
(Linear) decadic	a, K	$a = A/l$	m^{-1}
(Linear) napierian	α	$\alpha = B/l$	m^{-1}
Molar (decadic)	ε	$\varepsilon = a/c = A/cl$	m^2 · mol^{-1}
Molar napierian	κ	$\kappa = a/c = B/cl$	m^2 · mol^{-1}
Absorption index	k	$k = \alpha/4\pi\ \tilde{\nu}$	l
Complex refractive index	$\hat{n}$	$\hat{n} = n + ik$	1
Molar refraction	R, R_m	$R = \dfrac{(n^2-1)}{(n^2+2)} V_m$	m^3 · mol^{-1}
Angle of optical rotation	α		1, rad
		Solid State	
Lattice vector	R, R_0		m
Fundamental translation vectors for the crystal lattice	a_1; a_2; a_3, a; b; c	$R = n_1a_1 + n_2a_2 + n_3a_3$	m
(Circular) reciprocal lattice vector	G	$G \cdot R = 2\pi m$	m^{-1}
(Circular) fundamental translation vectors for the reciprocal lattice	b_1; b_2; b_3, a^*; b^*; c^*	$a_1 \cdot b_k = 2\pi\delta_{ik}$	m^{-1}
Lattice plane spacing	d		m
Bragg angle	θ	$n\lambda = 2d \sin\theta$	1, rad
Order of reflection	n		1
Order parameters			
Short range	σ		1
Long range	s		1
Burgers vector	b		m
Particle position vectort	r, R_j		m
Equilibrium position vector of an ion	R_0		m
Displacement vector of an ion	u	$u = R - R_0$	m
Debye-Waller factor	B, D		1
Debye circular wavenumber	q_D		m^{-1}
Debye circular frequency	ω_D		s^{-1}
Grüneisen parameter	γ,Γ	$\gamma = aV/\kappa C_v$	1

Name	Symbol	Definition	SI Unit
		Classical Mechanics	
Madelung constant	α, M	$E_{\text{coul}} = \dfrac{\alpha N_A z + z - e^2}{4\pi\varepsilon_0 R_0}$	1
Density of states	N_E	$N_E = dN(E)/dE$	$J^{-1} \cdot m^{-3}$
(Spectral) density of vibrational modes	N_ω, g	$N_\omega = dN(\omega)/d\omega$	$s \cdot m^{-3}$
Resistivity tensor	ρ_{ik}	$E = \rho \cdot j$	$\Omega \cdot m$
Conductivity tensor	σ_{ik}	$\sigma = \rho^{-1}$	$S \cdot m^{-1}$
Thermal conductivity tensor	λ_{ik}	$J_q = -\lambda \cdot \text{grad}\, T$	$W \cdot m^{-1} \cdot K^{-1}$
Residual resistivity	ρ_R		$\Omega \cdot m$
Relaxation time	τ	$\tau = l/v_F$	s
Lorenz coefficient	L	$L = \lambda/\sigma T$	$V^2 \cdot K^{-2}$
Hall coefficient	A_H, R_H	$E = \rho \cdot j + R_H(B \times j)$	$m^3 \cdot C^{-1}$
Thermoelectric force	E		V
Peltier coefficient	Π		V
Thomson coefficient	μ, (τ)		$V \cdot K^{-1}$
Work function	Φ	$\Phi = E_\infty - E_F$	J
Number density, number concentration	n, (p)		m^{-3}
Gap energy	E_g		J
Donor ionization energy	E_d		J
Acceptor ionization energy	E_a		J
Fermi energy	E_F, ε_F		J
Circular wave vector, propagation vector	k, q	$k = 2\pi/\lambda$	m^{-1}
Bloch function	$u_k(r)$	$\psi(r) = u_k(r) \exp(ik \cdot r)$	$m^{-3/2}$
Charge density of electrons	ρ	$\rho(r) = -e\psi^*(r)\psi(r)$	$C \cdot m^{-3}$
Effective mass	m^*		kg
Mobility	m	$\mu = v_{\text{drift}}/E$	$m^2 \cdot V^{-1} \cdot s^{-1}$
Mobility ratio	b	$b = \mu_n/\mu_p$	1
Diffusion coefficient	D	$dN/dt = -DA(dn/dx)$	$m^2 \cdot s^{-1}$
Diffusion length	L	$L = \sqrt{D\tau}$	m
Characteristic (Weiss) temperature	ϕ, ϕ_w		K
Curie temperature	T_c		K
Neel temperature	T_N		K

Elementary Algebra and Geometry

Fundamental Properties (Real Numbers)

$a + b = b + a$	Commutative law for addition
$(a + b) + c = a + (b + c)$	Associative law for addition
$a + 0 = 0 + a$	Identity law for addition
$a + (-a) = (-a) + a = 0$	Inverse law for addition
$a(bc) = (ab)c$	Associative law for multiplication
$a\left(\frac{1}{a}\right) = \left(\frac{1}{a}\right)a = 1,\ a \neq 0$	Inverse law for multiplication
$(a)(1) = (1)(a) = a$	Identity law for multiplication
$ab = ba$	Commutative law for multiplication
$a(b + c) = ab + ac$	Distributive law

Division by zero is not defined.

Exponents

For integers m and n,

$$a^n a^m = a^{n+m}$$

$$a^n / a^m = a^{n-m}$$

$$\left(a^n\right)^m = a^{nm}$$

$$(ab)^m = a^m b^m$$

$$(a/b)^m = a^m / b^m$$

Fractional Exponents

$$a^{p/q} = \left(a^{1/q}\right)^p$$

where $a^{1/q}$ is the positive qth root of a if $a > 0$ and the negative qth root of a if a is negative and q is odd. Accordingly, the five rules of exponents given above (for integers) are also valid if m and n are fractions, provided a and b are positive.

Irrational Exponents

If an exponent is irrational (e.g., $\sqrt{2}$), the quantity, such as $a^{\sqrt{2}}$, is the limit of the sequence $a^{1.4}$, $a^{1.41}$, $a^{1.414}$,…

Operations with Zero

$$0^m = 0 \qquad a^0 = 1$$

Logarithms

If x, y, and b are positive $b \neq 1$,

$$\log_b(xy) = \log_b x + \log_b y$$

$$\log_b(x/y) = \log_b x - \log_b y$$

$$\log_b x^p = p\log_b x$$

$$\log_b(1/x) = -\log_b x$$

$$\log_b b = 1$$

$$\log_b 1 = 0 \qquad \textit{Note: } b^{\log_b x} = x$$

Change of Base ($a \neq 1$)

$$\log_b x = \log_a x \log_b a$$

Factorials

The factorial of a positive integer n is the product of all the positive integers less than or equal to the integer n and is denoted $n!$. Thus,

$$n! = 1 \cdot 2 \cdot 3 \cdot \cdots \cdot n$$

Factorial 0 is defined: 0! = 1.

Stirling's Approximation

$$\lim_{n\to\infty}\left(n/e^n\right)^n \sqrt{2\pi n} = n!$$

Binomial Theorem

For positive integer n

$$(x+y)^n = x^n + nx^{n-1}y + \frac{n(n-1)}{2!}x^{n-2}y^2 + \frac{n(n-1)(n-2)}{3!}x^{n-3}y^3 + \cdots + nxy^{n-1} + y^n$$

Factors and Expansion

$$(a+b)^2 = a^2 + 2ab + b^2$$

$$(a-b)^2 = a^2 - 2ab + b^2$$

$$(a+b)^3 = a^3 + 3a^2b + 3ab^2 + b^3$$

$$(a-b)^3 = a^3 - 3a^2b + 3ab^2 - b^3$$

$$\left(a^2 - b^2\right) = (a-b)(a+b)$$

$$\left(a^3 - b^3\right) = (a-b)\left(a^2 + ab + b^2\right)$$

$$\left(a^3 + b^3\right) = (a+b)\left(a^2 - ab + b^2\right)$$

Progression

An *arithmetic progression* is a sequence in which the difference between any term and the preceding term is a constant (d):

$$a, a+d, a+2d, \ldots, a+(n-1)d$$

If the last term is denoted l [$= a + (n - 1)d$], then the sum is

$$s = \frac{n}{2}(a+l)$$

A *geometric progression* is a sequence in which the ratio of any term to the preceding term is a constant r. Thus, for n terms,

$$a, ar, ar^2, \ldots, ar^{n-1}$$

The sum is

$$S = \frac{a - ar^n}{1-r}$$

Complex Numbers

A complex number is an ordered pair of real numbers (a, b).

Equality: $(a,b) = (c,d)$ if and only if $a = c$ and $b = d$

Addition: $(a,b) + (c,d) = (a + c, b + d)$

Multiplication: $(a,b)(c,d) = (ac - bd, ad + bc)$

The first element (a, b) is called the *real* part, the second the *imaginary* part. An alternative notation for (a, b) is $a + bi$, where $i^2 = (-1, 0)$, and $i = (0, 1)$ or $0 + 1i$ is written for this complex number as a convenience. With this understanding, i behaves as a number, that is, $(2 - 3i)(4 + i) = 8 + 2i - 12i - 3i^2 = 11 - 10i$. The conjugate of a $a + bi$ is $a - bi$, and the product of a complex number and its conjugate is $a^2 + b^2$. Thus, *quotients* are computed by multiplying numerator and denominator by the conjugate of the denominator, as illustrated below:

$$\frac{2+3i}{4+2i} = \frac{(4-2i)(2+3i)}{(4-2i)(4+2i)} = \frac{14+8i}{20} = \frac{7+4i}{10}$$

Polar Form

The complex number $x + iy$ may be represented by a plane vector with components x and y:

$$x + iy = r(\cos\theta + i\sin\theta)$$

(See Figure 19.1.1.). Then, given two complex numbers $z_1 = \tau_1(\cos\theta_1 + i \sin\theta_1)$ and $z_2 = \tau_2(\cos\theta_2 + i \sin\theta_2)$, the product and quotient are:

Product: $$z_1 z_2 = r_1 r_2\left[\cos(\theta_1+\theta_2)+i\sin(\theta_1+\theta_2)\right]$$

Quotient: $$z_1/z_2 = (r_1/r_2)\left[\cos(\theta_1-\theta_2)+i\sin(\theta_1-\theta_2)\right]$$

Powers: $$z^n = \left[r(\cos\theta+i\sin\theta)\right]^n = r^n[\cos n\theta+i\sin n\theta]$$

Roots: $$z^{1/n} = \left[r(\cos\theta+i\sin\theta)\right]^{1/n}$$

$$= r^{1/n}\left[\cos\frac{\theta+k\cdot 360}{n}+i\sin\frac{\theta+k\cdot 360}{n}\right]$$

$$k = 0,\ 1,\ 2,\ \ldots,\ n-1$$

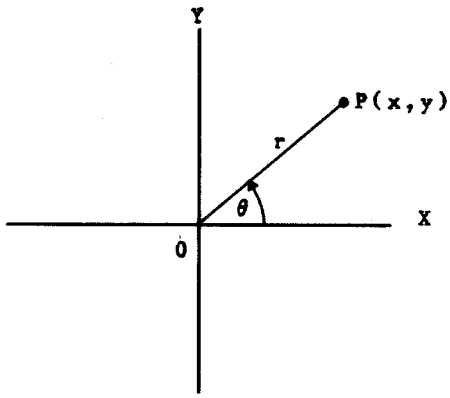


FIGURE 19.1.1 Polar form of complex number.

Permutations

A permutation is an ordered arrangement (sequence) of all or part of a set of objects. The number of permutations of n objects taken r at a time is

$$p(n,r) = n(n-1)(n-2)\cdots(n-r+1)$$

$$= \frac{n!}{(n-r)!}$$

A permutation of positive integers is "even" or "odd" if the total number of inversions is an even integer or an odd integer, respectively. Inversions are counted relative to each integer j in the permutation by counting the number of integers that follow j and are less than j. These are summed to give the total number of inversions. For example, the permutation 4132 has four inversions: three relative to 4 and one relative to 3. This permutation is therefore even.

Combinations

A combination is a selection of one or more objects from among a set of objects regardless of order. The number of combinations of n different objects taken r at a time is

$$C(n,r) = \frac{P(n,r)}{r!} = \frac{n!}{r!(n-r)!}$$

Algebraic Equations

Quadratic

If $ax^2 + bx + c = 0$, and $a \neq 0$, then roots are

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

Cubic

To solve $\mathbf{x}^2 + bx^2 + cx + d = 0$, let $x = y - b/3$. Then the *reduced cubic* is obtained:

$$y^3 + py + q = 0$$

where $p = c - (1/3)b^2$ and $q = d - (1/3)bc + (2/27)b^3$. Solutions of the original cubic are then in terms of the reduced curbic roots y_1, y_2, y_3:

$$x_1 = y_1 - (1/3)b \qquad x_2 = y_2 - (1/3)b \qquad x_3 = y_3 - (1/3)b$$

The three roots of the reduced cubic are

$$y_1 = (A)^{1/3} + (B)^{1/3}$$

$$y_2 = W(A)^{1/3} + W^2(B)^{1/3}$$

$$y_3 = W^2(A)^{1/3} + W(B)^{1/3}$$

where

$$A = -\tfrac{1}{2}q + \sqrt{(1/27)p^3 + \tfrac{1}{4}q^2}$$

$$B = -\tfrac{1}{2}q - \sqrt{(1\ 27)p^3 + \tfrac{1}{4}q^2}$$

$$W = \frac{-1 + i\sqrt{3}}{2}, \qquad W^2 = \frac{-1 - i\sqrt{3}}{2}$$

When $(1/27)p^3 + (1/4)q^2$ is negative, A is complex; in this case A should be expressed in trigonometric form: $A = r(\cos\theta + i \sin\theta)$ where θ is a first or second quadrant angle, as q is negative or positive. The three roots of the reduced cubic are

$$y_1 = 2(r)^{1/3}\cos(\theta/3)$$

$$y_2 = 2(r)^{1/3}\cos\left(\frac{\theta}{3} + 120°\right)$$

$$y_3 = 2(r)^{1/3}\cos\left(\frac{\theta}{3} + 240°\right)$$

Geometry

Figures 19.1.2 to 19.1.12 are a collection of common geometric figures. Area (A), volume (V), and other measurable features are indicated.

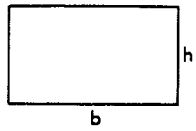


FIGURE 19.1.2 Rectangle. $A = bh$.

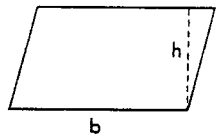


FIGURE 19.1.3 Parallelogram. $A = bh$.

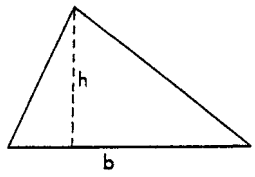


FIGURE 19.1.4 Triangle. $A = 1/2\ bh$.

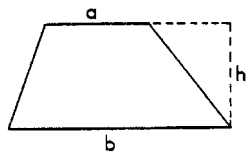


FIGURE 19.1.5 Trapezoid. $A = 1/2\ (a + b)h$.

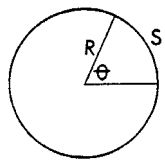


FIGURE 19.1.6 Circle. $A = \pi R^2$; circumference $= 2\pi R$, arc length $S = R\ \theta$ (θ in radians).

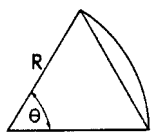


FIGURE 19.1.7 Sector of circle. $A_{sector} = 1/2R^2\theta$; $A_{segment} = 1/2R^2\ (\theta - \sin\theta)$.

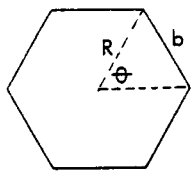


FIGURE 19.1.8 Regular polygon of n sides. $A = (n/4)b^2\ \mathrm{ctn}(\pi/n)$; $R = (b/2)\ \csc(\pi/n)$.

Table of Derivatives

In the following table, a and n are constants, e is the base of the natural logarithms, and u and v denote functions of x.

Additional Relations with Derivatives

$$\frac{d}{dt}\int_a^t f(x)\,dx = f(t) \qquad \frac{d}{dt}\int_t^a f(x)\,dx = -f(t)$$

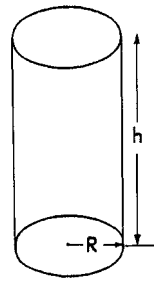


FIGURE 19.1.9 Right circular cylinder. $V = \pi R^2 h$; lateral surface area = $2\pi Rh$.

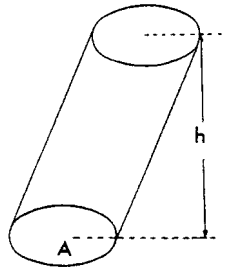


FIGURE 19.1.10 Cylinder (or prism) with parallel bases. $V = Ah$.

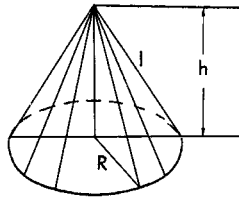


FIGURE 19.1.11 Right circular cone. $V = 1/3\ \pi R^2 h$; lateral surface area = $\pi Rl = \pi R\sqrt{R^2 + h^2}$.

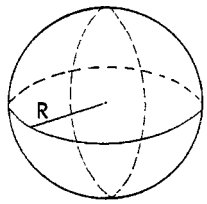


FIGURE 19.1.12 Sphere $V = 4/3\ \pi R^3$; surface area = $4\pi R^2$.

If $x = f(y)$, then $\dfrac{dy}{dx} = \dfrac{1}{dx/dy}$

If $y = f(u)$ and $u = g(x)$, then $\dfrac{dy}{dx} = \dfrac{dy}{du} \cdot \dfrac{du}{dx}$ (chain rule)

If $x = f(t)$ and $y = g(t)$, then $\dfrac{dy}{dx} = \dfrac{g'(t)}{f'(t)}$, and $\dfrac{d^2y}{dx^2} = \dfrac{f'(t)g''(t) - g'(t)f''(t)}{\left[f'(t)\right]^3}$

(Note: Exponent in denominator is 3.)

1. $\frac{d}{dx}(a) = 0$
2. $\frac{d}{dx}(x) = 1$
3. $\frac{d}{dx}(au) = a\frac{du}{dx}$
4. $\frac{d}{dx}(u+v) = \frac{du}{dx} + \frac{dv}{dx}$
5. $\frac{d}{dx}(uv) = u\frac{dv}{dx} + v\frac{du}{dx}$
6. $\frac{d}{dx}(u/v) = \frac{v\frac{du}{dx} - u\frac{dv}{dx}}{v^2}$
7. $\frac{d}{dx}(u^n) = nu^{n-1}\frac{du}{dx}$
8. $\frac{d}{dx}e^u = e^u\frac{du}{dx}$
9. $\frac{d}{dx}a^u = (\log_e a)a^u\frac{du}{dx}$
10. $\frac{d}{dx}\log_e u = (1/u)\frac{du}{dx}$
11. $\frac{d}{dx}\log_a u = (\log_a e)(1/u)\frac{du}{dx}$
12. $\frac{d}{dx}u^v = vu^{v-1}\frac{du}{dx} + u^v(\log_e u)\frac{dv}{dx}$
13. $\frac{d}{dx}\sin u = \cos u\frac{du}{dx}$
14. $\frac{d}{dx}\cos u = -\sin u\frac{du}{dx}$
15. $\frac{d}{dx}\tan u = \sec^2 u\frac{du}{dx}$
16. $\frac{d}{dx}\operatorname{ctn} u = -\csc^2 u\frac{du}{dx}$
17. $\frac{d}{dx}\sec u = \sec u\tan u\frac{du}{dx}$
18. $\frac{d}{dx}\csc u = -\csc u\operatorname{ctn} u\frac{du}{dx}$
19. $\frac{d}{dx}\sin^{-1} u = \frac{1}{\sqrt{1-u^2}}\frac{du}{dx}$, $\left(-\frac{1}{2}\pi \le \sin^{-1} u \le \frac{1}{2}\pi\right)$
20. $\frac{d}{dx}\cos^{-1} u = \frac{-1}{\sqrt{1-u^2}}\frac{du}{dx}$, $\left(0 \le \cos^{-1} u \le \pi\right)$
21. $\frac{d}{dx}\tan^{-1} u = \frac{1}{1+u^2}\frac{du}{dx}$
22. $\frac{d}{dx}\operatorname{ctn}^{-1} u = \frac{-1}{1+u^2}\frac{du}{dx}$
23. $\frac{d}{dx}\sec^{-1} u = \frac{1}{u\sqrt{u^2-1}}\frac{du}{dx}$,
$\left(-\pi \le \sec^{-1} u < -\frac{1}{2}\pi;\quad 0 \le \sec^{-1} u \le \frac{1}{2}\pi\right)$
24. $\frac{d}{dx}\csc^{-1} u = \frac{-1}{u\sqrt{u^2-1}}\frac{du}{dx}$,
$\left(-\pi < \csc^{-1} u \le -\frac{1}{2}\pi;\quad 0 < \csc^{-1} u \le \frac{1}{2}\pi\right)$
25. $\frac{d}{dx}\sinh u = \cosh u\frac{du}{dx}$
26. $\frac{d}{dx}\cosh u = \sinh u\frac{du}{dx}$
27. $\frac{d}{dx}\tanh u = \operatorname{sech}^2 u\frac{du}{dx}$
28. $\frac{d}{dx}\operatorname{ctnh} u = -\operatorname{csch}^2 u\frac{du}{dx}$
29. $\frac{d}{dx}\operatorname{sech} u = -\operatorname{sech} u\tanh u\frac{du}{dx}$
30. $\frac{d}{dx}\operatorname{csch} u = -\operatorname{csch} u\operatorname{ctnh} u\frac{du}{dx}$
31. $\frac{d}{dx}\sin^{-1} u = \frac{1}{\sqrt{u^2+1}}\frac{du}{dx}$
32. $\frac{d}{dx}\cosh^{-1} u = \frac{1}{\sqrt{u^2-1}}\frac{du}{dx}$
33. $\frac{d}{dx}\tanh^{-1} u = \frac{1}{1-u^2}\frac{du}{dx}$
34. $\frac{d}{dx}\operatorname{ctnh}^{-1} u = \frac{-1}{u^2-1}\frac{du}{dx}$
35. $\frac{d}{dx}\operatorname{sech}^{-1} u = \frac{-1}{u\sqrt{1-u^2}}\frac{du}{dx}$
36. $\frac{d}{dx}\operatorname{csch}^{-1} u = \frac{-1}{u\sqrt{u^2+1}}\frac{du}{dx}$

Integrals

Elementary Forms (Add an arbitrary constant to each integral)

1. $\int a\,dx = ax$
2. $\int a \cdot f(x)\,dx = a\int f(x)\,dx$
3. $\int \phi(y)\,dx = \int \frac{\phi(y)}{y'}\,dy, \quad \text{where } y' = \frac{dy}{dx}$
4. $\int (u+v)\,dx = \int u\,dx + \int v\,dx, \quad \text{where } u \text{ and } v \text{ are any functions of } x$
5. $\int u\,dv = u\int dv - \int v\,du = uv - \int v\,du$
6. $\int u\frac{dv}{dx}\,dx = uv - \int v\frac{du}{dx}\,dx$
7. $\int x^n\,dx = \frac{x^{n+1}}{n+1}, \quad \text{except } n = -1$
8. $\int \frac{f'(x)\,dx}{f(x)} = \log f(x), \quad \left[df(x) = f'(x)\,dx\right]$
9. $\int \frac{dx}{x} = \log x$
10. $\int \frac{f'(x)\,dx}{2\sqrt{f(x)}} = \sqrt{f(x)}, \quad \left[df(x) = f'(x)\,dx\right]$
11. $\int e^x\,dx = e^x$
12. $\int e^{ax}\,dx = e^{ax}/a$
13. $\int b^{ax}\,dx = \frac{b^{ax}}{a\log b}, \quad (b > 0)$
14. $\int \log x\,dx = x\log x - x$
15. $\int a^x \log a\,dx = a^x, \quad (a > 0)$
16. $\int \frac{dx}{a^2 + x^2} = \frac{1}{a}\tan^{-1}\frac{x}{a}$
17. $\int \frac{dx}{a^2 - x^2} = \begin{cases} \frac{1}{a}\tan^{-1}\frac{x}{a} \\ \quad \text{or} \\ \frac{1}{2a}\log\frac{a+x}{a-x}, \quad \left(a^2 > x^2\right) \end{cases}$
18. $\int \frac{dx}{x^2 - a^2} = \begin{cases} -\frac{1}{a}\operatorname{ctnh}^{-1}\frac{x}{a} \\ \quad \text{or} \\ \frac{1}{2a}\log\frac{x-a}{x+a}, \quad \left(x^2 > a^2\right) \end{cases}$

19. $$\int \frac{dx}{\sqrt{a^2 - x^2}} = \begin{cases} \sin^{-1}\dfrac{x}{|a|} \\ \text{or} \\ -\cos^{-1}\dfrac{x}{|a|}, \quad (a^2 > x^2) \end{cases}$$

20. $$\int \frac{dx}{\sqrt{x^2 \pm a^2}} = \log\left(x + \sqrt{x^2 \pm a^2}\right)$$

21. $$\int \frac{dx}{x\sqrt{x^2 - a^2}} = \frac{1}{|a|}\sec^{-1}\frac{x}{|a|}$$

22. $$\int \frac{dx}{x\sqrt{a^2 \pm x^2}} = -\frac{1}{a}\log\left(\frac{a + \sqrt{a^2 \pm x^2}}{x}\right)$$

Forms Containing $(a + bx)$

For forms containing $a + bx$, but not listed in the table, the substitution $u = (a + bx)\,x$ may prove helpful.

23. $$\int (a+bx)^n\,dx = \frac{(a+bx)^{n+1}}{(n+1)b}, \quad (n \neq -1)$$

24. $$\int x(a+bx)^n\,dx = \frac{1}{b^2(n+2)}(a+bx)^{n+2} - \frac{a}{b^2(n+1)}(a+bx)^{n+1}, \quad (n \neq -1, -2)$$

25. $$\int x^2(a+bx)^n\,dx = \frac{1}{b^3}\left[\frac{(a+bx)^{n+3}}{n+3} - 2a\frac{(a+bx)^{n+2}}{n+2} + a^2\frac{(a+bx)^{n+1}}{n+1}\right]$$

26. $$\int x^m(a+bx)^n\,dx = \begin{cases} \dfrac{x^{m+1}(a+bx)^n}{m+n+1} + \dfrac{an}{m+n+1}\displaystyle\int x^m(a+bx)^{n-1}\,dx \\ \text{or} \\ \dfrac{1}{a(n+1)}\left[-x^{m+1}(a+bx)^{n+1} + (m+n+2)\displaystyle\int x^m(a+bx)^{n+1}\,dx\right] \\ \text{or} \\ \dfrac{1}{b(m+n+1)}\left[x^m(a+bx)^{n+1} - ma\displaystyle\int x^{m-1}(a+bx)^n\,dx\right] \end{cases}$$

27. $$\int \frac{dx}{a+bx} = \frac{1}{b}\log(a+bx)$$

28. $$\int \frac{dx}{(a+bx)^2} = -\frac{1}{b(a+bx)}$$

29. $$\int \frac{dx}{(a+bx)^3} = -\frac{1}{2b(a+bx)^2}$$

30. $$\int \frac{x\,dx}{a+bx} = \begin{cases} \dfrac{1}{b^2}\left[a + bx - a\log(a+bx)\right] \\ \text{or} \\ \dfrac{x}{b} - \dfrac{a}{b^2}\log(a+bx) \end{cases}$$

31. $$\int \frac{x\,dx}{(a+bx)^2} = \frac{1}{b^2}\left[\log(a+bx) + \frac{a}{a+bx}\right]$$

32. $\displaystyle\int \frac{x\,dx}{(a+bx)^n} = \frac{1}{b^2}\left[\frac{-1}{(n-2)(a+bx)^{n-2}} + \frac{a}{(n-1)(a+bx)^{n-1}}\right], \quad n \neq 1, 2$

33. $\displaystyle\int \frac{x^2\,dx}{a+bx} = \frac{1}{b^3}\left[\frac{1}{2}(a+bx)^2 - 2a(a+bx) + a^2\log(a+bx)\right]$

34. $\displaystyle\int \frac{x^2\,dx}{(a+bx)^2} = \frac{1}{b^3}\left[a+bx-2a\log(a+bx) - \frac{a^2}{a+bx}\right]$

35. $\displaystyle\int \frac{x^2\,dx}{(a+bx)^3} = \frac{1}{b^3}\left[\log(a+bx) + \frac{2a}{a+bx} - \frac{a^2}{2(a+bx)^2}\right]$

36. $\displaystyle\int \frac{x^2\,dx}{(a+bx)^n} = \frac{1}{b^3}\left[\frac{-1}{(n-3)(a+bx)^{n-3}} + \frac{2a}{(n-2)(a+bx)^{n-2}} - \frac{a}{(n-1)(a+bx)^{n-1}}\right], \quad n \neq 1, 2, 3$

37. $\displaystyle\int \frac{dx}{x(a+bx)} = -\frac{1}{a}\log\frac{a+bx}{x}$

38. $\displaystyle\int \frac{dx}{x(a+bx)^2} = \frac{1}{a(a+bx)} - \frac{1}{a^2}\log\frac{a+bx}{x}$

39. $\displaystyle\int \frac{dx}{x(a+bx)^3} = \frac{1}{a^3}\left[\frac{1}{2}\left(\frac{2a+bx}{a+bx}\right)^2 + \log\frac{x}{a+bx}\right]$

40. $\displaystyle\int \frac{dx}{x^2(a+bx)} = -\frac{1}{ax} + \frac{b}{a^2}\log\frac{a+bx}{x}$

41. $\displaystyle\int \frac{dx}{x^3(a+bx)} = \frac{2bx-a}{2a^2x^2} + \frac{b^2}{a^3}\log\frac{x}{a+bx}$

42. $\displaystyle\int \frac{dx}{x^2(a+bx)^2} = -\frac{a+2bx}{a^2x(a+bx)} + \frac{2b^2}{a^3}\log\frac{a+bx}{x}$

The Fourier Transforms

For a piecewise continuous function $F(x)$ over a finite interval $0 \le x \le \pi$, the *finite Fourier cosine transform* of $F(x)$ is

$$f_c(n) = \int_0^\pi F(x)\cos nx\,dx \quad (n = 0,\ 1,\ 2,\ \ldots) \tag{19.1.1}$$

If x ranges over the interval $0 \le x \le L$, the substitution $x' = \pi x/L$ allows the use of this definition also. The inverse transform is written

$$\overline{F}(x) = \frac{1}{\pi}f_c(0) + \frac{2}{\pi}\sum_{n=1}^{\infty} f_c(n)\cos nx \quad (0 < x < \pi) \tag{19.1.2}$$

where $\overline{F}(x) = [F(x+0) + F(x-0)]/2$. We observe that $\overline{F}(x) = F(x)$ at points of continuity. The formula

$$\begin{aligned} f_c^{(2)}(n) &= \int_0^\pi F''(x)\cos nx\,dx \\ &= -n^2 f_c(n) - F'(0) + (-1)^n F'(\pi) \end{aligned} \tag{19.1.3}$$

makes the finite Fourier cosine transform useful in certain boundary value problems.

Analogously, the *finite Fourier sine transform* of $F(x)$ is

$$f_s(n) = \int_0^{\pi} F(x)\sin nx\, dx \quad (n = 1,\ 2,\ 3,\ \ldots) \tag{19.1.4}$$

and

$$\bar{F}(x) = \frac{2}{\pi}\sum_{n=1}^{\infty} f_s(n)\sin nx \quad (0 < x < \pi) \tag{19.1.5}$$

Corresponding to Equation (19.1.6), we have

$$\begin{aligned} f_s^{(2)}(n) &= \int_0^{\pi} F''(x)\sin nx\, dx \\ &= -n^2 f_s(n) - nF(0) - n(-1)^n F(\pi) \end{aligned} \tag{19.1.6}$$

Fourier Transforms

If $F(x)$ is defined for $x \geq 0$ and is piecewise continuous over any finite interval, and if

$$\int_0^{\infty} F(x)\, dx$$

is absolutely convergent, then

$$f_c(\alpha) = \sqrt{\frac{2}{\pi}}\int_0^{\infty} F(x)\cos(\alpha x)\, dx \tag{19.1.7}$$

is the *Fourier cosine transform* of $F(x)$. Furthermore,

$$\bar{F}(x) = \sqrt{\frac{2}{\pi}}\int_0^{\infty} f_c(\alpha)\cos(\alpha x)\, d\alpha \tag{19.1.8}$$

If $\lim_{x\to\infty} d^nF/dx^n = 0$, an important property of the Fourier cosine transform,

$$\begin{aligned} f_c^{(2r)}(\alpha) &= \sqrt{\frac{2}{\pi}}\int_0^{\infty}\left(\frac{d^{2r}F}{dx^{2r}}\right)\cos(\alpha x)\, dx \\ &= -\sqrt{\frac{2}{\pi}}\sum_{n=0}^{r-1}(-1)^n a_{2r-2n-1}\alpha^{2n} + (-1)^r\alpha^{2r} f_c(\alpha) \end{aligned} \tag{19.1.9}$$

where $\lim_{x\to 0} d^rF/dx^r = a_r$, makes it useful in the solution of many problems.

Under the same conditions,

$$f_s(\alpha) = \sqrt{\frac{2}{\pi}} \int_0^\infty F(x) \sin(\alpha x)\, dx \tag{19.1.10}$$

defines the *Fourier sine transform* of *F(x)*, and

$$\bar{F}(x) = \sqrt{\frac{2}{\pi}} \int_0^\infty f_s(\alpha) \sin(\alpha x)\, d\alpha \tag{19.1.11}$$

Corresponding to Equation (19.1.9) we have

$$\begin{aligned} f_s^{(2r)}(\alpha) &= \sqrt{\frac{2}{\pi}} \int_0^\infty \frac{d^{2r}F}{dx^{2r}} \sin(\alpha x)\, dx \\ &= -\sqrt{\frac{2}{\pi}} \sum_{n=1}^{r} (-1)^n \alpha^{2n-1} a_{2r-2n} + (-1)^{r-1} \alpha^{2r} f_s(\alpha) \end{aligned} \tag{19.1.12}$$

Similarly, if *F(x)* is defined for $-\infty < x < \infty$, and if $\int_{-\infty}^{\infty} F(x)\, dx$ is absolutely convergent, then

$$f(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} F(x) e^{i\alpha x}\, dx \tag{19.1.13}$$

is the *Fourier transform* of *F(x)*, and

$$\bar{F}(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(\alpha) e^{-i\alpha x}\, d\alpha \tag{19.1.14}$$

Also, if

$$\lim_{|x| \to \infty} \left| \frac{d^n F}{dx^n} \right| = 0 \quad (n = 1,\ 2,\ \ldots,\ r-1)$$

then

$$f^{(r)}(\alpha) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} F^{(r)}(x) e^{i\alpha x}\, dx = (-i\alpha)^r f(\alpha) \tag{19.1.15}$$

Finite Sine Transforms

	$f_s(n)$	$F(x)$
1.	$f_s(n) = \int_0^{\pi} F(x)\sin nx\, dx \quad (n = 1,\ 2,\ \ldots)$	$F(x)$
2.	$(-1)^{n+1} f_s(n)$	$F(\pi - x)$
3.	$\dfrac{1}{n}$	$\dfrac{\pi - x}{\pi}$
4.	$\dfrac{(-1)^{n+1}}{n}$	$\dfrac{x}{\pi}$
5.	$\dfrac{1-(-1)^n}{n}$	1
6.	$\dfrac{2}{n^2}\sin\dfrac{n\pi}{2}$	$\begin{cases} x & \text{when } 0 < x < \pi/2 \\ \pi - x & \text{when } \pi/2 < x < \pi \end{cases}$
7.	$\dfrac{(-1)^{n+1}}{n^3}$	$\dfrac{x(\pi^2 - x^2)}{6\pi}$
8.	$\dfrac{1-(-1)^n}{n^3}$	$\dfrac{x(\pi - x)}{2}$
9.	$\dfrac{\pi^2(-1)^{n-1}}{n} - \dfrac{2[1-(-1)^n]}{n^3}$	x^2
10.	$\pi(-1)^n\left(\dfrac{6}{n^3} - \dfrac{\pi^2}{n}\right)$	x^3
11.	$\dfrac{n}{n^2 + c^2}[1-(-1)^n e^{c\pi}]$	e^{cx}
12.	$\dfrac{n}{n^2 + c^2}$	$\dfrac{\sinh c(\pi - x)}{\sinh c\pi}$
13.	$\dfrac{n}{n^2 - k^2} \quad (k \neq 0,\ 1,\ 2,\ \ldots)$	$\dfrac{\sinh k(\pi - x)}{\sinh k\pi}$
14.	$\begin{cases} \dfrac{\pi}{2} & \text{when } n = m \\ 0 & \text{when } n \neq m \end{cases} \quad (m = 1,\ 2,\ \ldots)$	$\sin mx$
15.	$\dfrac{n}{n^2 - k^2}[1-(-1)^n \cos k\pi] \quad (k \neq 1,\ 2,\ \ldots)$	$\cos kx$
16.	$\begin{cases} \dfrac{n}{n^2 - m^2}[1-(-1)^{n+m}] & \text{when } n \neq m = 1,\ 2,\ \ldots \\ 0 & \text{when } n = m \end{cases}$	$\cos mx$
17.	$\dfrac{n}{(n^2 - k^2)^2} \quad (k \neq 0,\ 1,\ 2,\ \ldots)$	$\dfrac{\pi \sin kx}{2k\sin^2 kx} - \dfrac{x\cos k(\pi - x)}{2k\sin k\pi}$
18.	$\dfrac{b^n}{n} \quad (\lvert b\rvert \leq 1)$	$\dfrac{2}{\pi}\arctan\dfrac{b\sin x}{1 - b\cos x}$
19.	$\dfrac{1-(-1)^n}{n} b^n \quad (\lvert b\rvert \leq 1)$	$\dfrac{2}{\pi}\arctan\dfrac{2b\sin x}{1 - b^2}$

Finite Cosine Transforms

	$f_c(n)$	$F(x)$
1.	$f_c(n)=\int_0^{\pi} F(x)\cos nx\,dx \quad (n=0,\ 1,\ 2,\ \ldots)$	$F(x)$
2.	$(-1)^n f_c(n)$	$F(\pi - x)$
3.	0 when $n=1,\ 2,\ \ldots;\quad f_c(0)=\pi$	1
4.	$\frac{2}{n}\sin\frac{n\pi}{2};\quad f_c(0)=0$	$\begin{cases}1 & \text{when } 0<x<\pi/2\\ -1 & \text{when } \pi/2<x<\pi\end{cases}$
5.	$-\frac{1-(-1)^n}{n^2};\quad f_c(0)=\frac{\pi^2}{2}$	x
6.	$\frac{(-1)^n}{n^2};\quad f_c(0)=\frac{\pi^2}{6}$	$\frac{x^2}{2\pi}$
7.	$\frac{1}{n^2};\quad f_c(0)=0$	$\frac{(\pi-x)^2}{2\pi}-\frac{\pi}{6}$
8.	$3\pi^2\frac{(-1)^n}{n^2}-6\frac{1-(-1)^n}{n^4};\quad f_c(0)=\frac{\pi^4}{4}$	x^3
9.	$\frac{(-1)^n e^c\pi - 1}{n^2+c^2}$	$\frac{1}{c}e^{cx}$
10.	$\frac{1}{n^2+c^2}$	$\frac{\cosh c(\pi-x)}{c\sinh c\pi}$
11.	$\frac{k}{n^2-k^2}\left[(-1)^n\cos\pi k-1\right] \quad (k\neq 0,\ 1,\ 2,\ \ldots)$	$\sin kx$
12.	$\frac{(-1)^{n+m}-1}{n^2-m^2};\quad f_c(m)=0 \quad (m=1,\ 2,\ \ldots)$	$\frac{1}{m}\sin mx$
13.	$\frac{1}{n^2-k^2} \quad (k\neq 0,\ 1,\ 2,\ \ldots)$	$-\frac{\cos k(\pi-x)}{k\sin k\pi}$
14.	0 when $n=1,\ 2,\ \ldots;\quad f_c(m)=\frac{\pi}{2} \quad (m=1,\ 2,\ \ldots)$	$\cos mx$

Fourier Sine Transforms

	F(x)	$f_s(\alpha)$
1.	$\begin{cases}1 & (0<x<a)\\ 0 & (x>a)\end{cases}$	$\sqrt{\frac{2}{\pi}}\left[\frac{1-\cos\alpha}{\alpha}\right]$
2.	$x^{p-1} \quad (0<p<1)$	$\sqrt{\frac{2}{\pi}}\frac{\Gamma(p)}{\alpha^p}\sin\frac{p\pi}{2}$
3.	$\begin{cases}\sin x & (0<x<a)\\ 0 & (x>a)\end{cases}$	$\frac{1}{\sqrt{2\pi}}\left[\frac{\sin[a(1-\alpha)]}{1-\alpha}-\frac{\sin[a(1+\alpha)]}{1+\alpha}\right]$
4.	e^{-x}	$\sqrt{\frac{2}{\pi}}\left[\frac{\alpha}{1+\alpha^2}\right]$
5.	$xe^{-x^2/2}$	$\alpha e^{-\alpha^2/2}$

F(x)	$f_s(\alpha)$
6. $\cos\frac{x^2}{2}$	$\sqrt{2}\left[\sin\frac{\alpha^2}{2}C\left(\frac{\alpha^2}{2}\right)-\cos\frac{\alpha^2}{2}S\left(\frac{\alpha^2}{2}\right)\right]^*$
7. $\sin\frac{x^2}{2}$	$\sqrt{2}\left[\cos\frac{\alpha^2}{2}C\left(\frac{\alpha^2}{2}\right)+\sin\frac{\alpha^2}{2}S\left(\frac{\alpha^2}{2}\right)\right]^*$

* $C(y)$ and $S(y)$ are the Fresnel integrals

$$C(y)=\frac{1}{\sqrt{2\pi}}\int_0^y\frac{1}{\sqrt{t}}\cos t\,dt$$

$$S(y)=\frac{1}{\sqrt{2\pi}}\int_0^y\frac{1}{\sqrt{t}}\sin t\,dt$$

Fourier Cosine Transforms

F(x)	$f_c(\alpha)$
1. $\begin{cases}1 & (0<x<a)\\ 0 & (x>a)\end{cases}$	$\sqrt{\frac{2}{\pi}}\frac{\sin a\alpha}{\alpha}$
2. $x^{p-1}\quad(0<p<1)$	$\sqrt{\frac{2}{\pi}}\frac{\Gamma(p)}{\alpha^p}\cos\frac{p\pi}{2}$
3. $\begin{cases}\cos x & (0<x<a)\\ 0 & (x>a)\end{cases}$	$\frac{1}{\sqrt{2\pi}}\left[\frac{\sin[a(1-\alpha)]}{1-\alpha}+\frac{\sin[a(1+\alpha)]}{1+\alpha}\right]$
4. e^{-x}	$\sqrt{\frac{2}{\pi}}\left(\frac{1}{1+\alpha^2}\right)$
5. $e^{-x^2/2}$	$e^{-\alpha^2/2}$
6. $\cos\frac{x^2}{2}$	$\cos\left(\frac{\alpha^2}{2}-\frac{\pi}{4}\right)$
7. $\sin\frac{x^2}{2}$	$\cos\left(\frac{\alpha^2}{2}-\frac{\pi}{4}\right)$

Fourier Transforms

	$F(x)$	$f(\alpha)$
1.	$\dfrac{\sin ax}{x}$	$\begin{cases}\sqrt{\dfrac{\pi}{2}} & \lvert\alpha\rvert < a \\ 0 & \lvert\alpha\rvert > a\end{cases}$
2.	$\begin{cases}e^{iwx} & (p < x < q) \\ 0 & (x < p, x > q)\end{cases}$	$\dfrac{i}{\sqrt{2\pi}}\dfrac{e^{ip(w+\alpha)} - e^{iq(w+\alpha)}}{(w+\alpha)}$
3.	$\begin{cases}e^{-cx+iwx} & (x > 0) \\ 0 & (x < 0)\end{cases}$ $(c > 0)$	$\dfrac{i}{\sqrt{2\pi}(w+\alpha+ic)}$
4.	e^{-px^2} $R(p) > 0$	$\dfrac{1}{\sqrt{2p}}e^{-\alpha^2/4p}$
5.	$\cos px^2$	$\dfrac{1}{\sqrt{2p}}\cos\left[\dfrac{\alpha^2}{4p} - \dfrac{\pi}{4}\right]$
6.	$\sin px^2$	$\dfrac{1}{\sqrt{2p}}\cos\left[\dfrac{\alpha^2}{4p} + \dfrac{\pi}{4}\right]$
7.	$\lvert x\rvert^{-p}$ $(0 < p < 1)$	$\sqrt{\dfrac{2}{\pi}}\dfrac{\Gamma(1-p)\sin\dfrac{p\pi}{2}}{\lvert\alpha\rvert^{(1-p)}}$
8.	$\dfrac{e^{-a\lvert x\rvert}}{\sqrt{\lvert x\rvert}}$	$\dfrac{\sqrt{\sqrt{(a^2+\alpha^2)}+a}}{\sqrt{a^2+\alpha^2}}$
9.	$\dfrac{\cosh ax}{\cosh \pi x}$ $(-\pi < a < \pi)$	$\sqrt{\dfrac{2}{\pi}}\dfrac{\cos\dfrac{a}{2}\cosh\dfrac{\alpha}{2}}{\cosh\alpha + \cos a}$
10.	$\dfrac{\sinh ax}{\sinh \pi x}$ $(-\pi < a < \pi)$	$\dfrac{1}{\sqrt{2\pi}}\dfrac{\sin a}{\cosh\alpha + \cos a}$
11.	$\dfrac{1}{\sqrt{a^2-x^2}}$ $(\lvert x\rvert < a)$ 0 $(\lvert x\rvert > a)$	$\sqrt{\dfrac{\pi}{2}}J_0(a\alpha)$
12.	$\dfrac{\sin\left[b\sqrt{a^2+x^2}\right]}{\sqrt{a^2+x^2}}$	$\begin{cases}0 & (\lvert\alpha\rvert > b) \\ \sqrt{\dfrac{\pi}{2}}J_0\left(\sqrt{b^2-\alpha^2}\right) & (\lvert\alpha\rvert < b)\end{cases}$
13.	$\begin{cases}P_n(x) & (\lvert x\rvert < 1) \\ 0 & (\lvert x\rvert > 1)\end{cases}$	$\dfrac{i^n}{\sqrt{\alpha}}j_{n+1/2}(\alpha)$
14.	$\begin{cases}\dfrac{\cos\left[b\sqrt{a^2-x^2}\right]}{\sqrt{a^2-x^2}} & (\lvert x\rvert < a) \\ 0 & (\lvert x\rvert > a)\end{cases}$	$\sqrt{\dfrac{\pi}{2}}J_0\left(a\sqrt{a^2+b^2}\right)$
15.	$\begin{cases}\dfrac{\cosh\left[b\sqrt{a^2-x^2}\right]}{\sqrt{a^2-x^2}} & (\lvert x\rvert < a) \\ 0 & (\lvert x\rvert > a)\end{cases}$	$\sqrt{\dfrac{\pi}{2}}J_0\left(a\sqrt{a^2-b^2}\right)$

The following functions appear among the entries of the tables on transforms.

Function	Definition	Name
Ei(x)	$\int_{-\infty}^{x} \frac{e^v}{v}\,dv$; or sometimes defined as $-\mathrm{Ei}(-x) = \int_{x}^{\infty} \frac{e^{-v}}{v}\,dv$	Exponential integral function
Si(x)	$\int_{0}^{x} \frac{\sin v}{v}\,dv$	Sine integral function
Ci(x)	$\int_{\infty}^{x} \frac{\cos v}{v}\,dv$; or sometimes defined as negative of this integral	Cosine integral function
erf(x)	$\frac{2}{\sqrt{\pi}} \int_{0}^{x} e^{-v^2}\,dv$	Error function
erfc(x)	$1 - erf(x) = \frac{2}{\sqrt{\pi}} \int_{x}^{\infty} e^{-v^2}\,dv$	Complementary function to error function
$L_n(x)$	$\frac{e^x}{n!} \frac{d^n}{dx^n}\left(x^n e^{-x}\right), \quad n = 0,\ 1,\ \ldots$	Laguerre polynomial of degree n

Bessel Functions

Bessel Functions of the First Kind, $J_n(x)$ (Also Called Simply *Bessel Functions*) (Figure 19.1.13)

Domain: [$x > 0$]
Recurrence relation:

$$J_{n+1}(x) = \frac{2n}{x} J_n(x) - J_{n-1}(x), \quad n = 0,\ 1,\ 2,\ \ldots$$

Symmetry: $J_{-n}(x) = (-1)^n J_n(x)$

0. $J_0(20x)$
1. $J_1(20x)$
2. $J_2(20x)$
3. $J_3(20x)$
4. $J_4(20x)$
5. $J_5(20x)$

Bessel Functions of the Second Kind, $Y_n(x)$ (Also Called *Neumann Functions* or *Weber Functions*) (Figure 19.1.14)

Domain: [$x > 0$]
Recurrence relation:

$$Y_{n+1}(x) = \frac{2n}{x} Y_n(x) - Y_{n-1}(x), \quad n = 0,\ 1,\ 2,\ \ldots$$

Symmetry: $Y_{-n}(x) = (-1)^n Y_n(x)$

0. $Y_0(20x)$
1. $Y_1(20x)$
2. $Y_2(20x)$
3. $Y_3(20x)$
4. $Y_4(20x)$
5. $Y_5(20x)$

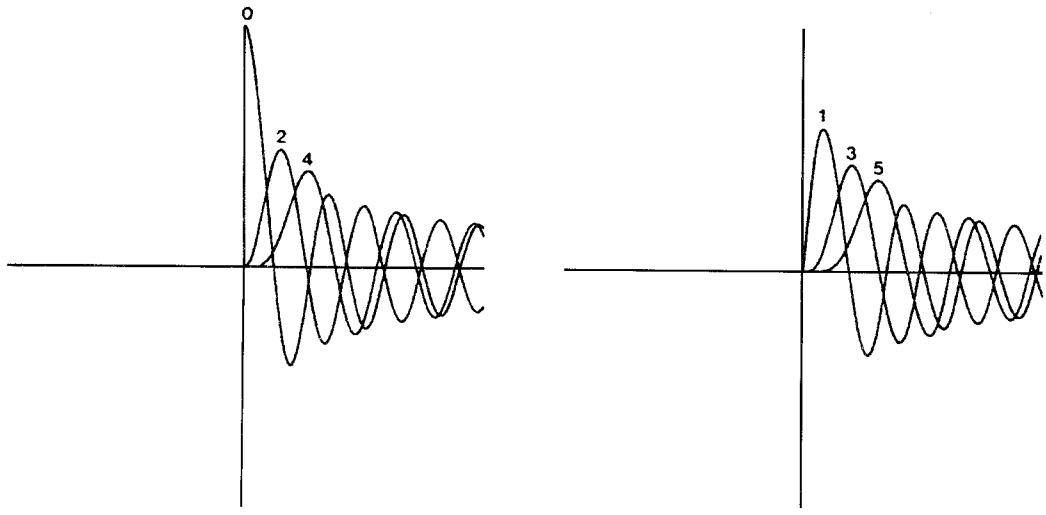


FIGURE 19.1.13 Bessel functions of the first kind.

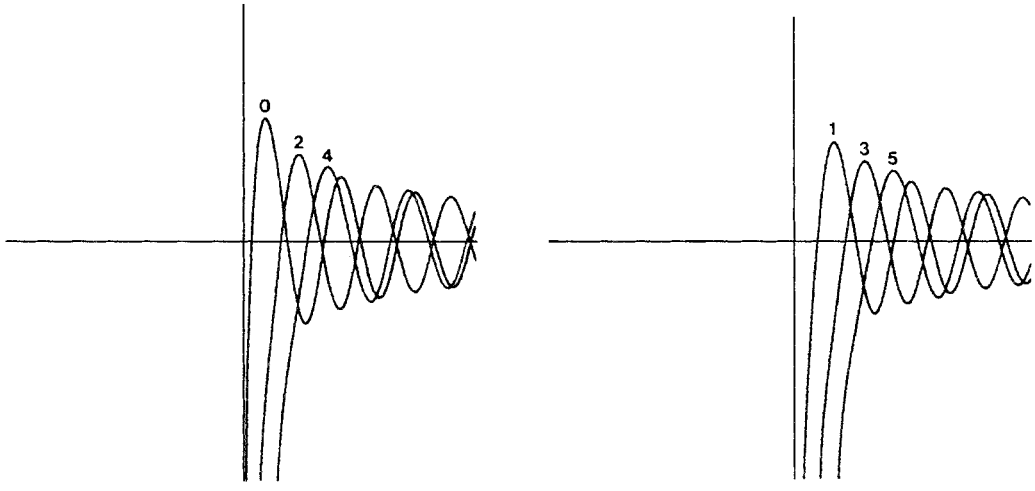


FIGURE 19.1.14 Bessel functions of the second kind.

Legendre Functions

Associated Legendre Functions of the First Kind, $P_n^m(x)$ (Figure 19.1.15)

Domain: $[-1 < x < 1]$
Recurrence relations:

$$P_{n+1}^m(x) = \frac{(2n+1)xP_n^m - (n+m)P_{n-1}^m(x)}{n-m+1}, \quad n = 1, 2, 3, \ldots$$

$$P_n^{m+1}(x) = \left(x^2 - 1\right)^{-1/2}\left[(n-m)xP_n^m(x) - (n+m)P_{n-1}^m(x)\right], \quad m = 0, 1, 2, \ldots$$

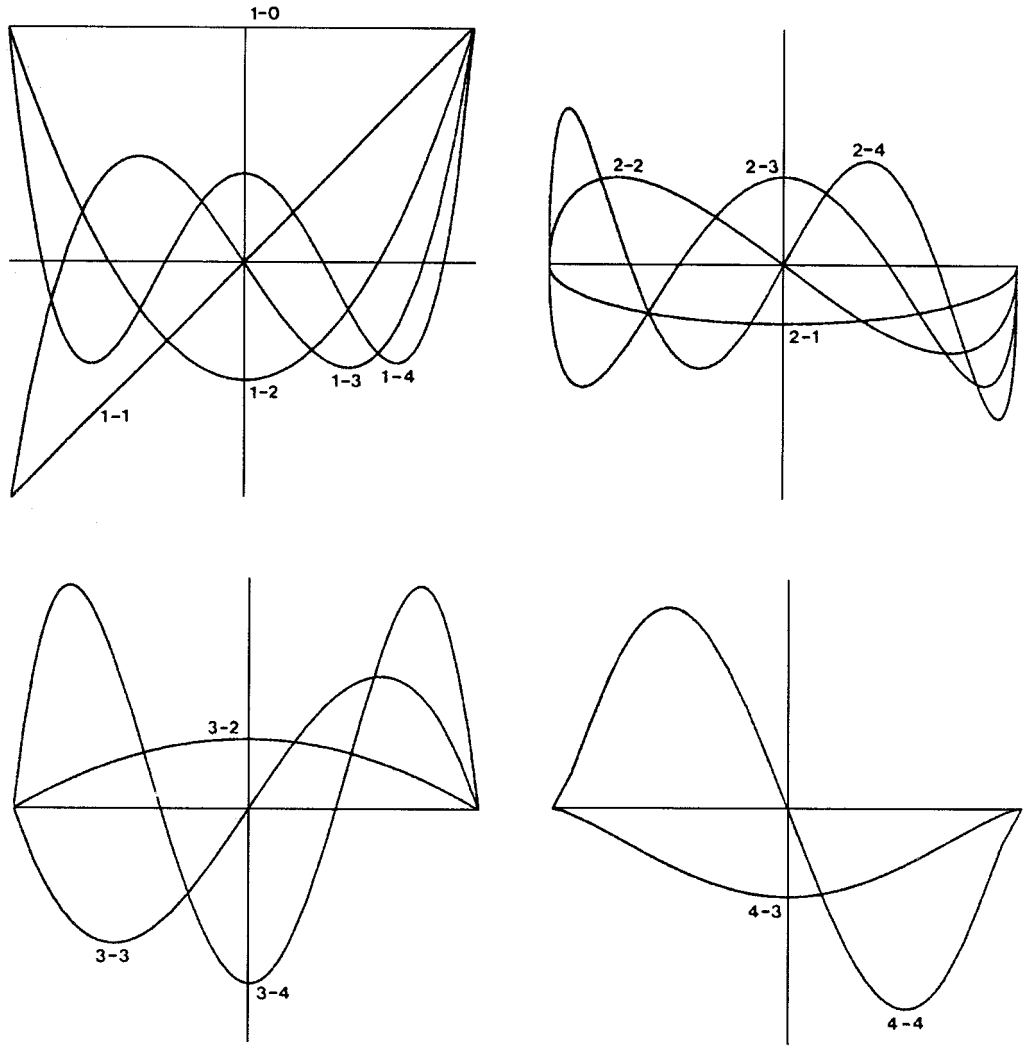


FIGURE 19.1.15 Legendre functions of the first kind.

with

$$P_0^0 = 1 \qquad P_1^0 = x$$

Special case: P_n^0 = Legendre polynomials

1-0. $P_0^0(x)$
1-1. $P_1^0(x)$ 2-1. 0.25 $P_1^1(x)$
1-2. $P_2^0(x)$ 2-2. 0.25 $P_2^1(x)$ 3-2. 0.10 $P_2^2(x)$
1-3. $P_3^0(x)$ 2-3. 0.25 $P_3^1(x)$ 3-3. 0.10 $P_3^2(x)$ 4-3. 0.025 $P_3^3(x)$
1-4. $P_4^0(x)$ 2-4. 0.25 $P_4^1(x)$ 3-4. 0.10 $P_4^2(x)$ 4-4. 0.025 $P_4^3(x)$

Table of Differential Equations

Equation	Solution
1. $y' = \dfrac{dy}{dx} = f(x)$	$y = \int f(x)\,dx + c$
2. $y' + p(x)y = q(x)$	$y = \exp[-\int p(x)\,dx]\{c + \int \exp[\int p(x)\,dx]q(x)dx\}$
3. $y' + p(x)y = q(x)y^{\alpha}$ $\alpha \neq 0,\ \alpha \neq 1$	Set $z = y^{1-\alpha} \rightarrow z' + (1-\alpha)p(x)z = (1-\alpha)q(x)$ and use 2
4. $y' = f(x)g(y)$	Integrate $\dfrac{dy}{g(y)} = f(x)\,dx$ (separable)
5. $\dfrac{dy}{dx} = f(x/y)$	Set $y = xu \rightarrow u + x\dfrac{du}{dx} = f(u)$
	$\displaystyle\int \frac{1}{f(u)-u}\,du = \ln\lvert x\rvert + c$
6. $y' = f\left(\dfrac{a_1x + b_1y + c_1}{a_2x + b_2y + c_2}\right)$	Set $x = X + \alpha, y = Y + \beta$ Choose $\begin{cases} a_1\alpha + b_1\beta = -c_2 \\ a_2\alpha + b_2\beta = -c_2 \end{cases} \rightarrow Y' = f\left(\dfrac{a_1X + b_1Y}{a_2X + b_2Y}\right)$ If $a_1b_2 - a_2b_1 \neq 0$, set $Y = Xu \rightarrow$ separable form $u + Xu' = f\left(\dfrac{a_1 + b_1u}{a_2 + b_2u}\right)$ If $a_1b_2 - a_2b_1 = 0$, set $u = a_1x + b_1y \rightarrow$ $\dfrac{du}{dx} = a_1 + b_1 f\left(\dfrac{u + c_1}{ku + c_2}\right)$ since $a_2x + b_2y = k(a_1x + a_2y)$
7. $y'' + a^2y = 0$	$y = c_1 \cos ax + c_2 \sin ax$
8. $y'' - a^2y = 0$	$y = c_1e^{ax} + c_2e^{-ax}$
9. $y'' + ay' + by = 0$	Set $y = e^{-(a/2)x}u \rightarrow u'' + \left(b - \dfrac{a^2}{4}\right)u = 0$
10. $y'' + a(x)y' + b(x)y = 0$	Set $y = e^{-(1/2)\int a(x)\,dx} \rightarrow u'' + \left[b(x) - \dfrac{a^2}{4} - \dfrac{a'}{2}\right]u = 0$
11. $x^2y'' + xy' + (x^2 - a^2)y = 0$ $a \geq 0$ (Bessel)	i. If a is not an integer $y = c_1J_a(x) + c_2J_{-a}(x)$ (Bessel functions of first kind) ii. If a is an integer (say, n) $y = c_1J_n(x) + c_2Y_n(x)$ (Y_n is Bessel function of second kind)
12. $(1 - x^2)y'' - 2xy' + a(a + 1)y = 0$ a is real (Legendre)	$y(x) = c_1p_a(x) + c_1q_a(x)$ (Legendre functions)
13. $y' + ay^2 = bx^n$ (integrable Riccati) a, b, n real	Set $u' = ayu \rightarrow u'' - abx^nu = 0$ and use 14
14. $y'' - ax^{-1}y' + b^2x^{\mu}y = 0$	$y = x^p[c_1J_\nu(kx^q) + c_2J_{-\nu}(kx^q)$ where $p = (a + 1)/2$, $\nu = (a + 1)/(\mu + 2)$, $k = 2b/(\mu + 2)$, $q = (\mu + 2)/2$
15. Item 13 shows that the Riccati equation is linearized by raising the order of the equation. The *Riccati chain*, which is linearizable by raising the order, is $u' = uy,\quad u'' = u[y^1 + y^2],\quad u''' = u[y'' + 3yy' + y^3],$ $u^{(iv)} = u[y''' + 4yy'' + 6y^2y' + 3(y')^2 + y^4],\ldots$ To use this consider the second-order equation $y'' + 3yy' + y^3 = f(x)$. The Riccati transformation $u' = yu$ transforms this equation to the linear for $u''' = uf(x)$!	

References

Kanke, E. 1956. *Differentialgleichungen Lösungsmethoden und Lösungen*, Vol. I. Akad. Verlagsges., Leipzig.

Murphy, G. M. 1960. *Ordinary Differential Equations and Their Solutions*, Van Nostrand, New York.

Zwillger, D. 1992. *Handbook of Differential Equations*, 2nd ed. Academic Press, San Diego.

19.2 Linear Algebra and Matrices

George Cain

Basic Definitions

A *Matrix* **A** is a rectangular array of numbers (real or complex)

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \vdots & & & \\ a_{n1} & a_{n2} & \cdots & a_{nm} \end{bmatrix}$$

The *size* of the matrix is said to be $n \times m$. The $1 \times m$ matrices $[a_{i1} \cdots a_{im}]$ are called rows of A, and the $n \times 1$ matrices

$$\begin{bmatrix} a_{1j} \\ a_{2j} \\ \vdots \\ a_{nj} \end{bmatrix}$$

are called *columns* of **A**. An $n \times m$ matrix thus consists of n rows and m columns; a_{ij} denotes the *element*, or *entry*, of **A** in the ith row and jth column. A matrix consisting of just one row is called a *row vector*, whereas a matrix of just one column is called a *column vector*. The elements of a vector are frequently called *components* of the vector. When the size of the matrix is clear from the context, we sometimes write $\mathbf{A} = (a_{ij})$.

A matrix with the same number of rows as columns is a *square* matrix, and the number of rows and columns is the *order* of the matrix. The diagonal of an $n \times n$ square matrix **A** from a_{11} to a_{nn} is called the *main*, or *principal*, *diagonal*. The word *diagonal* with no modifier usually means the main diagonal. The *transpose* of a matrix **A** is the matrix that results from interchanging the rows and columns of **A**. It is usually denoted by $\mathbf{A}^T$. A matrix **A** such that $\mathbf{A} = \mathbf{A}^T$ is said to be *symmetric*. The *conjugate transpose* of **A** is the matrix that results from replacing each element of $\mathbf{A}^T$ by its complex conjugate, and is usually denoted by $\mathbf{A}^H$. A matrix such that $\mathbf{A} = \mathbf{A}^H$ is said to be *Hermitian*.

A square matrix $\mathbf{A} = (a_{ij})$ is *lower triangular* if $a_{ij} = 0$ for $j > i$ and is *upper triangular* if $a_{ij} = 0$ for $j < i$. A matrix that is both upper and lower triangular is a *diagonal* matrix. The $n \times n$ *identity matrix* is the $n \times n$ diagonal matrix in which each element of the main diagonal is 1. It is traditionally denoted $\mathbf{I}_n$, or simply **I** when the order is clear from the context.

Algebra of Matrices

The sum and difference of two matrices **A** and **B** are defined whenever **A** and **B** have the same size. In that case $\mathbf{C} = \mathbf{A} \pm \mathbf{B}$ is defined by $\mathbf{C} = (c_{ij}) = (a_{ij} \pm b_{ij})$. The product $t\mathbf{A}$ of a scalar t (real or complex number) and a matrix **A** is defined by $t\mathbf{A} = (ta_{ij})$. If **A** is an $n \times m$ matrix and **B** is an $m \times p$ matrix, the product $\mathbf{C} = \mathbf{AB}$ is defined to be the $n \times p$ matrix $\mathbf{C} = (c_{ij})$ given by $c_{ij} = \sum_{k=1}^{m} a_{ik}b_{kj}$. Note that the product of an $n \times m$ matrix and an $m \times p$ matrix is an $n \times p$ matrix, and the product is defined only when the number of columns of the first factor is the same as the number of rows of the second factor. Matrix multiplication is, in general, associative: $\mathbf{A}(\mathbf{BC}) = (\mathbf{AB})\mathbf{C}$. It also distributes over addition (and subtraction):

$$\mathbf{A}(\mathbf{B}+\mathbf{C}) = \mathbf{AB}+\mathbf{AC} \quad \text{and} \quad (\mathbf{A}+\mathbf{B})\mathbf{C} = \mathbf{AC}+\mathbf{BC}$$

It is, however, not in general true that $\mathbf{AB} = \mathbf{BA}$, even in case both products are defined. It is clear that $(\mathbf{A} + \mathbf{B})^{\mathrm{T}} = \mathbf{A}^{\mathrm{T}} + \mathbf{B}^{\mathrm{T}}$ and $(\mathbf{A} + \mathbf{B})^{\mathrm{H}} = \mathbf{A}^{\mathrm{H}} + \mathbf{B}^{\mathrm{H}}$. It is also true, but not so obvious perhaps, that $(\mathbf{AB})^{\mathrm{T}} = \mathbf{B}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}$ and $(\mathbf{AB})^{\mathrm{H}} = \mathbf{B}^{\mathrm{H}}\mathbf{A}^{\mathrm{H}}$.

The $n \times n$ identity matrix **I** has the property that $\mathbf{IA} = \mathbf{AI} = \mathbf{A}$ for every $n \times n$ matrix **A**. If **A** is square, and if there is a matrix **B** such at $\mathbf{AB} = \mathbf{BA} = \mathbf{I}$, then **B** is called the *inverse* of **A** and is denoted $\mathbf{A}^{-1}$. This terminology and notation are justified by the fact that a matrix can have at most one inverse. A matrix having an inverse is said to be *invertible*, or *nonsingular,* while a matrix not having an inverse is said to be *noninvertible*, or *singular.* The product of two invertible matrices is invertible and, in fact, $(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$. The sum of two invertible matrices is, obviously, not necessarily invertible.

Systems of Equations

The system of n linear equations in m unknowns

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + \ldots + a_{1m}x_m &= b_1 \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_3 + \ldots + a_{2m}x_m &= b_2 \\ &\vdots \\ a_{n1}x_1 + a_{n2}x_2 + a_{n3}x_3 + \ldots + a_{nm}x_m &= b_n \end{aligned}$$

may be written $\mathbf{Ax} = \mathbf{b}$, where $\mathbf{A} = (a_{ij})$, $\mathrm{x} = [\mathrm{x}_1\ \mathrm{x}_2\ \cdots\ \mathrm{x}_\mathrm{m}]^{\mathrm{T}}$, and $\mathrm{b} = [\mathrm{b}_1\ \mathrm{b}_2\ \cdots\ \mathrm{b}_\mathrm{n}]^{\mathrm{T}}$. Thus A is an $n \times m$ matrix, and x and b are column vectors of the appropriate sizes.

The matrix **A** is called the *coefficient matrix* of the system. Let us first suppose the coefficient matrix is square; that is, there are an equal number of equations and unknowns. If **A** is upper triangular, it is quite easy to find all solutions of the system. The ith equation will contain only the unknowns x_i, x_{i+1}, …, x_n, and one simply solves the equations in reverse order: the last equation is solved for x_n; the result is substituted into the $(n - 1)$st equation, which is then solved for x_{n-1}; these values of x_n and x_{n-1} are substituted in the $(n - 2)$th equation, which is solved for x_{n-2}, and so on. This procedure is known as *back substitution.*

The strategy for solving an arbitrary system is to find an upper-triangular system equivalent with it and solve this upper-triangular system using back substitution. First suppose the element $a_{11} \neq 0$. We may rearrange the equations to ensure this, unless, of course the first column of **A** is all 0s. In this case proceed to the next step, to be described later. For each $i \geq 2$ let $m_{i1} = a_{i1}/a_{11}$. Now replace the ith equation by the result of multiplying the first equation by m_{i1} and subtracting the new equation from the ith equation. Thus,

$$a_{i1}x_1 + a_{i2}x_2 + a_{i3}x_3 + \ldots + a_{im}x_m = b_i$$

is replaced by

$$0 \cdot x_1 + \left(a_{i2} + m_{i1}a_{12}\right)x_2 + \left(a_{i3} + m_{i1}a_{13}\right)x_3 + \ldots + \left(a_{im} + m_{i1}a_{1m}\right)x_m = b_i + m_{i1}b_1$$

After this is done for all $i = 2, 3, \ldots, n$, there results the equivalent system

$$\begin{aligned}
a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + \ldots + a_{1n}x_n &= b_1 \\
0 \cdot x_1 + a'_{22}x_2 + a'_{23}x_3 + \ldots + a'_{2n}x_n &= b'_2 \\
0 \cdot x_1 + a'_{32}x_2 + a'_{33}x_3 + \ldots + a'_{3n}x_n &= b'_3 \\
&\vdots \\
0 \cdot x_1 + a'_{n2}x_2 + a'_{n3}x_3 + \ldots + a'_{nn}x_n &= b'_n
\end{aligned}$$

in which all entries in the first column below a_{11} are 0. (Note that if all entries in the first column were 0 to begin with, then $a_{11} = 0$ also.) This procedure is now repeated for the $(n - 1) \times (n - 1)$ system

$$\begin{aligned}
a'_{22}x_2 + a'_{23}x_3 + \ldots + a'_{2n}x_n &= b'_2 \\
a'_{32}x_2 + a'_{33}x_3 + \ldots + a'_{3n}x_n &= b'_3 \\
&\vdots \\
a'_{n2}x_2 + a'_{n3}x_3 + \ldots + a'_{nn}x_n &= b'_n
\end{aligned}$$

to obtain an equivalent system in which all entries of the coefficient matrix below a'_{22} are 0. Continuing, we obtain an upper-triangular system Ux = c equivalent with the original system. This procedure is known as *Gaussian elimination*. The number m_{ij} are known as the *multipliers*.

Essentially the same procedure may be used in case the coefficient matrix is not square. If the coefficient matrix is not square, we may make it square by appending either rows or columns of 0s as needed. Appending rows of 0s and appending 0s to make b have the appropriate size equivalent to appending equations 0 = 0 to the system. Clearly the new system has precisely the same solutions as the original system. Appending columns of 0s and adjusting the size of x appropriately yields a new system with additional unknowns, each appearing only with coefficient 0, thus not affecting the solutions of the original system. In either case we may assume the coefficient matrix is square, and apply the Gauss elimination procedure.

Suppose the matrix **A** is invertible. Then if there were no row interchanges in carrying out the above Gauss elimination procedure, we have the *LU factorization* of the matrix **A**:

$$\mathbf{A} = \mathbf{LU}$$

where U is the upper-triangular matrix produced by elimination and L is the lower-triangular matrix given by

$$\mathbf{L} = \begin{bmatrix} 1 & 0 & \ldots & \ldots & 0 \\ m_{21} & 1 & 0 & \ldots & 0 \\ \vdots & & \ddots & & \\ m_{n1} & m_{n2} & \ldots & & 1 \end{bmatrix}$$

A *permutation* P_{ij} matrix is an $n \times n$ matrix such that P_{ij} A is the matrix that results from exchanging row i and j of the matrix A. The matrix P_{ij} is the matrix that results from exchanging rows i and j of the identity matrix. A product P of such matrices P_{ij} is called a *permutation* matrix. If row interchanges are required in the Gauss elimination procedure, then we have the factorization

$$\mathbf{PA} = \mathbf{LU}$$

where P is the permutation matrix giving the required row exchanges.

Vector Spaces

The collection of all column vectors with n real components is *Euclidean n-space*, and is denoted R^n. The collection of column vectors with n complex components is denoted C^n. We shall use *vector space* to mean either R^n or C^n. In discussing the space R^n, the word *scalar* will mean a real number, and in discussing the space C^n, it will mean a complex number. A subset S of a vector space is a *subspace* such that if **u** and **v** are vectors in S, and if c is any scalar, then **u** + **v** and c**u** are in S. We shall sometimes use the word *space* to mean a subspace. If $B = \{v_1, v_2, \ldots, v_k\}$ is a collection of vectors in a vector space, then the set S consisting of all vectors $c_1v_1 + c_2v_2 + \cdots + c_mv_m$ for all scalars $c_1, c_2, \ldots, c_m$ is a subspace, called the *span* of B. A collection $\{v_1, v_2, \ldots, v_m\}$ of vectors $c_1v_1 + c_2v_2 + \cdots + c_mv_m$ is a *linear combination* of B. If S is a subspace and $B = \{v_1, v_2, \ldots, v_m\}$ is a subset of S such that S is the span of B, then B is said to *span* S.

A collection $\{v_1, v_2, \ldots, v_m\}$ of n-vectors is *linearly dependent* if there exist scalars $c_1, c_2, \ldots, c_m$, not all zero, such that $c_1v_1 + c_2v_2 + \cdots + c_mv_m = 0$. A collection of vectors that is not linearly dependent is said to be *linearly independent*. The modifier *linearly* is frequently omitted, and we speak simply of dependent and independent collections. A linearly independent collection of vectors in a space S that spans S is a *basis* of S. Every basis of a space S contains the same number of vectors; this number is the *dimension* of S. The dimension of the space consisting of only the zero vector is 0. The collection $B = \{e_1, e_2, \ldots, e_n\}$, where $e_1 = [1,0,0,\ldots, 0]^T$, $e_2 = [0,1,0,\ldots, 0]^T$, and so forth ($e_i$ has 1 as its ith component and zero for all other components) is a basis for the spaces R^n and C^n. This is the *standard basis* for these spaces. The dimension of these spaces is thus n. In a space S of dimension n, no collection of fewer than n vectors can span S, and no collection of more than n vectors in S can be independent.

Rank and Nullity

The *column space* of an $n \times m$ matrix **A** is the subspace of R^n or C^n spanned by the columns of **A**. The *row space* is the subspace of R^m or C^m spanned by the rows or **A**. Note that for any vector $x = [x_1\, x_2, \cdots x_m]^T$,

$$\mathbf{Ax} = x_1\begin{bmatrix} a_{11} \\ a_{21} \\ \vdots \\ a_{n1} \end{bmatrix} + x_2\begin{bmatrix} a_{12} \\ a_{22} \\ \vdots \\ a_{n2} \end{bmatrix} + \ldots + x_m\begin{bmatrix} a_{1m} \\ a_{2m} \\ \vdots \\ a_{nm} \end{bmatrix}$$

so that the column space is the collection of all vectors, **Ax**, and thus the system **Ax** = **b** has a solution if and only *if* **b** is a member of the column space of **A**.

The dimension of the column space is the *rank* of **A**. The row space has the same dimension as the column space. The set of all solutions of the system **Ax** = 0 is a subspace called the *null space* of **A**, and the dimension of this null space is the *nullity* of **A**. A fundamental result in matrix theory is the fact that, for an $n \times m$ matrix **A**.

$$\text{rank } \mathbf{A} + \text{nullity } \mathbf{A} = m$$

The difference of any two solutions of the linear system $\mathbf{Ax} = \mathbf{b}$ is a member of the null space of **A**. Thus this system has at most one solution if and only if the nullity of **A** is zero. If the system is square (that is, if **A** is $n \times n$), then there will be a solution for every right-hand side **b** if and only if the collection of columns of **A** is linearly independent, which is the same as saying the rank of **A** is n. In this case the nullity must be zero. Thus, for any **b**, the square system $\mathbf{Ax} = \mathbf{b}$ has exactly one solution if and only if rank **A** = n. In other words the $n \times n$ matrix **A** is invertible if and only if rank **A** = n.

Orthogonality and Length

The *inner product* of two vectors x and y is the scalar $x^H y$. The *length*, or *norm*, $\|x\|$, of the vector **x** is given by $\|x\| = \sqrt{\mathbf{x}^H \mathbf{x}}$. A *unit vector* is a vector of norm 1. Two vectors x and y are *orthogonal* if $x^H y = 0$. A collection of vectors $\{v_1, v_2, \ldots, v_m\}$ in a space S is said to be an *orthonormal* collection if $\mathbf{v}_i^H \mathbf{v}_j = 0$ for $i \neq j$ and $\mathbf{v}_i^H \mathbf{v}_i = 1$. An orthonormal collection is necessarily linearly independent. If S is a subspace (of R^n or C^n) spanned by the orthonormal collection $\{v_1, v_2, \ldots, v_m\}$, then the *projection* of a vector x onto S is the vector

$$\text{proj}(\mathbf{x}; \mathbf{S}) = (\mathbf{x}^H \mathbf{v}_1)\mathbf{v}_1 + (\mathbf{x}^H \mathbf{v}_2)\mathbf{v}_2 + \ldots + (\mathbf{x}^H \mathbf{v}_m)\mathbf{v}_m$$

The projection of x onto S minimizes the function $f(y) = \|x - y\|^2$ for $y \in S$. In other words the projection of x onto S is the vector in S that is "closest" to x.

If **b** is a vector and **A** is an $n \times m$ matrix, then a vector x minimizes $\|b - Ax\|^2$ if only if it is a solution of $A^H Ax = A^H b$. This system of equations is called the *system of normal equations* for the least-squares problem of minimizing $\|b - Ax\|^2$.

If **A** is an $n \times m$ matrix, and rank **A** = k, then there is a $n \times k$ matrix Q whose columns form an orthonormal basis for the column space of **A** and a $k \times m$ upper-triangular matrix **R** of rank k such that

$$\mathbf{A} = \mathbf{QR}$$

This is called the *QR factorization* of **A**. It now follows that x minimizes $\|b - Ax\|^2$ if and only if it is a solution of the upper-triangular system $Rx = Q^H b$.

If $\{w_1, w_2, \ldots, w_m\}$ is a basis for a space S, the following procedure produces an orthonormal basis $\{v_1, v_2, \ldots, v_m\}$ for S.

Set $\mathbf{v}_1 = \mathbf{w}_1 / \|\mathbf{w}_1\|$.
Let $\tilde{\mathbf{v}}_2 = \mathbf{w}_2 - \text{proj}(\mathbf{w}_2; \mathbf{S}_1)$, where $\mathbf{S}_1$ is the span of $\{\mathbf{v}_1\}$; set $\mathbf{v}_2 = \tilde{\mathbf{v}}_2 / \|\tilde{\mathbf{v}}_2\|$.
Next, let $\tilde{\mathbf{v}}_3 = \mathbf{w}_3 - \text{proj}(\mathbf{w}_3; \mathbf{S}_2)$, where $\mathbf{S}_2$ is the span of $\{\mathbf{v}_1, \mathbf{v}_2\}$; set $\mathbf{v}_3 = \tilde{\mathbf{v}}_3 / \|\tilde{\mathbf{v}}_3\|$.

And, so on: $\tilde{\mathbf{v}}_i = w_i - \text{proj}(w_i; S_{i-1})$, where S_{i-1} is the span of $\{v_1, v_2, \ldots, v_{i-1}\}$; set $v_i = \tilde{\mathbf{v}}_i / \|\tilde{\mathbf{v}}_i\|$. This the *Gram-Schmidt procedure.*

If the collection of columns of a square matrix is an orthonormal collection, the matrix is called a *unitary matrix*. In case the matrix is a real matrix, it is usually called an *orthogonal matrix*. A unitary matrix U is invertible, and $\mathbf{U}^{-1} = \mathbf{U}^H$. (In the real case an orthogonal matrix **Q** is invertible, and $\mathbf{Q}^{-1} = \mathbf{Q}^T$.)

Determinants

The *determinant* of a square matrix is defined inductively. First, suppose the determinant $\det \mathbf{A}$ has been defined for all square matrices of order $< n$. Then

$$\det \mathbf{A} = a_{11}\mathbf{C}_{11} + a_{12}\mathbf{C}_{12} + \ldots + a_{1n}\mathbf{C}_{1n}$$

where the numbers C_{ij} are *cofactors* of the matrix **A**:

$$\mathbf{C}_{ij} = (-1)^{i+j} \det \mathbf{M}_{ij}$$

where M_{ij} is the $(n-1) \times (n-1)$ matrix obtained by deleting the ith row and jth column of **A**. Now det**A** is defined to be the only entry of a matrix of order 1. Thus, for a matrix of order 2, we have

$$\det \begin{bmatrix} a & b \\ c & d \end{bmatrix} = ad - bc$$

There are many interesting but not obvious properties of determinants. It is true that

$$\det \mathbf{A} = a_{i1}\mathbf{C}_{i1} + a_{i2}\mathbf{C}_{i2} + \ldots + a_{in}\mathbf{C}_{in}$$

for any $1 \le i \le n$. It is also true that det**A** = det$\mathbf{A}^{\mathrm{T}}$, so that we have

$$\det \mathbf{A} = a_{1j}\mathbf{C}_{1j} + a_{2j}\mathbf{C}_{2j} + \ldots + a_{nj}\mathbf{C}_{nj}$$

for any $1 \le j \le n$.

If **A** and **B** are matrices of the same order, then det**AB** = (det**A**)(det**B**), and the determinant of any identity matrix is 1. Perhaps the most important property of the determinant is the fact that a matrix in invertible if and only if its determinant is not zero.

Eigenvalues and Eigenvectors

If **A** is a square matrix, and **Av** = λv for a scalar λ and a nonzero v, then λ is an *eigenvalue* of **A** and v is an *eigenvector* of **A** that *corresponds* to λ. Any nonzero linear combination of eigenvectors corresponding to the same eigenvalue λ is also an eigenvector corresponding to λ. The collection of all eigenvectors corresponding to a given eigenvalue λ is thus a subspace, called an *eigenspace* of **A**. A collection of eigenvectors corresponding to different eigenvalues is necessarily linear-independent. It follows that a matrix of order n can have at most n distinct eigenvectors. In fact, the eigenvalues of **A** are the roots of the nth degree polynomial equation

$$\det(\mathbf{A} - \lambda \mathbf{I}) = 0$$

called the *characteristic equation* of **A**. (Eigenvalues and eigenvectors are frequently called *characteristic values* and *characteristic vectors*.)

If the nth order matrix **A** has an independent collection of n eigenvectors, then **A** is said to have a *full set* of eigenvectors. In this case there is a set of eigenvectors of **A** that is a basis for R^n or, in the complex case, C^n. In case there are n distinct eigenvalues of **A**, then, of course, **A** has a full set of eigenvectors. If there are fewer than n distinct eigenvalues, then **A** may or may not have a full set of eigenvectors. If there is a full set of eigenvectors, then

$$\mathbf{D} = \mathbf{S}^{-1}\mathbf{A}\mathbf{S} \quad \text{or} \quad \mathbf{A} = \mathbf{S}\mathbf{D}\mathbf{S}^{-1}$$

where **D** is a diagonal matrix with the eigenvalues of **A** on the diagonal, and **S** is a matrix whose columns are the full set of eigenvectors. If **A** is symmetric, there are n real distinct eigenvalues of **A** and the corresponding eigenvectors are orthogonal. There is thus an orthonormal collection of eigenvectors that span R^n, and we have

$$\mathbf{A} = \mathbf{Q}\mathbf{D}\mathbf{Q}^{\mathrm{T}} \quad \text{and} \quad \mathbf{D} = \mathbf{Q}^{\mathrm{T}}\mathbf{A}\mathbf{Q}$$

where **Q** is a real orthogonal matrix and **D** is diagonal. For the complex case, if **A** is Hermitian, we have

$$\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{U}^{H} \quad \text{and} \quad \mathbf{D} = \mathbf{U}^{H}\mathbf{A}\mathbf{U}$$

where **U** is a unitary matrix and **D** is a *real* diagonal matrix. (A Hermitian matrix also has *n* distinct real eigenvalues.)

References

Daniel, J. W. and Nobel, B. 1988. *Applied Linear Algebra.* Prentice Hall, Englewood Cliffs, NJ.
Strang, G. 1993. *Introduction to Linear Algebra*. Wellesley-Cambridge Press, Wellesley, MA.

19.3 Vector Algebra and Calculus

George Cain

Basic Definitions

A vector is a directed line segment, with two vectors being equal if they have the same length and the same direction. More precisely, a *vector* is an equivalence class of directed line segments, where two directed segments are equivalent if they have the same length and the same direction. The *length* of a vector is the common length of its directed segments, and the *angle between* vectors is the angle between any of their segments. The length of a vector u is denoted $|\mathbf{u}|$. There is defined a distinguished vector having zero length, which is usually denoted **0**. It is frequently useful to visualize a directed segment as an arrow; we then speak of the nose and the tail of the segment. The *sum* $\mathbf{u} + \mathbf{v}$ of two vectors **u** and **v** is defined by taking directed segments from **u** and **v** and placing the tail of the segment representing **v** at the nose of the segment representing **u** and defining $\mathbf{u} + \mathbf{v}$ to be the vector determined by the segment from the tail of the **u** representative to the nose of the **v** representative. It is easy to see that $\mathbf{u} + \mathbf{v}$ is well defined and that $\mathbf{u} + \mathbf{v} = \mathbf{v} + \mathbf{u}$. Subtraction is the inverse operation of addition. Thus the *difference* $\mathbf{u} - \mathbf{v}$ of two vectors is defined to be the vector that when added to **v** gives **u**. In other words, if we take a segment from **u** and a segment from **v** and place their tails together, the difference is the segment from the nose of v to the nose of **u**. The zero vector behaves as one might expect; $\mathbf{u} + \mathbf{0} = \mathbf{u}$, and $\mathbf{u} - \mathbf{u} = \mathbf{0}$. Addition is associative: $\mathbf{u} + (\mathbf{v} + \mathbf{w}) = (\mathbf{u} + \mathbf{v}) + \mathbf{w}$.

To distinguish them from vectors, the real numbers are called *scalars*. The product $t\mathbf{u}$ of a scalar t and a vector **u** is defined to be the vector having length $|t|\,|\mathbf{u}|$ and direction the same as **u** if $t > 0$, the opposite direction if $t < 0$, If $t = 0$, then $t\mathbf{u}$ is defined to be the zero vector. Note that $t(\mathbf{u} + \mathbf{v}) = t\mathbf{u} + t\mathbf{v}$, and $(t + s)\mathbf{u} = t\mathbf{u} + s\mathbf{u}$. From this it follows that $\mathbf{u} - \mathbf{v} = \mathbf{u} + (-1)\mathbf{v}$.

The *scalar product* $\mathbf{u} \cdot \mathbf{v}$ of two vectors is $|\mathbf{u}||\mathbf{v}| \cos \theta$, where θ is the angle between **u** and **v**. The scalar product is frequently called the *dot product.* The scalar product distributes over addition:

$$\mathbf{u} \cdot (\mathbf{v} + \mathbf{w}) = \mathbf{u} \cdot \mathbf{v} + \mathbf{u} \cdot \mathbf{w}$$

and it is clear that $(t\mathbf{u}) \cdot \mathbf{v} = t(\mathbf{u} \cdot \mathbf{v})$. The *vector product* $\mathbf{u} \times \mathbf{v}$ of two vectors is defined to be the vector perpendicular to both **u** and **v** and having length $|\mathbf{u}||\mathbf{v}| \sin \theta$, where θ is the angle between **u** and **v**. The direction of $\mathbf{u} \times \mathbf{v}$ is the direction a right-hand threaded bolt advances if the vector **u** is rotated to **v**. The vector is frequently called the *cross product.* The vector product is both associative and distributive, but not commutative: $\mathbf{u} \times \mathbf{v} = -\mathbf{v} \times \mathbf{u}$.

Coordinate Systems

Suppose we have a right-handed Cartesian coordinate system in space. For each vector, **u**, we associate a point in space by placing the tail of a representative of **u** at the origin and associating with **u** the point at the nose of the segment. Conversely, associated with each point in space is the vector determined by the directed segment from the origin to that point. There is thus a one-to-one correspondence between the points in space and all vectors. The origin corresponds to the zero vector. The coordinates of the point associated with a vector **u** are called *coordinates* of **u**. One frequently refers to the vector **u** and writes **u** = (*x*, *y*, *z*), which is, strictly speaking, incorrect, because the left side of this equation is a vector and the right side gives the coordinates of a point in space. What is meant is that (*x*, *y*, *z*) are the coordinates of the point associated with **u** under the correspondence described. In terms of coordinates, for $\mathbf{u} = (u_1, u_2, u_3)$ and $\mathbf{v} = (v_1, v_2, v_3)$, we have

$$\mathbf{u} + \mathbf{v} = \left(u_1 + v_1, u_2 + v_2, u_3 + v_3\right)$$

$$t\mathbf{u} = \left(tu_1, tu_2, tu_3\right)$$

$$\mathbf{u} \cdot \mathbf{v} = u_1v_1 + u_2v_2 + u_3v_3$$

$$\mathbf{u} \times \mathbf{v} = \left(u_2v_3 - v_2u_3, u_3v_1 - v_3u_1, u_1v_2 - v_1u_2\right)$$

The *coordinate vectors* **i**, **j**, and **k** are the unit vectors i = (1, 0, 0), j = (0, 1, 0), and k = (0, 0, 1). Any vector $\mathbf{u} = (u_1, u_2, u_3)$ is thus a linear combination of these coordinate vectors: $u = u_1i + u_2j + u_3k$. A convenient form for the vector product is the formal determinant

$$\mathbf{u} \times \mathbf{v} = \det\begin{bmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ u_1 & u_2 & u_3 \\ v_1 & v_2 & v_2 \end{bmatrix}$$

Vector Functions

A *vector function* **F** *of one variable* is a rule that associates a vector **F**(*t*) with each real number *t* in some set, called the *domain* of **F**. The expression $\lim_{t \to t_0} \mathbf{F}(t) = \mathbf{a}$ means that for any $\varepsilon > 0$, there is a $\delta > 0$ such that $|\mathbf{F}(t) - \mathbf{a}| < \varepsilon$ whenever $0 < |t - t_0| < \delta$. If $\mathbf{F}(t) = [x(t), y(t), z(t)]$ and $\mathbf{a} = (a_1, a_2, a_3)$, then $\lim_{t \to t_0} \mathbf{F}(t) = \mathbf{a}$ if and only if

$$\lim_{t \to t_0} x(t) = a_1$$

$$\lim_{t \to t_0} y(t) = a_2$$

$$\lim_{t \to t_0} z(t) = a_3$$

A vector function **F** is *continuous* at t_0 if $\lim_{t \to t_0} \mathbf{F}(t) = \mathbf{F}(t_0)$. The vector function **F** is continuous at t_0 if and only if each of the coordinates $x(t)$, $y(t)$, and $z(t)$ is continuous at t_0.

The function **F** is *differentiable* at t_0 if the limit

$$\lim_{h \to 0} \frac{1}{h}\left[\mathbf{F}(t + h) - \mathbf{F}(t)\right]$$

exists. This limit is called the *derivative* of **F** at t_0 and is usually written $\mathbf{F}'(t_0)$, or $(d\mathbf{F}/dt)(t_0)$. The vector function **F** is differentiable at t_0 if and only if each of its coordinate functions is differentiable at t_0. Moreover, $(d\mathbf{F}/dt)(t_0) = [(dx/dt)(t_0), (dy/dt)(t_0), (dz/dt)(t_0)]$. The usual rules for derivatives of real valued functions all hold for vector functions. Thus if **F** and **G** are vector functions and *s* is a scalar function, then

$$\frac{d}{dt}(\mathbf{F}+\mathbf{G}) = \frac{d\mathbf{F}}{dt} + \frac{d\mathbf{G}}{dt}$$

$$\frac{d}{dt}(s\mathbf{F}) = s\frac{d\mathbf{F}}{dt} + \frac{ds}{dt}\mathbf{F}$$

$$\frac{d}{dt}(\mathbf{F}\cdot\mathbf{G}) = \mathbf{F}\cdot\frac{d\mathbf{G}}{dt} + \frac{d\mathbf{F}}{dt}\cdot\mathbf{G}$$

$$\frac{d}{dt}(\mathbf{F}\times\mathbf{G}) = \mathbf{F}\times\frac{d\mathbf{G}}{dt} + \frac{d\mathbf{F}}{dt}\times\mathbf{G}$$

If **R** is a vector function defined for *t* in some interval, then, as *t* varies, with the tail of **R** at the origin, the nose traces out some object *C* in space. For nice functions **R**, the object *C* is a *curve*. If $\mathbf{R}(t) = [x(t), y(t), z(t)]$, then the equations

$$x = x(t)$$

$$y = y(t)$$

$$z = z(t)$$

are called *parametric equations* of *C*. At points where **R** is differentiable, the derivative $d\mathbf{R}/dt$ is a vector *tangent* to the curve. The unit vector $\mathbf{T} = (d\mathbf{R}/dt)/|d\mathbf{R}/dt|$ is called the *unit tangent vector*. If **R** is differentiable and if the length of the arc of curve described by **R** between $\mathbf{R}(a)$ and $\mathbf{R}(t)$ is given by $s(t)$, then

$$\frac{ds}{dt} = \left|\frac{d\mathbf{R}}{dt}\right|$$

Thus the length *L* of the arc from $\mathbf{R}(t_0)$ to $\mathbf{R}(t_1)$ is

$$L = \int_{t_0}^{t_1}\frac{ds}{dt}dt = \int_{t_0}^{t_1}\left|\frac{d\mathbf{R}}{dt}\right|dt$$

The vector $d\mathbf{T}/ds = (d\mathbf{T}/dt)/(ds/dt)$ is perpendicular to the unit tangent **T**, and the number $\kappa = |d\mathbf{T}/ds|$ is the *curvature* of *C*. The unit vector $\mathbf{N} = (1/\kappa)(d\mathbf{T}/ds)$ is the *principal normal*. The vector $\mathbf{B} = \mathbf{T}\times\mathbf{N}$ is the *binormal*, and $d\mathbf{B}/ds = -\tau N$. The number τ is the *torsion*. Note that *C* is a plane curve if and only if τ is zero for all *t*.

A *vector function* **F** *of two variables* is a rule that assigns a vector $\mathbf{F}(s, t)$ in some subset of the plane, called the *domain* of **F**. If $\mathbf{R}(s, t)$ is defined for all (s, t) in some region *D* of the plane, then as the point (s, t) varies over *D*, with its rail at the origin, the nose of $\mathbf{R}(s, t)$ traces out an object in space. For a nice function **R**, this object is a *surface*, *S*. The partial derivatives $(\partial\mathbf{R}/\partial s)(s, t)$ and $(\partial\mathbf{R}/\partial t)(s, t)$ are tangent to the surface at $\mathbf{R}(s, t)$, and the vector $(\partial\mathbf{R}/\partial s)\times(\partial\mathbf{R}/\partial t)$ is thus *normal* to the surface. Of course, $(\partial\mathbf{R}/\partial t)\times(\partial\mathbf{R}/\partial s) = -(\partial\mathbf{R}/\partial s)\times(\partial\mathbf{R}/\partial t)$ is also normal to the surface and points in the direction opposite that of $(\partial\mathbf{R}/\partial s)\times(\partial\mathbf{R}/\partial t)$. By electing one of these normal, we are choosing an *orientation* of the surface.

A surface can be oriented only if it has two sides, and the process of orientation consists of choosing which side is "positive" and which is "negative."

Gradient, Curl, and Divergence

If $f(x, y, z)$ is a scalar field defined in some region D, the *gradient* of **f** is the vector function

$$\text{grad } f = \frac{\partial f}{\partial x}\mathbf{i} + \frac{\partial f}{\partial y}\mathbf{j} + \frac{\partial f}{\partial z}\mathbf{k}$$

If $\mathbf{F}(x, y, z) = F_1(x, y, z)\mathbf{i} + F_2(x, y, z)\mathbf{j} + F_3(x, y, z)\mathbf{k}$ is a vector field defined in some region D, then the *divergence* of **F** is the scalar function

$$\text{div } \mathbf{F} = \frac{\partial F_1}{\partial x} + \frac{\partial F_2}{\partial y} + \frac{\partial F_3}{\partial z}$$

The curl is the vector function

$$\text{curl } \mathbf{F} = \left(\frac{\partial F_3}{\partial y} - \frac{\partial F_2}{\partial z}\right)\mathbf{i} + \left(\frac{\partial F_1}{\partial z} - \frac{\partial F_3}{\partial x}\right)\mathbf{j} + \left(\frac{\partial F_2}{\partial x} - \frac{\partial F_1}{\partial y}\right)\mathbf{k}$$

In terms of the vector operator *del*, $\nabla = \mathbf{i}(\partial/\partial x) + \mathbf{j}(\partial/\partial y) + \mathbf{k}(\partial/\partial z)$, we can write

$$\text{grad } f = \nabla f$$

$$\text{div } \mathbf{F} = \nabla \cdot \mathbf{F}$$

$$\text{curl } \mathbf{F} = \nabla \times \mathbf{F}$$

The *Laplacian operator* is div (grad) $= \nabla \cdot \nabla = \nabla^2 = (\partial^2/\partial x^2) + (\partial^2/\partial y^2) + (\partial^2/\partial z^2)$.

Integration

Suppose C is a curve from the point (x_0, y_0, z_0) to the point (x_1, y_1, z_1) and is described by the vector function $\mathbf{R}(t)$ for $t_0 \le t \le t_1$. If f f is a scalar function (sometimes called a *scalar field*) defined on C, then the integral of f over C is

$$\int_C f(x,y,z)\, ds = \int_{t_0}^{t_1} f\left[\mathbf{R}(t)\right]\left|\frac{d\mathbf{R}}{dt}\right| dt$$

If **F** is a vector function (sometimes called a *vector field*) defined on C, then the integral of **F** over C is

$$\int_C \mathbf{F}(x,y,z)\cdot d\mathbf{R} = \int_{t_0}^{t_1} \mathbf{F}\left[\mathbf{R}(t)\right]\frac{d\mathbf{R}}{dt}\, dt$$

These integrals are called *line integrals*.

In case there is a scalar function f such that $\mathbf{F} = \text{grad } f$, then the line integral

$$\int_C \mathbf{F}(x,y,z)\cdot d\mathbf{R} = f\left[\mathbf{R}(t_1)\right] - f\left[\mathbf{R}(t_0)\right]$$

The value of the integral thus depends only on the end points of the curve C and not on the curve C itself. The integral is said to be *path-independent.* The function f is called a *potential function* for the vector field **F**, and **F** is said to be a *conservative field.* A vector field **F** with domain D is conservative if and only if the integral of **F** around every closed curve in D is zero. If the domain D is simply connected (that is, every closed curve in D can be continuously deformed in D to a point), then **F** is conservative if and only if curl **F** = 0 in D.

Suppose S is a surface described by $\mathbf{R}(s, t)$ for (s, t) in a region D of the plane. If f is a scalar function defined on D, then the integral of f over S is given by

$$\iint_S f(x,y,z)\, dS = \iint_D f[\mathbf{R}(s,t)] \left| \frac{\partial \mathbf{R}}{\partial s} \times \frac{\partial \mathbf{R}}{\partial t} \right| ds\, dt$$

If **F** is a vector function defined on S, and if an orientation for S is chosen, then the integral **F** over S, sometimes called the flux of **F** through S, is

$$\iint_S \mathbf{F}(x,y,z) \cdot d\mathbf{S} = \iint_D \mathbf{F}[\mathbf{R}(s,t)] \left| \frac{\partial \mathbf{R}}{\partial s} \times \frac{\partial \mathbf{R}}{\partial t} \right| ds\, dt$$

Integral Thorems

Suppose **F** is a vector field with a closed domain D bounded by the surface S oriented so that the normal points out from D. Then the *divergence theorem* states that

$$\iiint_D \text{div } \mathbf{F}\, dV = \iint_S \mathbf{F} \cdot \mathbf{dS}$$

If S is an orientable surface bounded by a closed curve C, the orientation of the closed curve C is chosen to be consistent with the orientation of the surface S. Then we have *Stoke's theorem:*

$$\iint_S (\text{curl } \mathbf{F}) \cdot d\mathbf{S} = \oint_C \mathbf{F} \cdot \mathbf{ds}$$

References

Davis, H. F. and Snider, A. D. 1991. *Introduction to Vector Analysis,* 6th ed., Wm. C. Brown, Dubuque, IA.
Wylie, C. R. 1975. *Advanced Engineering Mathematics,* 4th ed., McGraw-Hill, New York.

Further Information

More advanced topics leading into the theory and applications of tensors may be found in J. G. Simmonds, *A Brief on Tensor Analysis* (1982, Springer-Verlag, New York).

19.4 Difference Equations

William F. Ames

Difference equations are equations involving *discrete variables*. They appear as natural descriptions of natural phenomena and in the study of discretization methods for differential equations, which have continuous variables.

Let $y_n = y(nh)$, where n is an integer and h is a real number. (One can think of measurements taken at equal intervals, h, $2h$, $3h$, …, and y_n describes these). A typical equation is that describing the famous Fibonacci sequence — $y_{n+2} - y_{n+1} - y_n = 0$. Another example is the equation $y_{n+2} - 2zy_{n+1} + y_n = 0$, $z \in C$, which describes the Chebyshev polynomials.

First-Order Equations

The general first-order equation $y_{n+1} = f(y_n)$, $y_0 = y(0)$ is easily solved, for as many terms as are needed, by *iteration*. Then $y_1 = f(y_0)$; $y_2 = f(y_1)$,…. An example is the logistic equation $y_{n+1} = ay_n(1 - y_n) = f(y_n)$. The logistic equation has two fixed (critical or equilibrium) points where $y_{n+1} = y_n$. They are 0 and $\bar{y} = (a - 1)/a$. This has physical meaning only for $a > 1$. For $1 < a < 3$ the equilibrium $\bar{y}$ is asymptotically stable, and for $a > 3$ there are two points y_1 and y_2, called a *cycle of period two*, in which $y_2 = f(y_1)$ and $y_1 = f(y_2)$. This study leads into chaos, which is outside our interest. By iteration, with $y_0 = {}^1/_2$, we have $y_1 = (a/2)(1/2) = a/2^2$, $y_2 = a(a/2^2)(1 - a/2^2) = (a^2/2^2)(1 - a/2^2)$, ….

With a constant, the equation $y_{n+1} = ay_n$ is solved by making the assumption $y_n = A\lambda^n$ and finding λ so that the equation holds. Thus $A\lambda^{n+1} = aA\lambda^n$, and hence $\lambda = 0$ or $\lambda = a$ and A is arbitrary. Discarding the trivial solution 0 we find $y_n = Aa^{n+1}$ is the desired solution. By using a method called the *variation of constants*, the equation $y_{n+1} - ay_n = g_n$ has the solution $y_n = y_0a^n + \Sigma_{j=0}^{n-1} g_j a^{n-j-1}$, with y_0 arbitrary.

In various applications we find the first-order equation of *Riccati type* $y_ny_{n-1} + ay_n + by_{n-1} + c = 0$ where a, b, and c are real constants. This equation can be transformed to a linear second-order equation by setting $y_n = z_n/z_{n-1} - a$ to obtain $z_{n+1} + (b + a)z_n + (c - ab)z_{n-1} = 0$, which is solvable as described in the next section.

Second-Order Equations

The second-order linear equation with constant coefficients $y_{n+2} + ay_{n+1} + by_n = f_n$ is solved by first solving the homogeneous equation (with right-hand side zero) and adding to that solution any solution of the inhomogeneous equation. The *homogeneous equation* $y_{n+2} + ay_{n+1}$ $by_n = 0$ is solved by assuming $y_n = \lambda^n$, whereupon $\lambda^{n+2} + a\lambda^{n+1} + b\lambda^n = 0$ or $\lambda = 0$ (rejected) or $\lambda^2 + a\lambda + b = 0$. The roots of this quadratic are $\lambda_1 = {}^1/_2\,(-a + \sqrt{a^2 - 4b})$, $\lambda_2 = -{}^1/_2\,(a + \sqrt{a^2 - 4b})$ and the solution of the homogeneous equation is $y_n = c_1\lambda_1^n + c_2\lambda_2^n$. As an example consider the Fibonacci equation $y_{n+2} - y_{n+1} - y_n = 0$. The roots of $\lambda^2 - \lambda - 1 = 0$ are $\lambda_1 = {}^1/_2\,(1 + \sqrt{5})$, $\lambda_2 = {}^1/_2\,(1 - \sqrt{5})$, and the solution $y_n = c_1[(1 + \sqrt{5})/2]^n + c_2[(1 - \sqrt{5})/2]^n$ is known as the *Fibonacci sequence.*

Many of the orthogonal polynomials of differential equations and numerical analysis satisfy a second-order difference equation (recurrence relation) involving a discrete variable, say n, and a continuous variable, say z. One such is the *Chebshev equation* $y_{n+2} - 2zy_{n+1} + y_n = 0$ with the initial conditions $y_0 = 1$, $y_1 = z$ (*first-kind* Chebyshev polynomials) and $y_{n-1} = 0$, $y_0 = 1$ (second-kind Chebyshev polynomials). They are denoted $T_n(z)$ and $V_n(z)$, respectively. By iteration we find

$$T_0(z) = 1, \quad T_1(z) = z, \quad T_2(z) = 2z^2 - 1,$$

$$T_3(z) = 4z^3 - 3z, \quad T_4(z) = 8z^4 - 8z^2 + 1$$

$$V_0(z) = 0, \quad V_1(z) = 1, \quad V_2(z) = 2z,$$

$$V_3(z) = 4z^2 - 1, \quad V_4(z) = 8z^3 - 4z$$

and the general solution is $y_n(z) = c_1T_n(z) + c_2V_{n-1}(z)$,

Linear Equations with Constant Coefficients

The genral kth-order linear equation with constant coefficients is $\Sigma_{i=0}^{k} p_i y_{n+k-i} = g_n$, $p_0 = 1$. The solution to the corresponding homogeneous equation (obtained by setting $g_n = 0$) is as follows. (a) $y_n = \Sigma_{i=1}^{k} c_i \lambda_i^n$ if the λ_i are the distinct roots of the characteristic polynomial $p(\lambda) = \Sigma_{i=0}^{k} p_i \lambda^{k-i} = 0$. (b) if m_s is the multiplicity of the root λ_s, then the functions $y_{n,s} = u_s(n)\lambda_s^n$, where $u_s(n)$ are polynomials in n whose degree does not exceed $m_s - 1$, are solutions of the equation. Then the general solution of the homogeneous equation is $y_n = \Sigma_{i=1}^{d} a_i u_i(n)\lambda_i^n = \Sigma_{i=1}^{d} a_i \Sigma_{j=0}^{m_i-1} c_j n^j \lambda_i^n$. To this solution one adds any particular solution to obtain the general solution of the general equation.

Example 19.4.1. A model equation for the price p_n of a product, at the nth time, is $p_n + b/a(1 + \rho)p_{n-1} - (b/a)\rho p_{n-2} + (s_0 - d_0)/a = 0$. The equilibrium price is obtained by setting $p_n = p_{n-1} = p_{n-2} = p_e$, and one finds $p_e = (d_0 - s_0)/(a + b)$. The homogeneous equation has the characteristic polynomial $\lambda^2 + (b/a)(1 + \rho)\lambda - (b/a)\rho = 0$. With λ_1 and λ_2 as the roots the general solution of the full equation is $p_n = c_1\lambda_1^n + c_2\lambda_2^n + p_e$, since p_e is a solution of the full equation. This is one method for finding the solution of the nonhomogeneous equation.

Generating Function (z Transform)

An elegant way of solving linear difference equations with constant coefficients, among other applications, is by use of *generating functions* or, as an alternative, the z transform. The generating function of a sequence $\{y_n\}$, $n = 0, 1, 2, \ldots$, is the function $f(x)$ given by the formal series $f(x) = \Sigma_{n=0}^{\infty} y_n x^n$. The z transform of the same sequence is $z(x) = \Sigma_{n=0}^{\infty} y_n x^{-n}$. Clearly, $z(x) = f(1/x)$. A table of some important sequences is given in Table 19.4.1.

Table 19.4.1 Important Sequences

y_n	$f(x)$	Convergence Domain
1	$(1 - x)^{-1}$	$\|x\| < 1$
n	$x(1 - x)^{-2}$	$\|x\| < 1$
n^m	$xp_m(x)(1 - x)^{-n-1}$*	$\|x\| < 1$
k^n	$(1 - kx)^{-1}$	$\|x\| < k^{-1}$
e^{an}	$(1 - e^a x)^{-1}$	$\|x\| < e^{-a}$
$k^n \cos an$	$\dfrac{1 - kx\cos a}{1 - 2kx\cos a + k^2x^2}$	$\|x\| < k^{-1}$
$k^n \sin an$	$\dfrac{kx\sin a}{1 - 2kx\cos a + k^2x^2}$	$\|x\| < k^{-1}$
$\binom{n}{m}$	$x^m(1 - x)^{-m-1}$	$\|x\| < 1$
$\binom{k}{n}$	$(1 + x)^k$	$\|x\| < 1$

* The term $p_m(z)$ is a polynomial of degree m satisfying $p_{m+1}(z) = (mz + 1) \cdot p_m(z) + z(1 - z)\ p'_m(x)$, $p_1 = 1$.

To solve the linear difference equation $\Sigma_{i=0}^{k} p_i y_{n+k-i} = 0$, $p_0 = 1$ we associate with it the two formal series $P = p_0 + p_1 x + \cdots + p_k x^k$ and $Y = y_0 + y_1 x + y_2 x^{2+} \cdots$. If $p(x)$ is the characteristic polynomial then $P(x) = x^k p(1/x) = \bar{p}(x)$. The *product* of the two series is $Q = YP = q_0 + q_1 x + \cdots + q_{k-1} x^{k-1} + q_k x^k + \cdots$ where $q_n = \Sigma_{i=0}^{n} p_i y_{n-i}$. Because $p_{k+1} = p_{k+2} = \cdots = 0$, it is obvious that $q_{k+1} = q_{k+2} = \cdots = 0$ — that is, Q is a polynomial (formal series with finite number of terms). Then $Y = P^{-1}Q = q(x)/\bar{p}(x) = q(x)/x^k p(1/x)$, where p is the characteristic polynomial and $q(x) = \Sigma_{i=0}^{k} q_i x^i$. The roots of $\bar{p}(x)$ are x_i^{-1} where the x_i are the roots of $p(x)$.

Theorem 1. If the roots of $p(x)$ are less than one in absolute value, then $Y(x)$ converges for $|x| < 1$.

Thorem 2. If $p(x)$ has no roots greater than one in absolute value and those on the unit circle are simple roots, then the coefficients y_n of Y are bounded. Now $q_k = g_0$, $q_{n+k} = g_n$, and $Q(x) = Q_1(x) + x^k Q_2(x)$. Hence $\Sigma_{i=1}^{\infty} y_i x^i = [Q_1(x) + x^k Q_2(x)]/[\bar{p}(x)]$.

Example 19.4.2. Consider the equation $y_{n+1} + y_n = -(n+1)$, $y_0 = 1$. Here $Q_1 = 1$, $Q_2 = -\Sigma_{n=0}^{\infty}(n+1)x^n = -1/(1-x)^2$.

$$G(x) = \frac{1 - x/(1-x)^2}{1+x} = \frac{5}{4}\frac{1}{1+x} - \frac{1}{4}\frac{1}{1-x} - \frac{1}{2}\frac{x}{(1-x)^2}$$

Using the table term by term, we find $\Sigma_{n=0}^{\infty} y_n x^n = \Sigma_{n=0}^{\infty}[{}^5/_4(-1)^n - {}^1/_4 - {}^1/_2 n]x^n$, so $y_n = {}^5/_4(-1)^n - {}^1/_4 - {}^1/_2 n$.

References

Fort, T. 1948. *Finite Differences and Difference Equations in the Real Domain.* Oxford University Press, London.

Jordan, C. 1950. *Calculus of Finite Differences,* Chelsea, New York.

Jury, E. I. 1964. *Theory and Applications of the Z Transform Method.* John Wiley & Sons, New York.

Lakshmikantham, V. and Trigrante, D. 1988. *Theory of Difference Equations.* Academic Press, Boston, MA.

Levy, H. and Lessman, F. 1961. *Finite Difference Equations.* Macmillan, New York.

Miller, K. S. 1968. *Linear Difference Equations,* Benjamin, New York.

Wilf, W. S. 1994. *Generating Functionology,* 2nd ed. Academic Press, Boston, MA.

19.5 Differential Equations

William F. Ames

Any equation involving derivatives is called a *differential equation.* If there is only one independent variable the equation is termed a *total differential equation* or an *ordinary differential equation*. If there is more than one independent variable the equation is called a *partial differential equation.* If the highest-order derivative is the *n*th then the equation is said to be *n*th order. If there is no function of the dependent variable and its derivatives other than the linear one, the equation is said to be *linear.* Otherwise, it is *nonlinear.* Thus $(d^3y/dx^3) + a(dy/dx) + by = 0$ is a *linear* third-order ordinary (total) differential equation. If we replace by with by^3, the equation becomes nonlinear. An example of a second-order linear partial differential equation is the famous wave equation $(\partial^2u/\partial x^2) - a^2(\partial^2u/\partial t^2) = f(x)$. There are two independent variables x and t and $a^2 > 0$ (of course). If we replace $f(x)$ by $f(u)$ (say u^3 or $\sin u$) the equation is nonlinear. Another example of a nonlinear third-order partial differential equation is $u_t + uu_x = au_{xxx}$. This chapter uses the common subscript notation to indicate the partial derivatives.

Now we briefly indicate some methods of solution and the solution of some commonly occurring equations.

Ordinary Differential Equations

First-Order Equations

The *general* first-order equation is $f(x, y, y') = 0$. Equation capable of being written in either of the forms $y' = f(x)g(y)$ or $f(x)g(y)y' + F(x)G(y) = 0$ are *separable* equations. Their solution is obtained by using $y' = dy/dx$ and writing the equations in differential form as $dy/g(y) = f(x)dx$ or $g(y)[dy/G(y)] = -F(x)[dx/f(x)]$ and integrating. An example is the famous *logistic* equation of inhibited growth $(dy/dt) = ay(1 - y)$. The integral of $dy/y(1 - y) = adt$ is $y = 1/[1 + (y_0^{-1} - 1)e^{-at}]$ for $t \geq 0$ and $y(0) = y_0$ (the initial state called the *initial condition*).

Equations may not have unique solutions. An example is $y' = 2y^{1/2}$ with the initial condition $y(0) = 0$. One solution by separation is $y = x^2$. But there are an *infinity* of others — namely, $y_a(x) = 0$ for $-\infty < x \leq a$, and $(x - a)^2$ for $a \leq x < \infty$.

If the equation $P(x, y)dy + Q(x, y)dy = 0$ is reducible to

$$\frac{dy}{dx} = f\left(\frac{y}{x}\right) \quad \text{or} \quad \frac{dy}{dx} = f\left(\frac{a_1x + b_1y + c_1}{a_2x + b_2y + c_2}\right)$$

the equation is called *homogenous* (nearly homogeneous). The first form reduces to the separable equation $u + x(du/dx) = f(u)$ with the substitution $y/x = u$. The nearly homogeneous equation is handled by setting $x = X + a$, $y = Y + \beta$, and choosing α and β so that $a_1\alpha + b_1\beta + c_1 = 0$ and $a_2\alpha + b_2\beta + c_2 = 0$. If $\begin{vmatrix} a_1 & b_1 \\ a_2 & b_2 \end{vmatrix} \neq 0$ this is always possible; the equation becomes $dY/dX = [a_1 + b_1(Y/X)]/[a_2 + b_2(Y/X)]$ and the substitution $Y = Xu$ gives a separable equation. If $\begin{vmatrix} a_1 & b_1 \\ a_2 & b_2 \end{vmatrix} = 0$ then $a_2x + b_2y = k(a_1x + b_1y)$ and the equation becomes $du/dx = a_1 + b_1(u + c_1)/(ku + c_2)$, with $u = a_1x + b_1y$. Lastly, any equation of the form $dy/dx = f(ax + by + c)$ transforms into the separable equation $du/dx = a + bf(u)$ using the change of variable $u = ax + by + c$.

The general first-order linear equation is expressible in the form $y' + f(x)y = g(x)$. It has the *general solution* (a solution with an arbitrary constant c)

$$y(x) = \exp\left[-\int f(x)\,dx\right]\left\{c + \int \exp\left[f(x)\right]g(x)\,dx\right\}$$

Two noteworthy examples of first-order equations are as follows:

1. An often-occurring nonlinear equation is the *Bernoulli equation*, $y' + p(x)y = g(x)y^{\alpha}$, with α real, $\alpha \neq 0$, $\alpha \neq 1$. The transformation $z = y^{1-\alpha}$ converts the equation to the linear first-order equation $z' + (1 - \alpha)p(x)z = (1 - \alpha)q(x)$.
2. The famous *Riccati equation*, $y' = p(x)y^2 + q(x)y + r(x)$, cannot in general be solved by integration. But some useful transformations are helpful. The substitution $y = y_1 + u$ leads to the equation $u' - (2py_1 + q)u = pu^2$, which is a Bernoulli equation for u. The substitution $y = y_1 + v^{-1}$ leads to the equation $v' + (2py_1 + q)v + p = 0$, which is a linear first-order equation for v. Once either of these equations has been solved, the general solution of the Riccati equation is $y = y_1 + u$ or $y = y_1 + v^{-1}$.

Second-Order Equations

The simplest of the second-order equations is $y'' + ay' + by = 0$ (a, b real), with the initial conditions $y(x_0) = y_0$, $y'(x_0) = y_0'$ or the boundary conditions $y(x_0) = y_0$, $y(x_1) = y_1$. The general solution of the equation is given as follows.

1. $a^2 - 4b > 0$, $\lambda_1 = {}^1\!/_2\,(-a + \sqrt{a^2 - 4b})$, $\lambda_2 = {}^1\!/_2\,(-a - \sqrt{a^2 - 4b})$
   $y = c_1 \exp(\lambda_1 x) + c_2 \exp(\lambda_2 x)$
2. $a^2 - 4b = 0$, $\lambda_1 = \lambda_2 = -\frac{a}{2}$, $y = (c_1 + c_2 x)\exp(\lambda_1 x)$
3. $a^2 - 4b < 0$, $\lambda_1 = {}^1\!/_2\,(-a + i\sqrt{4b - a^2})$, $\lambda_2 = {}^1\!/_2\,(-a - i\sqrt{4b - a^2})$,
   $i^2 = -1$
   With $p = -a/2$ and $q = {}^1\!/_2\,\sqrt{4b - a^2}$,

$$y = c_1 \exp\left[(p + iq)x\right] + c_2 \exp\left[(p - iq)x\right] = \exp(px)\left[A\sin qx + B\cos qx\right]$$

   The initial conditions or boundary conditions are used to evaluate the arbitrary constants c_1 and c_2 (or A and B).
   Note that a linear problem with specified data may not have a solution. This is especially serious if numerical methods are employed without serious thought.

For example, consider $y'' + y = 0$ with the boundary condition $y(0) = 1$ and $y(\pi) = 1$. The general solution is $y = c_1 \sin x + c_2 \cos x$. The first condition $y(0) = 1$ gives $c_2 = 1$, and the second condition requires $y(\pi) = c_1 \sin \pi + \cos \pi$ or "1 = –1," which is a *contradiction.*

Example 19.5.1 — The Euler Strut. When a strut of uniform construction is subject to a compressive load P it exhibits no transverse displacement until P exceeds some critical value P_1. When this load is exceeded, buckling occurs and large deflections are produced as a result of small load changes. Let the rod of length ℓ be placed as shown in Figure 19.5.1.

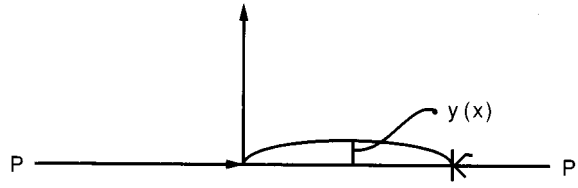


FIGURE 19.5.1

From the linear theory of elasticity (Timoshenko), the transverse displacement $y(x)$ satisfies the linear second-order equation $y'' + (Py/EI) = 0$, where E is the modulus of elasticity and I is the moment of inertia of the strut. The boundary conditions are $y(0) = 0$ and $y(a) = 0$. With $k^2 = P/EI$ the general solution is $y = c_1 \sin kx + c_2 \cos kx$. The condition $y(0) = 0$ gives $c_2 = 0$. The second condition gives $c_1 \sin ka = 0$. Since $c_1 = 0$ gives the trival solution $y = 0$ we must have $\sin ka = 0$. This occurs for $ka = n\pi$, $n = 0, 1, 2, \ldots$ (these are called *eigenvalues*). The first nontrivial solution occurs for $n = 1$ — that is, $k = \pi/a$ — whereupon $y_1 = c_1 \sin(\pi/a)$, with arbitrary c_1. Since $P = EIk^2$ the critical compressive load is $P_1 = EI\pi^2/a^2$. This is the buckling load. The weakness of the linear theory is its failure to model the situation when buckling occurs.

Example 19.5.2 — Some Solvable Nonlinear Equations. Many physical phenomena are modeled using nonlinear second-order equations. Some general cases are given here.

1. $y'' = f(y)$, *first integral* $(y')^2 = 2\int f(y)\,dy + c$.
2. $f(x, y', y'') = 0$. Set $p = y'$ and obtain a first-order equation $f(x, p, dp/dx) = 0$. Use first-order methods.
3. $f(y, y', y'') = 0$. Set $p = y'$ and then $y'' = p(dp/dy)$ so that a first-order equation $f[y, p, p(dp/dy) = 0$ for p as a function of y is obtained.
4. The *Riccati transformation* $du/dx = yu$ leads to the Riccati chain of equations, which linearize by raising the order. Thus,

Equation in y	Equation in u
1. $y' + y^2 = f(x)$	$u'' = f(x)u$
2. $y'' + 3yy' + y^3 = f(x)$	$u''' = f(x)u$
3. $y''' + 6y^2y' + 3(y')^2 + 4yy'' = f(x)$	$u^{(iv)} = f(x)u$

This method can be generalized to $u' = a(x)yu$ or $u' = a(x)f(u)y$.

Second-Order Inhomogeneous Equations

The general solution of $a_0(x)y'' + a_1(x)y' + a_2(x)y = f(x)$ is $y = y_H(x) + y_p(x)$ where $y_H(x)$ is the general solution of the homogeneous equation (with the right-hand side zero) and y_p is the particular integral of the equation. Construction of particular integrals can sometimes be done by the *method of undetermined coefficients*. See Table 19.5.1. This applies only to the linear constant coefficient case in which the function $f(x)$ is a linear combination of a polynomial, exponentials, sines and cosines, and some products of these functions. This method has as its base the observation that repeated differentiation of such functions gives rise to similar functions.

Table 19.5.1 Method of Undetermined Coefficients — Equation $L(y) = f(x)$ (Constant Coefficients)

Terms in $f(x)$	Terms To Be Included in $y_p(x)$
1. Polynomial of degree n	(i) If $L(y)$ contains y, try $y_p = a_0x^n + a_1x^{n-1} + \cdots + a_n$, (ii) If $L(y)$ does not contain y and lowest-order derivative is $y^{(r)}$, try $y_p = a_0x^{n+r} + \cdots + a_nx^r$.
2. $\sin qx$, $\cos qx$	(i) $\sin qx$ and/or $\cos qx$ are not in y_H; $y_p = B \sin qx + C \cos qx$. (ii) y_H contains terms of form $x^r \sin qx$ and/or $x^r \cos qx$ for $r = 0, 1, \ldots, m$; include in y_p terms of the form $a_0x^{m+1} \sin qx + a_1x^{m+1} \cos qx$.
3. e^{ax}	(i) y_H does not contain e^{ax}; include Ae^{ax} in y_p. (ii) y_H contains $e^{ax}, xe^{ax}, \ldots, x^ne^{ax}$; include in y_p terms of the form $Ax^{n+1}e^{ax}$,
4. $e^{px} \sin qx$, $e^{px} \cos qx$	(i) y_H does not contain these terms; in y_p include $Ae^{px} \sin qx + Be^{px} \cos qx$. (ii) y_H contains $x^r e^{px} \sin qx$ and/or $x^r e^{px} \cos qx$; $r = 0,1, \ldots, m$ include in y_p. $Ax^{m+1}e^{px} \sin qx + Bx^{m+1} e^{px} \cos qx$.

Example 19.5.3. Consider the equation $y'' + 3y' + 2y = \sin 2x$. The characteristic equation of the homogeneous equation $\lambda^2 + 3\lambda + 2 = 0$ has the two roots $\lambda_1 = -1$ and $\lambda_2 = -2$. Consequently, $y_H = c_1e^{-x} + c_2e^{-2x}$. Since $\sin 2x$ is not linearly dependent on the exponentials and since $\sin 2x$ repeats after two

differentiations, we assume a particular solution with undetermined coefficients of the form $y_p(x) = B \sin 2x + C \cos 2x$. Substituting into the original equation gives $-(2B + 6C) \sin 2x + (6B - 2C) \cos 2x = \sin 2x$. Consequently, $-(2B + 6C) = 1$ and $6B - 2C = 0$ to satisfy the equation. These two equations in two unknowns have the solution $B = -1/20$ and $C = -3/20$. Hence $y_p = -1/20 (\sin 2x + 3 \cos 2x)$ and $y = c_1e^{-x} + c_2e^{-2x} - 1/20 (\sin 2x + 3 \cos 2x)$.

A general method for finding $y_p(x)$ called *variation of parameters* uses as its starting point $y_H(x)$. This method applies to *all* linear differential equations irrespective of whether they have constant coefficients. But it assumes $y_H(x)$ is known. We illustrate the idea for $a(x)y'' + b(x)y' + c(x)y = f(x)$. If the solution of the homogeneous equation is $y_H(x) = c_1\phi_1(x) + c_2\phi_2(x)$, then vary the parameters c_1 and c_2 to seek $y_p(x)$ as $y_p(x) = u_1(x)\phi_1(x) + u_2(x)\phi_2(x)$. Then $y_P' = u_1\,\phi_1' + u_2\,\phi_2' + u_1'\,\phi_1 + u_2'\,\phi_2$ and choose $u_1'\,\phi_1 + u_2'\,\phi_2 = 0$. Calculating y_P'' and setting in the original equation gives $a(x)\,u_1'\phi_1' + a(x)\,u_2'\phi_2' = f$. Solving the last two equations for u_1' and u_2' gives $u_1' = -\phi_2 f/wa$, $u_2' = \phi_1 f/wa$, where $w = \phi_1\,\phi_2' - \phi_1'\,\phi_2 \neq 0$. Integrating the general solution gives $y = c_1\phi_1(x) + c_2\phi_2(x) - \{\int[\phi_2 f(x)]/wa\}\phi_1(x) + [\int(\phi_1 f/wa)dx]\phi_2(x)$.

Example 19.5.4. Consider the equations $y'' - 4y = \sin x/(1 + x^2)$ and $y_H = c_1e^{-2x} + c_2e^{-2x}$. With $\phi_1 = e^{2x}$, and $\phi_2 = e^{-2x}$, $w = 4$, so the general solution is

$$y = c_1e^{2x} + c_2e^{-2x} - \frac{e^{-2x}}{4}\int\frac{e^{2x}\sin x}{1+x^2}\,dx + \frac{e^{2x}}{4}\int\frac{e^{-2x}\sin x}{1+x^2}\,dx$$

The method of variation of parameters can be generalized as described in the references.

Higher-order systems of linear equations with constant coefficients are treated in a similar manner. Details can be found in the references.

Series Solution

The solution of differential equations can only be obtained in closed form in special cases. For all others, series or approximate or numerical solutions are necessary. In the simplest case, for an initial value problem, the solution can be developed as a Taylor series expansion about the point where the initial data are specified. The method fails in the *singular case* — that is, a point where the coefficient of the highest-order derivative is zero. The general method of approach is called the *Frobenius method.*

To understand the nonsingular case consider the equation $y'' + xy = x^2$ with $y(2) = 1$ and $y'(2) = 2$ (an initial value problem). We seek a series solution of the form $y(x) = a_0 + a_1(x - 2) + a_2(x - 2)^2 + \cdots$. To proceed, set $1 = y(2) = a_0$, which evaluates a_0. Next $y'(x) = a_1 + 2a_2(x - 2) + \cdots$, so $2 = y'(2) = a_1$ or $a_1 = 2$. Next $y''(x) = 2a_2 + 6a_3(x - 2) + \cdots$. and from the equation, $y'' = x^2 - xy$, so $y''(2) = 4 - 2y(2) = 4 - 2 = 2$. Hence $2 = 2a_2$ or $a_2 = 1$. Thus, to third-order $y(x) = 1 + 2(x - 2) + (x - 2)^2 + R_2(x)$, where the remainder $R_2(x)$ $[(x - 2)^3/3]y'''(\xi)$, where $2 < \xi < x$ can be bounded for each x by finding the maximum of $y'''(x) = 2x - y - xy'$. The third term of the series follows by evaluating $y'''(2) = 4 - 1 - 2 \cdot 2 = -1$, so $6a_3 = -1$ or $a_3 = -1/6$.

By now the nonsingular process should be familiar. The algorithm for constructing a series solution about a nonsingular (ordinary) point x_0 of the equation $P(x)y'' + Q(x)y' + R(x)y = f(x)$ (note that $P(x_0) \neq 0$) is as follows:

1. Substitute into the differential equation the expressions

$$y(x) = \sum_{n=0}^{\infty} a_n\left(x - x_0\right)^n, \qquad y'(x) = \sum_{n=1}^{\infty} na_n\left(x - x_0\right)^{n-1}, \qquad y''(x) = \sum_{n=2}^{\infty} n(n-1)a_n\left(x - x_0\right)^{n-2}$$

2. Expand $P(x)$, $Q(x)$, $R(x)$, and $f(x)$ about the point x_0 in a power series in $(x - x_0)$ and substitute these series into the equation.
3. Gather all terms involving the same power of $(x - x_0)$ to arrive at an identity of the form $\Sigma_{n=0}^{\infty}$ $A_n(x - x_0)^n \equiv 0$.

4. Equate to zero each coefficient A_n of step 3.
5. Use the expressions of step 4 to determine a_2, a_3, … in terms of a_0, a_1 (we need two arbitrary constants) to arrive at the general solution.
6. With the given initial conditions, determine a_0 and a_1.

If the equation has a regular singular point — that is, a point x_0 at which $P(x)$ vanishes and a series expansion is sought about that point — a solution is sought of the form $y(x) = (x - x_0)^r \sum_{n=0}^{\infty} a_n(x - x_0)^n$, $a_0 \neq 0$ and the index r and coefficients a_n must be determined from the equation by an algorithm analogous to that already described. The description of this Frobenius method is left for the references.

Partial Differential Equations

The study of partial differential equations is of continuing interest in applications. It is a vast subject, so the focus in this chapter will be on the most commonly occurring equations in the engineering literature — the second-order equations in two variables. Most of these are of the three basic types: elliptic, hyperbolic, and parabolic.

Elliptic equations are often called *potential equations* since they occur in potential problems where the potential may be temperature, voltage, and so forth. They also give rise to the steady solutions of parabolic equations. They require boundary conditions for the complete determination of their solution.

Hyperbolic equations are often called *wave equations* since they arise in the propagation of waves. For the development of their solutions, initial and boundary conditions are required. In principle they are solvable by the method of characteristics.

Parabolic equations are usually called *diffusion equations* because they occur in the transfer (diffusion) of heat and chemicals. These equations require initial conditions (for example, the initial temperature) and boundary conditions for the determination of their solutions.

Partial differential equations (PDEs) of the second order in two independent variables (x, y) are of the form $a(x, y)u_{xx} + b(x, y)u_{xy} + c(x, y)u_{yy} = E(x, y, u, u_x, u_y)$. If $E = E(x, y)$ the equation is linear; if E depends also on u, u_x, and u_y, it is said to be *quasilinear,* and if E depends only on x, y, and u, it is *semilinear.* Such equations are classified as follows: If $b^2 - 4ac$ is less than, equal to, or greater than zero at some point (x, y), then the equation is elliptic, parabolic, or hyperbolic, respectively, at that point. A PDE of this form can be transformed into canonical (standard) forms by use of new variables. These standard forms are most useful in analysis and numerical computations.

For hyperbolic equations the standard form is $u_{\xi\eta} = \phi(u, u_\eta, u_\xi, \eta, \xi)$, where $\xi_x/\xi_y = (-b + \sqrt{b^2 - 4ac})/2a$, and $\eta_x/\eta_y = (-b - \sqrt{b^2 - 4ac})/2a$. . The right-hand sides of these equations determine the so-called characteristics $(dy/dx)|_+ = (-b + \sqrt{b^2 - 4ac})/2a$, $(dy/dx)|_- = (-b - \sqrt{b^2 - 4ac})/2a$.

Example 19.5.5. Consider the equation $y^2u_{xx} - x^2u_{yy} = 0$, $\xi_x/\xi_y = -x/y$, $\eta_x/\eta_y = x/y$, so $\xi = y^2 - x^2$ and $\eta = y^2 + x^2$. In these new variables the equation becomes $u_{\xi\eta} = (\xi u_\eta - \eta u_\xi)/2(\xi^2 - \eta^2)$.

For parabolic equations the standard form is $u_{\xi\xi} = \phi(u, u_\eta, u_\xi, \eta, \xi)$ or $u_{\eta\eta} = \phi(u, u_\eta, u_\xi, \xi, \eta)$, depending upon how the variables are defined. In this case $\xi_x/\xi_y = -b/2a$ if $a \neq 0$, and $\xi_x/\xi_y = -b/2c$ if $c \neq 0$. Only ξ must be determined (there is only one characteristic) and η can be chosen as any function that is linearly independent of ξ.

Example 19.5.6. Consider the equation $y^2u_{xx} - 2xyu_{xy} + x^2u_{yy} + u_y = 0$. Clearly, $b^2 - 4ac = 0$. Neither a nor c is zero so either path can be chosen. With $\xi_x/\xi_y = -b/2a = x/y$, there results $\xi = x^2 + y^2$. With $\eta = x$, the equation becomes $u_{\eta\eta} = [2(\xi + \eta)u_\xi + u_\eta]/(\xi - \eta^2)$.

For *elliptic equations* the standard form is $u_{\alpha\alpha} + u_{\beta\beta} = \phi(u, u_\alpha, u_\beta, \alpha, \beta)$, where ξ and η are determined by solving the ξ and η equations of the hyperbolic system (they are complex) and taking $\alpha = (\eta + \xi)/2$, $\beta = (\eta - \xi)/2i (i^2 = -1)$. Since ξ and η are complex conjugates, both α and β are real.

Example 19.5.7. Consider the equation $y^2u_{xx} + x^2u_{yy} = 0$. Clearly, $b^2 - 4ac < 0$, so the equation is elliptic. Then $\xi_x/\xi_y = -ix/y$, $\eta_x/\eta_y = ix/y$, so $\alpha = (\eta + \xi)/2 = y^2$ and $\beta = (\eta - \xi)/2i = x^2$. The standard form is $u_{\alpha\alpha} + u_{\beta\beta} = -(u_\alpha/2\alpha + u_\beta/2\beta)$.

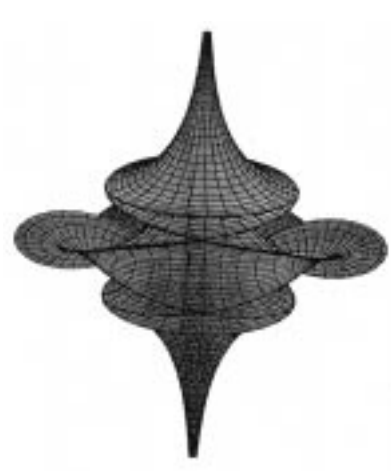


Figure 19.5.2

Figure 19.5.3

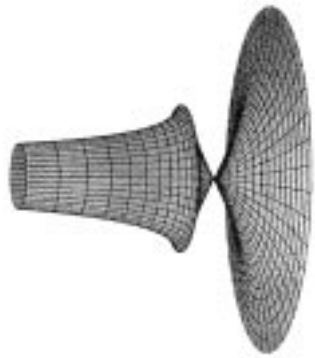


Figure 19.5.4

Figure 19.5.5

FIGURE 19.5.2 to 19.5.5 The mathematical equations used to generate these three-dimensional figures are worth a thousand words. The figures shown illustrate some of the nonlinear ideas of engineering, applied physics, and chemistry. Figure 19.5.2 represents a breather soliton surface for the sine-Gordon equation $w_{uv} = \sin w$ generated by a Backlund transformation. A single-soliton surface for the sine-Gordon equation $w_{uv} = \sin w$ is illustrated in Figure 19.5.3. Figure 19.5.4 represents a single-soliton surface for the Tzitzecia-Dodd-Bullough equation associated with an integrable anisentropic gas dynamics system. Figure 19.5.5 represents a single-soliton Bianchi surface. The solutions to the equations were developed by W. K. Schief and C. Rogers at the Center for Dynamical Systems and Nonlinear Studies at the Georgia Institute of Technology and the University of New South Wales in Sydney, Australia. All of these three-dimensional projections were generated using the MAPLE software package. (Figures courtesy of Schief and Rogers).

Methods of Solution

Separation of Variables. Perhaps the most elementary method for solving linear PDEs with homogeneous boundary conditions is the method of *separation of variables.* To illustrate, consider $u_t - u_{xx} = 0$, $u(x, 0) = f(x)$ (the initial condition) and $u(0, t) = u(1, t) = 0$ for $t > 0$ (the boundary conditions). A solution is assumed in "separated form" $u(x, t) = X(x)T(t)$. Upon substituting into the equation we find $\dot{T}/T = X''/X$ (where $\dot{T} = dT/dt$ and $X'' = d^2X/dx^2$). Since $T = T(t)$ and $X = X(x)$, the ratio must be constant, and for finiteness in t the constant must be negative, say $-\lambda^2$. The solutions of the separated equations $X'' + \lambda^2 X = 0$ with the boundary conditions $X(0) = 0$, $X(1) = 0$, and $\dot{T} = -\lambda^2 T$ are $X = A \sin \lambda x + B \cos \lambda x$ and $T = Ce^{-\lambda^2 t}$, where A, B, and C are arbitrary constants. To satisfy the boundary condition $X(0) = 0$, $B = 0$. An infinite number of values of λ (eigenvalues), say $\lambda_n = n\pi (n = 1, 2, 3, \ldots)$, permit all the eigenfunctions $X_n = b_n \sin \lambda_n x$ to satisfy the other boundary condition $X(1) = 0$. The solution of the

equation and boundary conditions (not the initial condition) is, by superposition, $u(x, t) = \Sigma_{n=1}^{\infty} b_n e^{-n^2\pi^2 t} \cdot \sin n\pi x$ (a Fourier sine series), where the b_n are arbitrary. These values are obtained from the initial condition using the orthogonality properties of the trigonometric function (e.g., $\int_{-\pi}^{\pi} \sin mx \sin n\,x\,d\,x$ is 0 for $m \neq n$ and is π for $m = n \neq 0$) to be $b_n = 2\int_0^1 f(r) \sin n\,\pi\,r\,d\,r$. Then the solution of the problem is $u(x, t) = \Sigma_{n=1}^{\infty}[2\int_0^1 f(r) \sin n\,\pi\,r\,d\,r]\, e^{-n^2\pi^2 t} \sin n\,\pi\,x$, which is a Fourier sine series.

If $f(x)$ is a piecewise smooth or a piecewise continuous function defined for $a \le x \le b$, then its Fourier series within $a \le x \le b$ as its fundamental interval (it is extended periodically ouside that interval) is

$$f(x) \sim \tfrac{1}{2}a_0 + \sum_{n=1}^{\infty} a_n \cos\left[2n\pi x/(b-a)\right] + b_n \sin\left[2n\pi x/(b-a)\right]$$

where

$$a_n = \left[\frac{2}{(b-a)}\right]\int_a^b f(x)\cos\left[2n\pi x/(b-a)\right] dx, \quad n = 0,\ 1,\ \ldots$$

$$b_n = \left[\frac{2}{(b-a)}\right]\int_a^b f(x)\sin\left[2n\pi x/(b-a)\right] dx, \quad n = 1,\ 2,\ \ldots$$

The Fourier sine series has $a_n \equiv 0$, and the Fourier cosine series has $b_n \equiv 0$. The symbol ~ means that the series converges to $f(x)$ at points of continuity, and at the (allowable) points of finite discontinuity the series converges to the *average value* of the discontinuous values.

Caution: This method *only* applies to linear equations with homogeneous boundary conditions. Linear equations with variable coefficients use other orthogonal functions, such as the Besel functions, Laguerre functions, Chebyshev functions, and so forth.

Some inhomogeneous boundary value problems can be transformed into homogeneous ones. Consider the problem $u_t - u_{xx} = 0$, $0 \le x \le 1$, $0 \le t < \infty$ with initial condition $u(x, 0) = f(x)$, and boundary conditions $u(0, t) = g(t)$, $u(1, t) = h(t)$. To homogenize the boundary conditions set $u(x, t) = w(x, t) + x[h(t) - g(t)] + g(t)$ and then solve $w_t - w_x = [\dot{g}(t) - \dot{h}(t)]x - \dot{g}(t)$ with the initial condition $w(x, 0) = f(x) - x[h(0) - g(0)] + g(0)$ and $w(0, t) = w(1, t) = 0$.

Operational Methods. A number of integral transforms are useful for solving a variety of linear problems. To apply the Laplace transform to the problem $u_t - u_{xx} = \delta(x)\,\delta(t)$, $-\infty < x < \infty$, $0 \le t$ with the initial condition $u(x, 0^-) = 0$, where δ is the Dirac delta function, we multiply by e^{-st} and integrate with respect to t from 0 to ∞. With the Laplace transform of $u(x, t)$ denoted by $U(x, s)$ — that is, $U(x, s) = \int_0^{\infty} e^{-st}\,u(x, t)\,dt$ — we have $sU - U_{xx} = \delta(x)$, which has the solution

$$U(x,s) = A(s)e^{-x\sqrt{s}} + B(s)e^{x\sqrt{s}} \qquad \text{for } x > 0$$

$$U(x,s) = C(s)e^{-x\sqrt{s}} + D(s)e^{x\sqrt{s}} \qquad \text{for } x < 0$$

Clearly, $B(s) = C(s) = 0$ for bounded solutions as $|x| \to \infty$. Then, from the boundary condition, $U(0^+, s) - U(0^-, s) = 0$ and integration of $sU - U_{xx} = \delta(x)$ from 0^- to 0^+ gives $U_x(0^+, s) - U(0^-, s) = -1$, so $A = D = 1/2\sqrt{s}$. Hence, $U(x, s) = (1/2\sqrt{s})e^{-\sqrt{s}|x|}$ and the inverse is $u(x, t) = (1/2\pi i)\int_{\Gamma} e^{st}U(x, s)\,ds$, where Γ is a Bromwich path, a vertical line taken to the right of all singularities of U on the sphere.

Similarity (Invariance). This very useful approach is related to dimensional analysis; both have their foundations in group theory. The three important transformations that play a basic role in Newtonian mechanics are translation, scaling, and rotations. Using two independent variables x and t and one dependent variable $u = u(x, t)$, the *translation group* is $\bar{x} = x + \alpha a$, $\bar{t} = t + \beta a$, $\bar{u} = u + \gamma a$; the *scaling*

group is $\bar{x} = a^{\alpha}x$, $\bar{t} = a^{\beta}t$, and $\bar{u} = a^{\gamma}u$; the *rotation group* is $\bar{x} = x\cos a + t\sin a$, $\bar{t} = t\cos a - x\sin a$, $\bar{u} = u$, with a nonnegative real number a. Important in which follows are the *invariants* of these groups. For the translation group there are two $\eta = x - \lambda t$, $\lambda = \alpha/\beta$, $f(\eta) = u - \varepsilon t$, $\varepsilon = \gamma/\beta$ or $f(\eta) = u - \theta x$, $\theta = \gamma/\alpha$; for the scaling group the invariants are $\eta = x/t^{\alpha/\beta}$ (or $t/x^{\beta/\alpha}$) and $f(\eta) = u/t^{\gamma/\beta}$ (or $u/x^{\gamma/\alpha}$); for the rotation group the invariants are $\eta = x^2 + t^2$ and $u = f(\eta) = f(x^2 + t^2)$.

If a PDE and its data (initial and boundary conditions) are left invariant by a transformation group, then similar (invariant) solutions are sought using the invariants. For example, if an equation is left invariant under scaling, then solutions are sought of the form $u(x, t) = t^{\gamma/\beta} f(\eta)$, $\eta = xt^{-\alpha/\beta}$ or $u(x, t) = x^{\gamma/\alpha} f(tx^{-\beta/\alpha})$; invariance under translation gives solutions of the form $u(x, t) = f(x - \lambda t)$; and invariance under rotation gives rise to solutions of the form $u(x, t) = f(x^2 + t^2)$.

Examples of invariance include the following:

1. The equation $u_{xx} + u_{yy} = 0$ is invariant under rotation, so we search for solutions of the form $u = f(x^2 + y^2)$. Substitution gives the ODE $f' + \eta f'' = 0$ or $(\eta f')' = 0$. The solution is $u(x, t) = c \ln \eta = c \ln(x^2 + t^2)$, which is the (so-called) fundamental solution of Laplace's equation.
2. The nonlinear diffusion equation $u_t = (u^n u_x)_x$ $(n > 0)$, $0 \le x$, $0 \le t$, $u(0, t) = ct^n$ is invariant under scaling with the similar form $u(x, t) = t^n f(\eta)$, $\eta = xt^{-(n+1)/2}$. Substituting into the PDE gives the equation $(f^n f')' + ((n + 1)/2)\eta f' - nf = 0$, with $f(0) = c$ and $f(\infty) = 0$. Note that the equation is an ODE.
3. The wave equation $u_{xx} - u_{tt} = 0$ is invariant under translation. Hence, solutions exist of the form $u = f(x - \lambda t)$. Substitution gives $f''(1 - \lambda^2) = 0$. Hence, $\lambda = \pm 1$ or f is linear. Rejecting the trivial linear solution we see that $u = f(x - t) + g(x + t)$, which is the general (d'Alembert) solution of the wave equation; the quantities $x - t = \alpha$, $x + t = \beta$ are the characteristics of the next section.

The construction of all transformations that leave a PDE invariant is a solved problem left for the references.

The study of "solitons" (solitary traveling waves with special properties) has benefited from symmetry considerations. For example, the nonlinear third-order (Korteweg-de Vries) equation $u_t + uu_x - au_{xxx} = 0$ is invariant under translation. Solutions are sought of the form $u = f(x - \lambda t)$, and f satisfies the ODE, in $\eta = x - \lambda t$, $-\lambda f' + ff' - af''' = 0$.

Characteristics. Using the characteristics the solution of the hyperbolic problem $u_{tt} - u_{xx} = p(x, t)$, $-\infty < x < \infty$, $0 \le t$, $u(x, 0) = f(x)$, $u_t(x, 0) = h(x)$ is

$$u(x,t) = \tfrac{1}{2}\int_0^t d\tau \int_{x-(t-\tau)}^{x+(t-\tau)} p(\xi,\tau)\, d\xi + \tfrac{1}{2}\int_{x-t}^{x+t} h(\xi)\, d\xi + \tfrac{1}{2}\big[f(x+t) + f(x-t)\big]$$

The solution of $u_{tt} - u_{xx} = 0$, $0 \le x < \infty$, $0 \le t < \infty$, $u(x, 0) = 0$, $u_t(x, 0) = h(x)$, $u(0, t) = 0$, $t > 0$ is $u(x, t) = {}^1/_2 \int_{-x+t}^{x+t} h(\xi)\, d\xi$.

The solution of $u_{tt} - u_{xx} = 0$, $0 \le x < \infty$, $0 \le t < \infty$, $u(x, 0) = 0$, $u_t(x, 0) = 0$, $u(0, t) = g(t)$, $t > 0$ is

$$u(x,t) = \begin{cases} 0 & \text{if } t < x \\ g(t-x) & \text{if } t > x \end{cases}$$

From time to time, lower-order derivatives appear in the PDE in use. To remove these from the equation $u_{tt} - u_{xx} + au_x + bu_t + cu = 0$, where a, b, and c are constants, set $\xi = x + t$, $\mu = t - x$, whereupon $u(x, t) = u[(\xi - \mu)/2, (\xi + \mu)/2] = U(\xi, \mu)$, where $U_{\xi\mu} + [(b + a)/4]\, U_{\xi} + [(b - a)/4]\, U_{\mu} + (c/4)U = 0$. The transformation $U(\xi, \mu) = W(\xi, \mu) \exp[-(b - a)\xi/4 - (b + a)\mu/4]$ reduces to satisfying $W_{\xi\mu} + \lambda W = 0$, where $\lambda = (a^2 - b^2 + 4c)/16$. If $\lambda \ne 0$, we lose the simple d'Alembert solution. But the equation for W is still easier to handle.

In linear problems discontinuities propagate along characteristics. In nonlinear problems the situation is usually different. The characteristics are often used as new coordinates in the numerical method of characteristics.

Green's Function. Consider the diffusion problem $u_t - u_{xx} = \delta(t)\delta(x - \xi)$, $0 \le x < \infty$, $\xi > 0$, $u(0, t) = 0$, $u(x, 0) = 0$ [$u(\infty, t) = u(\infty, 0) = 0$], a problem that results from a unit source somewhere in the domain subject to a homogeneous (zero) boundary condition. The solution is called a *Green's function of the first kind.* For this problem there is $G_1(x, \xi, t) = F(x - \xi, t) - F(x + \xi, t)$, where $F(x, t) = e^{-x^2/4t}/\sqrt{4\pi t}$ is the *fundamental* (invariant) *solution*. More generally, the solution of $u_t - u_{xx} = \delta(x - \xi)\,\delta(t - \tau)$, $\xi > 0$, $\tau > 0$, with the same conditions as before, is the Green's function of the first kind.

$$G_1(x,\xi,t-\tau) = \frac{1}{\sqrt{4\pi(t-\tau)}}\left[e^{-(x-\xi)^2/4(t-\tau)} - e^{-(x+\xi)^2/4(t-\tau)}\right]$$

for the semi-infinite interval.

The solution of $u_t - u_{xx} = p(x, t)$, $0 \le x < \infty$, $0 \le t < \infty$, with $u(x, 0) = 0$, $u(0, t) = 0$, $t > 0$ is $u(x, t) = \int_0^t d\tau \int_0^\infty p(\xi, \tau)G_1(x, \xi, t - \tau)]\, d\xi$, which is a superposition. Note that the Green's function and the desired solution must both satisfy a zero boundary condition at the origin for this solution to make sense.

The solution of $u_t - u_{xx} = 0$, $0 \le x < \infty$, $0 \le t < \infty$, $u(x, 0) = f(x)$, $u(0, t) = 0$, $t > 0$ is $u(x, t) = \int_0^\infty f(\xi)G_1(x, \xi, t)\, d\xi$.

The solution of $u_t - u_{xx} = 0$, $0 \le x < \infty$, $0 \le t < \infty$, $u(x, 0) = 0$, $u(0, t) = g(t)$, $t > 0$ (nonhomogeneous) is obtained by transforming to a new problem that has a homogeneous boundary condition. Thus, with $w(x, t) = u(x, t) - g(t)$ the equation for w becomes $w_t - w_{xx} = -\dot{g}(t) - g(0)\,\delta(t)$ and $w(x, 0) = 0$, $w(0, t) = 0$. Using G_1 above, we finally obtain $u(x, t) = (x/\sqrt{4\pi})\int_0^t g(t - \tau)e^{-x^2/4t}/\tau^{3/2}\, d\tau$.

The Green's function approach can also be employed for elliptic and hyperbolic problems.

Equations in Other Spatial Variables. The sperically symmetric wave equation $u_{rr} + 2u_r/r - u_{tt} = 0$ has the general solution $u(\tau, t) = [f(t - r) + g(t + r)]/r$.

The Poisson-Euler-Darboux equation, arising in gas dynamics,

$$u_{rs} + N\left(u_r + u_s\right)/(r+s) = 0$$

where N is a positive integer ≥ 1, has the general solution

$$u(r,s) = k + \frac{\partial^{N-1}}{\partial r^{N-1}}\left[\frac{f(r)}{(r+s)^N}\right] + \frac{\partial^{N-1}}{\partial s^{N-1}}\left[\frac{g(s)}{(r+s)^N}\right]$$

Here, k is an arbitrary constant and f and g are arbitrary functions whose form is determined from the problem initial and boundary conditions.

Conversion to Other Orthogonal Coordinate Systems. Let (x^1, x^2, x^3) be rectangular (Cartesian) coordinates and (u^1, u^2, u^3) be any orthogonal coordinate system related to the rectangular coordinates by $x^i = x^i(u^1, u^2, u^3)$, $i = 1, 2, 3$. With $(ds)^2 = (dx^1)^2 + (dx^2)^2 + (dx^3)^2 = g_{11}(du^1)^2 + g_{22}(du^2)^2 + g_{33}(du^3)^2$, where $g_{ii} = (\partial x^1/\partial u^i)^2 + (\partial x^2/\partial u^i)^2 + (\partial x^3/\partial u^i)^2$. In terms of these "metric" coefficients the basic operations of applied mathematics are expressible. Thus (with $g = g_{11}g_{22}g_{33}$)

$$dA = \left(g_{11}g_{22}\right)^{1/2} du^1\, du^2; \qquad dV = \left(g_{11}g_{22}g_{33}\right)^{1/2} du^1\, du^2\, du^3$$

$$\text{grad}\,\phi = \frac{\vec{a}_1}{(g_{11})^{1/2}}\frac{\partial\phi}{\partial u^1} + \frac{\vec{a}_2}{(g_{22})^{1/2}}\frac{\partial\phi}{\partial u^2} + \frac{\vec{a}_3}{(g_{33})^{1/2}}\frac{\partial\phi}{\partial u^3}$$

($\vec{a}_i$ are unit vectors in direction i);

$$\text{div}\,\vec{E} = g^{-1/2}\left\{\frac{\partial}{\partial u^1}\left[(g_{22}g_{33})^{1/2}E_1\right] + \frac{\partial}{\partial u^2}\left[(g_{11}g_{33})^{1/2}E_2\right] + \frac{\partial}{\partial u^3}\left[(g_{11}g_{22})^{1/2}E_3\right]\right\}$$

[here $\vec{E} = (E_1, E_2, E_3)$];

$$\begin{aligned}\text{curl}\,\vec{E} = g^{-1/2}\Bigg\{&\vec{a}_1(g_{11})^{1/2}\left(\frac{\partial}{\partial u^2}\left[(g_{33})^{1/2}E_3\right] - \frac{\partial}{\partial u^3}\left[(g_{22})^{1/2}E_2\right]\right)\\ &+ \vec{a}_2(g_{22})^{1/2}\left(\frac{\partial}{\partial u^3}\left[(g_{11})^{1/2}E_1\right] - \frac{\partial}{\partial u^1}\left[(g_{33})^{1/2}E_3\right]\right)\\ &+ \vec{a}_3(g_{33})^{1/2}\left(\frac{\partial}{\partial u^1}\left[(g_{22})^{1/2}E_2\right] - \frac{\partial}{\partial u^2}\left[(g_{11})^{1/2}E_1\right]\right)\Bigg\}\end{aligned}$$

$$\text{div grad}\,\psi = \nabla^2\psi = \text{Laplacian of } \psi = g^{-1/2}\sum_{i=1}^{3}\frac{\partial}{\partial u^i}\left[\frac{g^{1/2}}{g_{ii}}\frac{\partial\psi}{\partial u^i}\right]$$

Table 19.5.2 shows some coordinate systems.

Table 19.5.2 Some Coordinate Systems

Coordinate System		Metric Coefficients
Circular Cylindrical		
$x = r\cos\theta$	$u^1 = r$	$g_{11} = 1$
$y = r\sin\theta$	$u^2 = \theta$	$g_{22} = r^2$
$z = z$	$u^3 = z$	$g_{33} = 1$
Spherical		
$x = r\sin\psi\cos\theta$	$u^1 = r$	$g_{11} = 1$
$y = r\sin\psi\sin\theta$	$u^2 = \psi$	$g_{22} = r^2$
$z = r\cos\psi$	$u^3 = \theta$	$g_{33} = r^2\sin^2\psi$
Parabolic Coordinates		
$x = \mu\nu\cos\theta$	$u^1 = \mu$	$g_{11} = \mu^2 + \nu^2$
$y = \mu\nu\sin\theta$	$u^2 = \nu$	$g_{22} = \mu^2 + \nu^2$
$z = 1/2\,(\mu^2 - \nu^2)$	$u^3 = \theta$	$g_{33} = \mu^2\nu^2$

Other metric coefficients and so forth can be found in Moon and Spencer [1961].

References

Ames, W. F. 1965. *Nonlinear Partial Differential Equations in Science and Engineering, Volume 1.* Academic Press, Boston, MA.

Ames, W. F. 1972. *Nonlinear Partial Differential Equations in Science and Engineering, Volume 2.* Academic Press, Boston, MA.

Brauer, F. and Nohel, J. A. 1986. *Introduction to Differential Equations with Applications,* Harper & Row, New York.
Jeffrey, A. 1990. *Linear Algebra and Ordinary Differential Equations,* Blackwell Scientific, Boston, MA.
Kevorkian, J. 1990. *Partial Differential Equations.* Wadsworth and Brooks/Cole, Belmont, CA.
Moon, P. and Spencer, D. E. 1961. *Field Theory Handbook,* Springer, Berlin.
Rogers, C. and Ames, W. F. 1989. *Nonlinear Boundary Value Problems in Science and Engineering.* Academic Press, Boston, MA.
Whitham, G. B. 1974. *Linear and Nonlinear Waves.* John Wiley & Sons, New York.
Zauderer, E. 1983. *Partial Differential Equations of Applied Mathematics.* John Wiley & Sons, New York.
Zwillinger, D. 1992. *Handbook of Differential Equations.* Academic Press, Boston, MA.

Further Information

A collection of solutions for linear and nonlinear problems is found in E. Kamke, *Differential-gleichungen-Lösungsmethoden und Lösungen,* Akad. Verlagsges, Leipzig, 1956. Also see G. M. Murphy, *Ordinary Differential Equations and Their Solutions,* Van Nostrand, Princeton, NJ, 1960 and D. Zwillinger, *Handbook of Differential Equations,* Academic Press, Boston, MA, 1992. For nonlinear problems see

Ames, W. F. 1968. *Ordinary Differential Equations in Transport Phenomena.* Academic Press, Boston, MA.
Cunningham, W. J. 1958. *Introduction to Nonlinear Analysis.* McGraw-Hill, New York.
Jordan, D. N. and Smith, P. 1977. *Nonlinear Ordinary Differential Equations.* Clarendon Press, Oxford, UK.
McLachlan, N. W. 1955. *Ordinary Non-Linear Differential Equations in Engineering and Physical Sciences,* 2nd ed. Oxford University Press, London.
Zwillinger, D. 1992.

19.6 Integral Equations

William F. Ames

Classification and Notation

Any equation in which the unknown function $u(x)$ appears under the integral sign is called an *integral equation*. If $f(x)$, $K(x, t)$, a, and b are known then the integral equation for u, $\int_a^b K(x, t)\, u(t)\, dt = f(x)$ is called a *linear integral equation of the first kind of Fredholm type*. $K(x, t)$ is called the *kernel function* of the equation. If b is replaced by x (the independent variable) the equation is an equation of *Volterra type of the first kind*.

An equation of the form $u(x) = f(x) + \lambda \int_a^b K(x, t)u(t)\, dt$ is said to be a linear integral equation of *Fredholm type of the second kind*. If b is replaced by x it is of *Volterra type*. If $f(x)$ is not present the equation is homogeneous.

The equation $\phi(x)\, u(x) = f(x) + \lambda \int_a^{b \text{ or } x} K(x, t)u(t)\, dt$ is the *third kind equation* of Fredholm or Volterra type. If the unknown function u appears in the equation in any way other than to the first power then the integral equation is said to be *nonlinear*. Thus, $u(x) = f(x) + \int_a^b K(x, t) \sin u(t)\, dt$ is nonlinear. An integral equation is said to be *singular* when either or both of the limits of integration are infinite or if $K(x, t)$ becomes infinite at one or more points of the integration interval.

Example 19.6.1. Consider the singular equations $u(x) = x + \int_0^\infty \sin(x\, t)\, u(t)\, dt$ and $f(x) = \int_0^x [u(t)/(x - t)^2]\, dt$.

Relation to Differential Equations

The *Leibnitz rule* $(d/dx) \int_{a(x)}^{b(x)} F(x, t)\, dt \int_{a(x)}^{b(x)} (\partial F/\partial x)\, dt + F[x, b(x)](db/dx) - F[x, a(x)] \times (da/dx)$ is useful for differentiation of an integral involving a parameter (x in this case). With this, one can establish the relation

$$I_n(x) = \int_a^x (x - t)^{n-1} f(t)\, dt = (n-1)! \underbrace{\int_a^x \cdots \int_a^x}_{n \text{ times}} f(x) \underbrace{dx \ldots dx}_{n \text{ times}}$$

This result will be used to establish the relation of the second-order initial value problem to a Volterra integral equation.

The second-order differential equation $y''(x) + A(x)y'(x) + B(x)y = f(x), y(a) = y_0, y'(a) = y_0'$ is equivalent to the integral equations

$$y(x) = -\int_a^x \left\{A(t) + (x - t)\left[B(t) - A'(t)\right]\right\} y(t)\, dt + \int_a^x (x - t) f(t)\, dt + \left[A(a)y_0 + y_0'\right](x - a) + y_0$$

which is of the type $(x)y = \int_a^x K(x, t)y(t)\, dt + F(x)$ where $K(x, t) = (t - x)[B(t) - A'(t)] - A(t)$ and $F(x)$ includes the rest of the terms. Thus, this initial value problem is equivalent to a Volterra integral equation of the second kind.

Example 19.6.2. Consider the equation $y'' + x^2y' + xy = x$, $y(0) = 1$, $y'(0) = 0$. Here $A(x) = x^2$, $B(x) = x$, $f(x) = x$, $a = 0$, $y_0 = 1$, $y_0' = 0$. The integral equation is $y(x) = \int_0^x t(x - 2t)y(t)\, dt + (x^3/6) + 1$.

The expression for $I_n(x)$ can also be useful in converting boundary value problems to integral equations. For example, the problem $y''(x) + \lambda y = 0$, $y(0) = 0$, $y(a) = 0$ is equivalent to the Fredholm equation $y(x) = \lambda \int_0^a K(x, t)y(t)\, dt$, where $K(x, t) = (t/a)(a - x)$ when $t < x$ and $K(x, t) = (x/a)(a - t)$ when $t > x$.

In both cases the differential equation can be recovered from the integral equation by using the Leibnitz rule.

Nonlinear differential equations can also be transformed into integral equations. In fact this is one method used to establish properties of the equation and to develop approximate and numerical solutions. For example, the "forced pendulum" equation $y''(x) + a^2 \sin y(x) = f(x)$, $y(0) = y(1) = 0$ transforms into the nonlinear Fredholm equation.

$$y(x) = \int_0^1 K(x,t)\left[a^2 \sin y(t) - f(t)\right] dt$$

with $K(x, t) = x(1 - t)$ for $0 < x < t$ and $K(x, t) = t(1 - x)$ for $t < x < 1$.

Methods of Solution

Only the simplest integral equations can be solved exactly. Usually approximate or numerical methods are employed. The advantage here is that integration is a "smoothing operation," whereas differentiation is a "roughening operation." A few exact and approximate methods are given in the following sections. The numerical methods are found under 19.12.

Convolution Equations

The special convolution equation $y(x) = f(x) + \lambda \int_0^x K(x - t)y(t)\, dt$ is a special case of the Volterra equation of the second kind. $K(x - t)$ is said to be a *convolution kernel*. The integral part is the convolution integral discussed under 19.8. The solution can be accomplished by transforming with the Laplace transform: $L[y(x)] = L[f(x)] + \lambda L[y(x)]L[K(x)]$ or $y(x) = L^{-1}\{L[f(x)]/(1 - \lambda L[K(x)])\}$.

Abel Equation

The Volterra equation $f(x) = \int_0^x y(t)/(x - t)^\alpha\, dt$, $0 < \alpha < 1$ is the (singular) Abel equation. Its solution is $y(x) = (\sin \alpha\pi/\pi)(d/dx) \int_0^x F(t)/(x - t)^{1-\alpha} dt$.

Approximate Method (Picard's Method)

This method is one of successive approximations that is described for the equation $y(x) = f(x) + \lambda \int_a^x K(x, t)y(t)\, dt$. Beginning with an initial guess $y_0(t)$ (often the value at the initial point a) generate the next approximation with $y_1(x) = f(x) + \int_a^x K(x, t)y_0(t)\, dt$ and continue with the general iteration

$$y_n(x) = f(x) + \lambda \int_a^x K(x,t)y_{n-1}(t)\, dt$$

Then, by iterating, one studies the convergence of this process, as is described in the literature.

Example 19.6.3. Let $y(x) = 1 + \int_0^x xt[y(t)]^2\, dt$, $y(0) = 1$, With $y_0(t) = 1$ we find $y_1(x) = 1 + \int_0^x xt\, dt = 1 + (x^3/2)$ and $y_2(x) = 1 + \int_0^x xt[1 + (t^3/2)^2 dt$, and so forth.

References

Jerri, A. J. 1985. *Introduction to Integral Equations with Applications*, Marcel Dekker, New York.
Tricomi, F. G. 1958. *Integral Equations.* Wiley-Interscience, New York.
Yosida, K. 1960. *Lectures on Differential and Integral Equations.* Wiley-Interscience, New York.

19.7 Approximation Methods

William F. Ames

The term *approximation methods* usually refers to an analytical process that generates a symbolic approximation rather than a numerical one. Thus, $1 + x + x^2/2$ is an approximation of e^x for small x. This chapter introduces some techniques for approximating the solution of various operator equations.

Perturbation

Regular Perturbation

This procedure is applicable to *some* equations in which a small parameter, ε, appears. Use this procedure with care; the procedure involves expansion of the dependent variables and data in a power series in the small parameter. The following example illustrates the procedure.

Example 19.7.1. Consider the equation $y'' + \varepsilon y' + y = 0$, $y(0) = 1$, $y'(0) = 0$. Write $y(x; \varepsilon) = y_0(x) + \varepsilon y_1(x) + \varepsilon^2 y_2(x) + \cdots$, and the initial conditions (data) become

$$y_0(0) + \varepsilon y_1(0) + \varepsilon^2 y_2(0) + \cdots = 1$$

$$y_0'(0) + \varepsilon y_1'(0) + \varepsilon^2 y_2'(0) + \cdots = 0$$

Equating like powers of ε in all three equations yields the sequence of equations

$$O(\varepsilon^0): y_0'' + y_0 = 0, \qquad y_0(0) = 1, \qquad y_0'(0) = 0$$

$$O(\varepsilon^1): y_1'' + y_1 = -y_0', \qquad y_1(0) = 0, \qquad y_1'(0) = 0$$

$$\vdots$$

The solution for y_0 is $y_0 = \cos x$ and using this for y_1 we find $y_1(x) = 1/2 (\sin x - x \cos x)$. So $y(x; \varepsilon) = \cos x + \varepsilon(\sin x - x \cos x)/2 + O(\varepsilon^2)$. Appearance of the term $x \cos x$ indicates a *secular term* that becomes arbitrarily large as $x \to \infty$. Hence, this approximation is valid only for $x \ll 1/\varepsilon$ and for small ε. If an approximation is desired over a larger range of x then the method of multiple scales is required.

Singular Perturbation

The *method of multiple scales* is a singular method that is *sometimes* useful if the regular perturbation method fails. In this case the assumption is made that the solution depends on *two* (or more) different length (or time) scales. By trying various possibilities, one can determine those scales. The scales are treated as dependent variables when transforming the given ordinary differential equation into a partial differential equation, but then the scales are treated as independent variables when solving the equations.

Example 19.7.2. Consider the equation $\varepsilon y'' + y' = 2$, $y(0) = 0$, $y(1) = 1$. This is singular since (with $\varepsilon = 0$) the resulting first-order equation cannot satisfy both boundary conditions. For the problem the proper length scales are $u = x$ and $v = x/\varepsilon$. The second scale can be ascertained by substituting $\varepsilon^n x$ for x and requiring $\varepsilon y''$ and y' to be of the same order in the transformed equation. Then

$$\frac{d}{dx} = \frac{\partial}{\partial u}\frac{du}{dx} + \frac{\partial}{\partial v}\frac{dv}{dx} = \frac{\partial}{\partial u} + \frac{1}{\varepsilon}\frac{\partial}{\partial v}$$

and the equation becomes

$$\varepsilon\left(\frac{\partial}{\partial u}+\frac{1}{\varepsilon}\frac{\partial}{\partial v}\right)^2 y+\left(\frac{\partial}{\partial u}+\frac{1}{\varepsilon}\frac{\partial}{\partial v}\right)y=2$$

With $y(x;\varepsilon)=y_0(u,v)+\varepsilon y_1(u,v)+\varepsilon^2 y_2(u,v)+\cdots$ we have terms

$$O\left(\varepsilon^{-1}\right):\frac{\partial^2 y_0}{\partial v^2}+\frac{\partial y_0}{\partial v}=0 \qquad \left(\text{actually ODEs with parameter } u\right)$$

$$O\left(\varepsilon^{0}\right):\frac{\partial^2 y_1}{\partial v^2}+\frac{\partial y_1}{\partial v}=2-2\frac{\partial^2 y_0}{\partial u\,\partial v}-\frac{\partial y_0}{\partial u}$$

$$O\left(\varepsilon^{1}\right):\frac{\partial^2 y_2}{\partial v^2}+\frac{\partial y_2}{\partial v}=-2\frac{\partial^2 y_1}{\partial u\,\partial v}-\frac{\partial y_1}{\partial u}-\frac{\partial^2 y_0}{\partial u^2}$$

$$\vdots$$

Then $y_0(u, v) = A(u) + B(u)e^{-v}$ and so the second equation becomes $\partial^2 y_1/\partial v^2 + \partial y_1/\partial v = 2 - A'(u) + B'(u)e^{-v}$, with the solution $y_1(u, v) = [2 - A'(u)]v + vB'(u)e^{-v} + D(u) + E(u)e^{-v}$. Here A, B, D and E are still arbitrary. Now the solvability condition — "higher order terms must vanish no slower (as $\varepsilon \to 0$) than the previous term" (Kevorkian and Cole, 1981) — is used. For y_1 to vanish no slower than y_0 we must have $2 - A'(u) = 0$ and $B'(u) = 0$. If this were not true the terms in y_1 would be larger than those in y_0 ($v \gg 1$). Thus $y_0(u, v) = (2u + A_0) + B_0e^{-v}$, or in the original variables $y(x; \varepsilon) \approx (2x + A_0) + B_0e^{-x/\varepsilon}$ and matching to both boundary conditions gives $y(x; \varepsilon) \approx 2x - (1 - e^{-x/\varepsilon})$.

Boundary Layer Method

The boundary layer method is applicable to regions in which the solution is *rapidly varying.* See the references at the end of the chapter for detailed discussion.

Iterative Methods

Taylor Series

If it is known that the solution of a differential equation has a power series in the independent variable (t), then we may proceed from the initial data (the easiest problem) to compute the Taylor series by differentiation.

Example 19.7.3. Consider the equation $(d^2x/dt) = -x - x^2$, $x(0) = 1$, $x'(0) = 1$. From the differential equation, $x''(0) = -2$, and, since $x''' = -x' - 2xx'$, $x'''(0) = -1 - 2 = -3$, so the four term approximation for $x(t) \approx 1 + t - (2t^2/2!) - (3t^3/3!) = 1 + t - t^2 - t^3/2$. An estimate for the error at $t = t_1$, (see a discussion of series methods in any calculus text) is not greater than $|d^4x/dt^4|_{\max}[(t_1)^4/4!]$, $0 \le t \le t_1$.

Picard's Method

If the vector differential equation $x' = f(t, x)$, $x(0)$ given, is to be approximated by Picard iteration, we begin with an initial guess $x_0 = x(0)$ and calculate iteratively $x_i' = f(t, x_{i-1})$.

Example 19.7.4. Consider the equation $x' = x + y^2$, $y' = y - x^3$, $x(0) = 1$, $y(0) = 2$. With $x_0 = 1$, $y_0 = 2$, $x_1' = 5$, $y_1' = 1$, so $x_1 = 5t + 1$, $y_1 = t + 2$, since $x_i(0) = 1$, $y_i(0) = 2$ for $i \ge 0$. To continue, use $x_{i+1}' = x_i + y_i^2$, $y_{i+1}' = y_i - x_i^3$. A modification is the utilization of the first calculated term immediately in the second equation. Thus, the calculated value of $x_1 = 5t + 1$, when used in the second equation, gives $y_1' = y_0 - (5t + 1)^3 = 2 - (125t^3 + 75t^2 + 15t + 1)$, so $y_1 = 2t - (125t^4/4) - 25t^3 - (15t^2/2) - t + 2$. Continue with the iteration $x_{i+1}' = x_i + y_i^3$, $y_{i+1}' = y_i - (x_{i+1})^3$.

Another variation would be $x_{i+1}' = x_{i+1} + (y_i)^3$, $y_{i+1}' = y_{i+1} - (x_{i+1})^3$.

References

Ames, W. F. 1965. *Nonlinear Partial Differential Equations in Science and Engineering, Volume I.* Academic Press, Boston, MA.

Ames, W. F. 1968. *Nonlinear Ordinary Differential Equations in Transport Processes.* Academic Press, Boston, MA.

Ames, W. F. 1972. *Nonlinear Partial Differential Equations in Science and Engineering, Volume II.* Academic Press, Boston, MA.

Kevorkian, J. and Cole, J. D. 1981. *Perturbation Methods in Applied Mathematics,* Springer, New York.

Miklin, S. G. and Smolitskiy, K. L. 1967. *Approximate Methods for Solutions of Differential and Integral Equations.* Elsevier, New York.

Nayfeh, A. H. 1973. *Perturbation Methods.* John Wiley & Sons, New York.

Zwillinger, D. 1992. *Handbook of Differential Equations,* 2nd ed. Academic Press, Boston, MA.

19.8 Integral Transforms

William F. Ames

All of the integral transforms are special cases of the equation $g(s) = \int_a^b K(s, t)f(t)d\,t$, in which $g(s)$ is said to be the *transform* of $f(t)$, and $K(s, t)$ is called the *kernel* of the transform. Table 19.8.1 shows the more important kernels and the corresponding intervals (a, b).

Details for the first three transforms listed in Table 19.8.1 are given here. The details for the other are found in the literature.

Laplace Transform

The Laplace transform of $f(t)$ is $g(s) = \int_0^\infty e^{-st} f(t)\,dt$. It may be thought of as transforming one class of functions into another. The advantage in the operation is that under certain circumstances it replaces complicated functions by simpler ones. The notation $L[f(t)] = g(s)$ is called the *direct transform* and $L^{-1}[g(s)] = f(t)$ is called the *inverse transform.* Both the direct and inverse transforms are tabulated for many often-occurring functions. In general $L^{-1}[g(s)] = (1/2\,\pi i)\int_{\alpha-i\infty}^{\alpha+i\infty} e^{st}g(s)\,ds$, and to evaluate this integral requires a knowledge of complex variables, the theory of residues, and contour integration.

Properties of the Laplace Transform

Let $L[f(t)] = g(s)$, $L^{-1}[g(s)] = f(t)$.

1. The Laplace transform may be applied to a function $f(t)$ if $f(t)$ is continuous or piecewise continuous; if $t^n|f(t)|$ is finite for all t, $t \to 0$, $n < 1$; and if $e^{-at}|f(t)|$ is finite as $t \to \infty$ for some value of a, $a > 0$.
2. L and L^{-1} are unique.
3. $L[af(t) + bh(t)] = aL[f(t)] + bL[h(t)]$ (linearity).
4. $L[e^{at}f(t)] = g(s - a)$ (shift theorem).
5. $L[(-t)^k f(t)] = d^k g/ds^k$; k a positive integer.

Example 19.8.1. $L[\sin a\,t] = \int_0^\infty e^{-st} \sin a\,t\,d\,t = a/(s^2 + a^2)$, $s > 0$. By property 5,

$$\int_0^\infty e^{-st} t\sin at\,dt = L[t\sin at] = \frac{2as}{s^2 + a^2}$$

Table 19.8.1 Kernels and Intervals of Various Integral Transforms

Name of Transform	(a, b)	$K(s, t)$
Laplace	$(0, \infty)$	e^{-st}
Fourier	$(-\infty, \infty)$	$\frac{1}{\sqrt{2\pi}} e^{-ist}$
Fourier cosine	$(0, \infty)$	$\sqrt{\frac{2}{\pi}} \cos st$
Fourier sine	$(0, \infty)$	$\sqrt{\frac{2}{\pi}} \sin st$
Mellin	$(0, \infty)$	t^{s-1}
Hankel	$(0, \infty)$	$tJ_\nu(st)$, $\nu \geq -\frac{1}{2}$

6.
$$L[f'(t)] = sL[f(t)] - f(0)$$
$$L[f''(t)] = s^2 L[f(t)] - sf(0) - f'(0)$$
$$\vdots$$
$$L\left[f^{(n)}(t)\right] = s^n L[f(t)] - s^{n-1} f(0) - \cdots - sf^{(n-2)}(0) - f^{(n-1)}(0)$$

In this property it is apparent that the initial data are automatically brought into the computation.

Example 19.8.2. Solve $y'' + y = e^t$, $y(0) = 1$, $y'(0) = 1$. Now $L[y''] = s^2L[y] - sy(0) - y'(0) = s^2L[y] - s - 1$. Thus, using the linear property of the transform (property 3), $s^2L[y] + L[y] - s - 1 = L[e^t] = 1/(s - 1)$. Therefore, $L[y] = s^2/[(s - 1)(s^2 + 1)]$.

With the notations $\Gamma(n + 1) = \int_0^\infty x^n e^{-x}\, dx$ (gamma function) and $J_n(t)$ the Bessel function of the first kind of order n, a short table of Laplace transforms is given in Table 19.8.2.

7. $$L\left[\int_a^t f(t)\, dt\right] = \tfrac{1}{s} L[f(t)] + \tfrac{1}{s}\int_a^0 f(t)\, dt.$$

Example 19.8.3. Find $f(t)$ if $L[f(t)] = (1/s^2)[1/(s^2 - a^2)]$. $L[1/a \sinh a\ t] = 1/(s^2 - a^2)$. Therefore, $f(t) = \int_0^t [\int_0^t \frac{1}{a} \sinh a\ t\ d\ t]d\ t = 1/a^2[(\sinh a\ t)/a - t]$.

$$L\left[\frac{f(t)}{t}\right] = \int_s^\infty g(s)\, ds; \qquad L\left[\frac{f(t)}{t^k}\right] = \underbrace{\int_s^\infty \cdots \int_s^\infty}_{k \text{ integrals}} g(s)(ds)^k$$

Example 19.8.4. $L[(\sin a\ t)/t] = \int_s^\infty L[\sin a\ t]d\ s = \int_s^\infty [a\ d\ s/(s^2 + a^2)] = \cot^{-1}(s/a)$.

9. The *unit step function* $u(t - a) = 0$ for $t < a$ and 1 for $t > a$. $L[u(t - a) = e^{-as}/s$.
10. The *unit impulse function* is $\delta(a) = u'(t - a) = 1$ at $t = a$ and 0 elsewhere. $L[u'(t - a) = e^{-as}$.
11. $L^{-1}[e^{-as}g(s)] = f(t - a)u(t - a)$ (second shift theorem).
12. If $f(t)$ is *periodic* of period b — that is, $f(t + b) = f(t)$ — then $L[f(t)] = [1/(1 - e^{-bs})] \times \int_0^b e^{-st} f(t)\, dt$.

Example 19.8.5. The equation $\partial^2 y/(\partial t \partial x) + \partial y/\partial t + \partial y/\partial x = 0$ with $(\partial y/\partial x)(0, x) = y(0, x) = 0$ and $y(t, 0) + (\partial y/\partial t)(t, 0) = \delta(0)$ (see property 10) is solved by using the Laplace transform of y with respect to t. With $g(s, x) = \int_0^\infty e^{-st} y(t, x)\, dt$, the transformed equation becomes

Table 19.8.2 Some Laplace Transforms

$f(t)$	$g(s)$	$f(t)$	$g(s)$
1	$\frac{1}{s}$	$e^{-at}(1-a\,t)$	$\frac{s}{(s+a)^2}$
t^n, n is a + integer	$\frac{n!}{s^{n+1}}$	$\frac{t\sin at}{2a}$	$\frac{s}{\left(s^2+a^2\right)^2}$
t^n, $n \neq$ a + integer	$\frac{\Gamma(n+1)}{s^{n+1}}$	$\frac{1}{2a^2}\sin at\sinh at$	$\frac{s}{s^4+4a^4}$
$\cos a\,t$	$\frac{s}{s^2+a^2}$	$\cos a\,t\cosh a\,t$	$\frac{s^3}{s^4+4a^4}$
$\sin a\,t$	$\frac{a}{s^2+a^2}$	$\frac{1}{2a}(\sinh at+\sin at)$	$\frac{s^2}{s^4-a^4}$
$\cosh a\,t$	$\frac{s}{s^2-a^2}$	$\frac{1}{2}(\cosh at+\cos at)$	$\frac{s^3}{s^4-a^4}$
$\sinh a\,t$	$\frac{a}{s^2-a^2}$	$\frac{\sin at}{t}$	$\tan^{-1}\frac{a}{s}$
e^{-at}	$\frac{1}{s+a}$	$J_0(a\,t)$	$\frac{1}{\sqrt{s^2+a^2}}$
$e^{-bt}\cos a\,t$	$\frac{s+b}{(s+b)^2+a^2}$	$\frac{n}{a^n}\frac{J_n(at)}{t}$	$\frac{1}{\left(\sqrt{s^2+a^2}+s\right)^n}$
$e^{-bt}\sin a\,t$	$\frac{a}{(s+b)^2+a^2}$	$J_0\left(2\sqrt{at}\right)$	$\frac{1}{s}e^{-a/s}$

$$s\frac{\partial g}{\partial x}-\frac{\partial y}{\partial x}(0,x)+sg-y(0,x)+\frac{\partial g}{\partial x}=0$$

or

$$(s+1)\frac{\partial g}{\partial x}+sg=\frac{\partial y}{\partial x}(0,x)+y(0,x)=0$$

The second (boundary) condition gives $g(s, 0) + sg(s, 0) – y(0, 0) = 1$ or $g(s, 0) = 1/(1 + s)$. A solution of the preceding ordinary differential equation consistent with this condition is $g(s, x) = [1/(s + 1)]e^{-sx/(s+1)}$. Inversion of this transform gives $y(t, x) = e^{-(t+x)}I_0\,(2/\sqrt{tx})$, where I_0 is the zero-order Bessel function of an imaginary argument.

Convolution Integral

The *convolution integral* (*faltung*) of two functions $f(t)$, $r(t)$ is $x(t) = f(t)*r(t) = \int_0^t f(\tau)r(t-\tau)\,d\,\tau$.

Example 19.8.6. $t * \sin t = \int_0^t \tau\sin(t-\tau)\,d\,\tau = t - \sin t$.

13. $L[f(t)]L[h(t)] = L[f(t) * h(t)]$.

Fourier Transform

The *Fourier transform* is given by $F[f(t)] = (1/\sqrt{2\pi})\int_{-\infty}^{\infty} f(t)e^{-ist}\,d\,t = g(s)$ and its *inverse* by $F^{-1}[g(s)] = (1/\sqrt{2\pi})\int_{-\infty}^{\infty} g(s)e^{ist}\,d\,t = f(t)$. In brief, the condition for the Fourier transform to exist is that $\int_{-\infty}^{\infty}|f(t)|\,d\,t < \infty$, although certain functions may have a Fourier transform even if this is violated.

Example 19.8.7. The function $f(t) = 1$ for $-a \le t \le a$ and $= 0$ elsewhere has

$$F\left[f(t)\right] = \int_{-a}^{a} e^{-ist}\,dt = \int_{0}^{a} e^{ist}\,dt + \int_{0}^{a} e^{-ist}\,dt = 2\int_{0}^{a} \cos st\,dt = \frac{2\sin sa}{s}$$

Properties of the Fourier Transform

Let $F[f(t)] = g(s)$; $F^{-1}[g(s)] = f(t)$.

1. $F[f^{(n)}(t)] = (i\,s)^n\,F[f(t)]$
2. $F[af(t) + bh(t)] = aF[f(t)] + bF[h(t)]$
3. $F[f(-t)] = g(-s)$
4. $F[f(at)] = 1/a\;g(s/a), \quad a > 0$
5. $F[e^{-iwt}f(t)] = g(s + w)$
6. $F[f(t + t_1)] = e^{ist_1}\,g(s)$
7. $F[f(t)] = G(i\,s) + G(-i\,s)$ if $f(t) = f(-t)$($f(t)$ even)
   $F[f(t)] = G(i\,s) - G(-i\,s)$ if $f(t) = -f(-t)$(f odd)

where $G(s) = L[f(t)]$. This result allows the use of the Laplace transform tables to obtain the Fourier transforms.

Example 19.8.8. Find $F[e^{-a|t|}]$ by property 7. The term $e^{-a|t|}$ is even. So $L[e^{-at}] = 1/(s + a)$. Therefore, $F[e^{-a|t|}] = 1/(i\,s + a) + 1/(-i\,s + a) = 2a/(s^2 + a^2)$.

Fourier Cosine Transform

The *Fourier cosine transform* is given by $F_c[f(t)] = g(s) = \sqrt{(2/\pi)}\int_0^\infty f(t)\cos s\,t\,d\,t$ and its *inverse* by $F_c^{-1}[g(s)] = f(t) = \sqrt{(2/\pi)}\int_0^\infty g(s)\cos s\,t\,d\,s$. The *Fourier sine transform* F_s is obtainable by replacing the cosine by the sine in the above integrals.

Example 19.8.9. $F_c[f(t)]$, $f(t) = 1$ for $0 < t < a$ and 0 for $a < t < \infty$. $F_c\,[f(t)] = \sqrt{(2/\pi)}\int_0^a \cos s\,t\,d\,t = \sqrt{(2/\pi)}\,(\sin a\,s)/s$.

Properties of the Fourier Cosine Transform

$F_c[f(t)] = g(s)$.

1. $F_c[af(t) + bh(t)] = aF_c[f(t)] + bF_c[h(t)]$
2. $F_c[f(at)] = (1/a)\;g\;(s/a)$
3. $F_c[f(at)\cos bt] = 1/2a\;[g\;((s + b)/a) + g((s - b)/a)]$, $a, b > 0$
4. $F_c[t^{2n}f(t)] = (-1)^n(d^{2n}g)/(d\,s^{2n})$
5. $F_c[t^{2n+1}f(t)] = (-1)^n(d^{2n+1})/(d\,s^{2n+1})\;F_s[f(t)]$

Table 19.8.3 presents some Fourier cosine transforms.

Example 19.8.10. The temperature θ in the semiinfinite rod $0 \le x < \infty$ is determined by the differential equation $\partial\theta/\partial t = k(\partial^2\theta/\partial x^2)$ and the condition $\theta = 0$ when $t = 0$, $x \ge 0$; $\partial\theta/\partial x = -\mu$ = constant when $x = 0$, $t > 0$. By using the Fourier cosine transform, a solution may be found as $\theta(x, t) = (2\mu/\pi)\int_0^\infty (\cos px/p)\,(1 - e^{-kp2t})\;\;d\,p$.

References

Churchill, R. V. 1958. *Operational Mathematics.* McGraw-Hill, New York.

Ditkin, B. A. and Proodnikav, A. P. 1965. *Handbook of Operational Mathematics* (in Russian). Nauka, Moscow.

Doetsch, G. 1950–1956. *Handbuch der Laplace Transformation,* vols. I-IV (in German). Birkhauser, Basel.

Table 19.8.3 Fourier Cosine Transforms

$f(t)$	$\dfrac{g(s)}{\sqrt{2/\pi}}$
$\left.\begin{matrix} t & 0<t<1 \\ 2-t & 1<t<2 \\ 0 & 2<t<\infty \end{matrix}\right\}$	$\dfrac{1}{s^2}[2\cos s - 1 - \cos 2s]$
$t^{-1/2}$	$\pi^{1/2}(s)^{-1/2}$
$\left.\begin{matrix} 0 & 0<t<a \\ (t-a)^{-1/2} & a<t<\infty \end{matrix}\right\}$	$\pi^{1/2}(s)^{-1/2}\,[\cos a\,s - \sin a\,s]$
$(t^2+a^2)^{-1}$	$\frac{1}{2}\pi a^{-1}e^{-as}$
$e^{-at}, \quad a>0$	$\dfrac{a}{s^2+a^2}$
$e^{-at^2}, \quad a>0$	$\frac{1}{2}\pi^{1/2}a^{-1/2}e^{-s^2/4a}$
$\dfrac{\sin at}{t} \quad a>0$	$\begin{cases} \pi/2 & s<a \\ \pi/4 & s=a \\ 0 & s>a \end{cases}$

Nixon, F. E. 1960. *Handbook of Laplace Transforms.* Prentice-Hall, Englewood Cliffs, NJ.
Sneddon, I. 1951. *Fourier Transforms.* McGraw-Hill, New York.
Widder, D. 1946. *The Laplace Transform*, Princeton University Press, Princeton, NJ.

Further Information

The references citing G. Doetsch, *Handbuch der Laplace Transformation*, vols. I-IV, Birkhauser, Basel, 1950–1956 (in German) *and* B. A. Ditkin and A. P. Prodnikav, *Handbook of Operational Mathematics*, Moscow, 1965 (in Russian) are the most extensive tables known. The latter reference is 485 pages.

19.9 Calculus of Variations

William F. Ames

The basic problem in the *calculus of variations* is to determine a function such that a certain *functional*, often an integral involving that function and certain of its derivatives, takes on *maximum or minimum values*. As an example, find the function $y(x)$ such that $y(x_1) = y_1$, $y(x_2) = y_2$ and the integral (functional) $I = 2\pi \int_{x_1}^{x_2} y[1 + y')^2]^{1/2}\, d\,x$ is a minimum. A second example concerns the transverse deformation $u(x, t)$ of a beam. The energy functional $I = \int_{t_1}^{t_2} \int_0^L [1/2\ \rho\ (\partial u/\partial t)^2 - 1/2\ EI\ (\partial^2 u/\partial x^2)^2 + fu]\, d\,x\, d\,t$ is to be minimized.

The Euler Equation

The elementary part of the theory is concerned with a *necessary* condition (generally in the form of a differential equation with boundary conditions) that the required function must satisfy. To show mathematically that the function obtained actually maximizes (or minimizes) the integral is much more difficult than the corresponding problems of the differential calculus.

The *simplest case* is to determine a function $y(x)$ that makes the integral $I = \int_{x_1}^{x_2} F(x, y, y')\, dx$ stationary and that satisfies the prescribed end conditions $y(x_1) = y_1$ and $y(x_2) = y_2$. Here we suppose F has continuous second partial derivatives with respect to x, y, and $y' = dy/dx$. If $y(x)$ is such a function, then it must satisfy the *Euler equation* $(d/dx)(\partial F/\partial y') - (\partial F/\partial y) = 0$, which is the required necessary condition. The indicated partial derivatives have been formed by treating x, y, and y' as independent variables. Expanding the equation, the equivalent form $F_{y'y'}y'' + F_{y'y}y' + (F_{y'x} - F_y) = 0$ is found. This is second order in y unless $F_{y'y'} = (\partial^2 F)/[(\partial y')^2] = 0$. An alternative form $1/y'[d/dx(F - (\partial F/\partial y')(dy/dx)) - (\partial F/\partial x)] = 0$ is useful. Clearly, if F does not involve x explicitly $[(\partial F/\partial x) = 0]$ a first integral of Euler's equation is $F - y'(\partial F/\partial y') = c$. If F does not involve y explicitly $[(\partial F/\partial y) = 0]$ a first integral is $(\partial F/\partial y') = c$.

The Euler equation for $I = 2\pi \int_{x_1}^{x_2} y[1 + (y')^2]^{1/2}\, dx$, $y(x_1) = y_1$, $y(x_2) = y_2$ is $(d/dx)[yy'/[1 + (y')^2]^{1/2}] - [1 + (y')^2]^{1/2} = 0$ or after reduction $yy'' - (y')^2 - 1 = 0$. The solution is $y = c_1 \cosh(x/c_1 + c_2)$, where c_1 and c_2 are integration constants. Thus the required minimal surface, if it exists, must be obtained by revolving a catenary. Can c_1 and c_2 be chosen so that the solution passes through the assigned points? The answer is found in the solution of a transcendental equation that has two, one, or no solutions, depending on the prescribed values of y_1 and y_2.

The Variation

If $F = F(x, y, y')$, with x independent and $y = y(x)$, then the *first variation* δF of F is defined to be $\delta F = (\partial F/\partial x)\ \delta y + (\partial F/\partial y)\ \delta y'$ and $\delta y' = \delta\,(dy/dx) = (d/dx)\ (\delta y)$ — that is, they commute. Note that the first variation, δF, of a functional is a first-order change from curve to curve, whereas the differential of a function is a first-order approximation to the change in that function along a *particular curve*. The laws of δ are as follows: $\delta(c_1F + c_2G) = c_1\delta F + c_2\delta G$; $\delta(FG) = F\delta G + G\delta F$; $\delta(F/G) = (G\delta F - F\delta G)/G^2$; if x is an independent variable, $\delta x = 0$; if $u = u(x, y)$; $(\partial/\partial x)(\delta u) = \delta(\partial u/\partial x)$, $(\partial/\partial y)\ (\delta u) = \delta(\partial u/\partial y)$.

A necessary condition that the integral $I = \int_{x_1}^{x_2} F(x, y, y')\, dx$ be stationary is that its (first) variation vanish — that is, $\delta I = \delta \int_{x_1}^{x_2} F(x, y, y')\, dx = 0$. Carrying out the variation and integrating by parts yields of $\delta I = \int_{x_1}^{x_2} [(\partial F/\partial y) - (d/dx)(\partial F/\partial y')]\ \delta y\, dx + [(\partial F/\partial y')\ \partial y]_{x_1}^{x_2} = 0$. The arbitrary nature of δy means the square bracket must vanish and the last term constitutes the *natural boundary conditions.*

Example. The *Euler equation* of $\int_{x_1}^{x_2} F(x, y, y', y'')\, dx$ is $(d^2/dx^2)(\partial F/\partial y'') - (d/dx)(\partial F/\partial y') + (\partial F/\partial y) = 0$, with natural boundary conditions $\{[(d/dx)(\partial F/\partial y'') - (\partial F/\partial y')]\ \delta y\}_{x_1}^{x_2} = 0$ and $(\partial F/\partial y'')\ \delta y'|_{x_1}^{x_2} = 0$. The Euler equation of $\int_{x_1}^{x_2} \int_{y_1}^{y_2} F(x, y, u, u_x, u_y, u_{xx}, u_{xy}, u_{yy})\, dx\, dy$ is $(\partial^2/\partial x^2)(\partial F/\partial u_{xx}) + (\partial^2/\partial x \partial y)(\partial F/u_{xy}) + (\partial^2/\partial y^2)(\partial F/\partial u_{yy}) - (\partial/\partial x)(\partial F/\partial u_x) - (\partial/\partial y)(\partial F/\partial u_y) + (\partial F/\partial u)$, and the natural boundary conditions are

$$\left[\left(\frac{\partial}{\partial x}\left(\frac{\partial F}{\partial u_{xx}}\right)+\frac{\partial}{\partial y}\left(\frac{\partial F}{\partial u_{xy}}\right)-\frac{\partial F}{\partial u_x}\right)\delta u\right]_{x_1}^{x_2}=0, \qquad \left[\left(\frac{\partial F}{\partial u_{xx}}\right)\delta u_x\right]_{x_1}^{x_2}=0$$

$$\left[\left(\frac{\partial}{\partial y}\left(\frac{\partial F}{\partial u_{yy}}\right)+\frac{\partial}{\partial x}\left(\frac{\partial F}{\partial u_{xy}}\right)-\frac{\partial F}{\partial u_y}\right)\delta u\right]_{y_1}^{y_2}=0, \qquad \left[\left(\frac{\partial F}{\partial u_{yy}}\right)\delta u_y\right]_{y_1}^{y_2}=0$$

In the more general case of $I = \iint_R F(x, y, u, v, u_x, u_y, v_x, v_y)\, dx\, dy$, the condition $\delta I = 0$ gives rise to the two Euler equations $(\partial/\partial x)(\partial F/\partial u_x) + (\partial/\partial y)(\partial F/\partial u_y) - (\partial F/\partial u) = 0$ and $(\partial/\partial x)(\partial F/\partial v_x) + (\partial/\partial y)(\partial F/\partial v_y) - (\partial F/\partial v) = 0$. These are two PDEs in u and v that are linear or quasi-linear in u and v. The Euler equation for $I = \iiint_R (u_x^2 + u_y^2 + u_z^2)\, dx\, dy\, dz$, from $\delta I = 0$, is Laplace's equation $u_{xx} + u_{yy} + u_{zz} = 0$.

Variational problems are easily derived from the differential equation and associated boundary conditions by multiplying by the variation and integrating the appropriate number of times. To illustrate, let $F(x)$, $\rho(x)$, $p(x)$, and w be the tension, the linear mass density, the natural load, and (constant) angular velocity of a rotating string of length L. The equation of motion is $(d/dx)[F\,(dy/dx)] + \rho w^2 y + p = 0$. To formulate a corresponding variational problem, multiply all terms by a variation δy and integrate over $(0, L)$ to obtain

$$\int_0^L \frac{d}{dx}\left(F\frac{dy}{dx}\right)\delta y\, dx + \int_0^L \rho w^2 y \delta y\, dx + \int_0^L p\,\delta y\, dx = 0$$

The second and third integrals are the variations of $1/2\ \rho w^2 y^2$ and py, respectively. To treat the first integral, integrate by parts to obtain

$$\left[F\frac{dy}{dx}\delta y\right]_0^L - \int_0^L F\frac{dy}{dx}\delta\frac{dy}{dx}\,dx = \left[F\frac{dy}{dx}\delta y\right]_0^L - \int_0^L \frac{1}{2}F\delta\left(\frac{dy}{dx}\right)^2 dx = 0$$

So the variation formulation is

$$\delta\int_0^L\left[\frac{1}{2}\rho w^2 y^2 + py - \frac{1}{2}F\left(\frac{dy}{dx}\right)^2\right]dx + \left[F\frac{dy}{dx}\delta y\right]_0^L = 0$$

The last term represents the *natural boundary conditions*. The term $1/2\ \rho w^2 y^2$ is the kinetic energy per unit length, the term $-py$ is the potential energy per unit length due to the radial force $p(x)$, and the term $1/2\ F(dy/dx)^2$ is a first approximation to the potential energy per unit length due to the tension $F(x)$ in the string. Thus the integral is often called the *energy integral*.

Constraints

The variations in some cases cannot be arbitrarily assigned because of one or more auxiliary conditions that are usually called *constraints*. A typical case is the functional $\int_{x_1}^{x_2} F(x, u, v, u_x, v_x)\, dx$ with a constraint $\phi(u, v) = 0$ relating u and v. If the variations of u and v (δu and δv) vanish at the end points, then the variation of the integral becomes

$$\int_{x_1}^{x_2}\left\{\left[\frac{\partial F}{\partial u} - \frac{d}{dx}\left(\frac{\partial F}{\partial u_x}\right)\right]\delta u + \left[\frac{\partial F}{\partial v} - \frac{d}{dx}\left(\frac{\partial F}{\partial v_x}\right)\right]\delta v\right\}dx = 0$$

The variation of the constraint $\phi(u, v) = 0$, $\phi_u \delta u + \phi_v \delta v = 0$ means that the variations cannot both be assigned arbitrarily inside (x_1, x_2), so their coefficients need not vanish separately. Multiply $\phi_u \delta u + \phi_v \delta v = 0$ by a Lagrange multiplier λ (may be a function of x) and integrate to find $\int_{x_1}^{x_2} (\lambda\phi_u \delta u + \lambda\phi_v \delta v)\, dx = 0$. Adding this to the previous result yields

$$\int_{x_1}^{x_2} \left\{ \left[\frac{\partial F}{\partial u} - \frac{d}{dx}\left(\frac{\partial F}{\partial u_x} \right) + \lambda\phi_u \right] \delta u + \left[\frac{\partial F}{\partial v} - \frac{d}{dx}\left(\frac{\partial F}{\partial v_x} \right) + \lambda\phi_v \right] \delta v \right\} dx = 0$$

which must hold for any λ. Assign λ so the first square bracket vanishes. Then δv can be assigned to vanish inside (x_1, x_2) so the two systems

$$\frac{d}{dx}\left[\frac{\partial F}{\partial u_x} \right] - \frac{\partial F}{\partial u} - \lambda\phi_u = 0, \qquad \frac{d}{dx}\left[\frac{\partial F}{\partial v_x} \right] - \frac{\partial F}{\partial v} - \lambda\phi_v = 0$$

plus the constraint $\phi(u, v) = 0$ are three equations for u, v and λ.

References

Gelfand, I. M. and Fomin, S. V. 1963. *Calculus of Variations*. Prentice Hall, Englewood Cliffs, NJ.

Lanczos, C. 1949. *The Variational Principles of Mechanics.* Univ. of Toronto Press, Toronto.

Schechter, R. S. 1967. *The Variational Method in Engineering,* McGraw-Hill, New York.

Vujanovic, B. D. and Jones, S. E. 1989. *Variational Methods in Nonconservative Phenomena.* Academic Press, New York.

Weinstock, R. 1952. *Calculus of Variations, with Applications to Physics and Engineering.* McGraw-Hill, New York.

19.10 Optimization Methods

George Cain

Linear Programming

Let **A** be an $m \times n$ matrix, **b** a column vector with m components, and **c** a column vector with n components. Suppose $m < n$, and assume the rank of **A** is m. The standard linear programming problem is to find, among all nonnegative solutions of $\mathbf{Ax} = \mathbf{b}$, one that minimizes

$$\mathbf{c}^T\mathbf{x} = c_1x_1 + c_2x_2 + \cdots + c_nx_n$$

This problem is called a *linear* program. Each solution of the system $\mathbf{Ax} = \mathbf{b}$ is called a *feasible* solution, and the *feasible set* is the collection of all *feasible solutions.* The function $\mathbf{c}^T\mathbf{x} = c_1x_1 + c_2x_2 + \cdots + c_nx_n$ is the cost function, or the objective function. A solution to the linear program is called an *optimal feasible solution.*

Let **B** be an $m \times n$ submatrix of **A** made up of m linearly independent columns of **A**, and let **C** be the $m \times (n - m)$ matrix made up of the remaining columns of **A**. Let $\mathbf{x_B}$ be the vector consisting of the components of **x** corresponding to the columns of **A** that make up **B**, and let $\mathbf{x_C}$ be the vector of the remaining components of **x**, that is, the components of **x** that correspond to the columns of **C**. Then the equation $\mathbf{Ax} = \mathbf{b}$ may be written $\mathbf{Bx_B} + \mathbf{Cx_C} = \mathbf{b}$. A solution of $\mathbf{Bx_B} = \mathbf{b}$ together with $\mathbf{x_C} = \mathbf{0}$ gives a solution **x** of the system $\mathbf{Ax} = \mathbf{b}$. Such a solution is called a *basic solution*, and if it is, in addition, nonnegative, it is a *basic feasible solution.* If it is also optimal, it is an *optimal basic feasible solution.* The components of a basic solution are called *basic variables.*

The Fundamental Theorem of Linear Programming says that if there is a feasible solution, there is a basic feasible solution, and if there is an optimal feasible solution, there is an optimal basic feasible solution. The linear programming problem is thus reduced to searching among the set of basic solutions for an optimal solution. This set is, of course, finite, containing as many as $n!/[m!(n - m)!]$ points. In practice, this will be a very large number, making it imperative that one use some efficient search procedure in seeking an optimal solution. The most important of such procedures is the *simplex method,* details of which may be found in the references.

The problem of finding a solution of $\mathbf{Ax} \le \mathbf{b}$ that minimizes $\mathbf{c}^T\mathbf{x}$ can be reduced to the standard problem by appending to the vector **x** an additional m nonnegative components, called *slack variables*. The vector **x** is replaced by **z**, where $\mathbf{z}^T = [x_1,x_2\ldots x_n\ s_1,s_2\ldots x_m]$, and the matrix **A** is replaced by $\mathbf{B} = [\mathbf{A}\ \mathbf{I}]$, where **I** is the $m \times m$ identity matrix. The equation $\mathbf{Ax} = \mathbf{b}$ is thus replaced by $\mathbf{Bz} = \mathbf{Ax} + s = \mathbf{b}$, where $s^T = [s_1,s_2,\ldots,s_m]$. Similarly, if inequalities are reversed so that we have $\mathbf{Ax} \le \mathbf{b}$, we simply append $-s$ to the vector x. In this case, the additional variables are called *surplus variables.*

Associated with every linear programming problem is a corresponding dual problem. If the *primal* problem is to minimize $\mathbf{c}^T\mathbf{x}$ subject to $\mathbf{Ax} \ge \mathbf{b}$, and $\mathbf{x} \ge 0$, the corresponding *dual* problem is to maximize $\mathbf{y}^T\mathbf{b}$ subject to $\mathbf{t}^T\mathbf{A} \le \mathbf{c}^T$. If either the primal problem or the dual problem has an optimal solution, so also does the other. Moreover, if $\mathbf{x_p}$ is an optimal solution for the primal problem and $\mathbf{y_d}$ is an optimal solution for the corresponding dual problem $\mathbf{c}^T\mathbf{x_p} = \mathbf{y_d}^T\mathbf{b}$.

Unconstrained Nonlinear Programming

The problem of minimizing or maximizing a sufficiently smooth nonlinear function $f(x)$ of n variables, $\mathbf{x}^T = [x_1,x_2\ldots x_n]$, with no restrictions on x is essentially an ordinary problem in calculus. At a minimizer or maximizer x^*, it must be true that the gradient of f vanishes:

$$\nabla f\left(\mathbf{x}^*\right) = 0$$

Thus x^* will be in the set of all solutions of this system of n generally nonlinear equations. The solution of the system can be, of course, a nontrivial undertaking. There are many recipes for solving systems of nonlinear equations. A method specifically designed for minimizing f is the *method of steepest descent*. It is an old and honorable algorithm, and the one on which most other more complicated algorithms for unconstrained optimization are based. The method is based on the fact that at any point x, the direction of maximum decrease of f is in the direction of $-\nabla f(x)$. The algorithm searches in this direction for a minimum, recomputes $\nabla f(x)$ at this point, and continues iteratively. Explicitly:

1. Choose an initial point x_0.
2. Assume x_k has been computed; then compute $y_k = \nabla f(x_k)$, and let $t_k \geq 0$ be a local minimum of $g(t) = f(x_k - ty_k)$. Then $x_{k+1} = x_k - t_k y_k$.
3. Replace k by $k + 1$, and repeat step 2 until t_k is small enough.

Under reasonably general conditions, the sequence (x_k) converges to a minimum of f.

Constrained Nonlinear Programming

The problem of finding the maximum or minimum of a function $f(x)$ of n variables, subject to the constraints

$$\mathbf{a}(\mathbf{x}) = \begin{bmatrix} a_1(x_1, x_2, \ldots, x_n) \\ a_2(x_1, x_2, \ldots, x_n) \\ \vdots \\ a_m(x_1, x_2, \ldots, x_n) \end{bmatrix} = \begin{bmatrix} b_1 \\ b \\ \vdots \\ b_m \end{bmatrix} = \mathbf{b}$$

is made into an unconstrained problem by introducing the new function $L(x)$:

$$L(\mathbf{x}) = f(\mathbf{x}) + \mathbf{z}^{\mathrm{T}}\mathbf{a}(\mathbf{x})$$

where $\mathbf{z}^{\mathrm{T}} = [\lambda_1, \lambda_2, \ldots, \lambda_m]$ is the vector of *Lagrange multipliers*. Now the requirement that $\nabla L(x) = 0$, together with the constraints $a(x) = \mathbf{b}$, give a system of $n + m$ equations

$$\nabla f(\mathbf{x}) + \mathbf{z}^{\mathrm{T}}\nabla\mathbf{a}(\mathbf{x}) = 0$$

$$\mathbf{a}(\mathbf{x}) = \mathbf{b}$$

for the $n + m$ unknowns $x_1, x_2 \ldots, x_n, \lambda_1\lambda_2 \ldots, \lambda_m$ that must be satisfied by the minimizer (or maximizer) x.

The problem of inequality constraints is significantly more complicated in the nonlinear case than in the linear case. Consider the problem of minimizing $f(x)$ subject to m equality constraints $\mathbf{a}(\mathbf{x}) = \mathbf{b}$, and p inequality constraints $c(x) \leq \mathbf{d}$ [thus $\mathbf{a}(\mathbf{x})$ and $\mathbf{b}$ are vectors of m components, and $c(x)$ and $\mathbf{d}$ are vectors of p components.] A point x^* that satisfies the constraints is a *regular point* if the collection

$$\left\{\nabla a_1(\mathbf{x}^*), \nabla a_2(\mathbf{x}^*), \ldots, \nabla a_m(\mathbf{x}^*)\right\} \cup \left\{\nabla c_j(\mathbf{x}^*) : j \in J\right\}$$

where

$$J = \left\{j : c_j(\mathbf{x}^*) = d_j\right\}$$

is linearly independent. If x^* is a local minimum for the constrained problem and if it is a regular point, there is a vector $\boldsymbol{z}$ with m components and a vector $\mathbf{w} \geq \mathbf{0}$ with p components such that

$$\nabla f\left(\mathbf{x}^*\right) + \mathbf{z}^{\mathrm{T}} \nabla \mathbf{a}\left(\mathbf{x}^*\right) + \mathbf{w}^{\mathrm{T}} \nabla \mathbf{c}\left(\mathbf{x}^*\right) = 0$$

$$\mathbf{w}^{\mathrm{T}}\left(\mathbf{c}\left(\mathbf{x}^*\right) - \mathbf{d}\right) = 0$$

These are the *Kuhn-Tucker conditions*. Note that in order to solve these equations, one needs to know for which j it is true that $c_j(x^*) = 0$. (Such a constraint is said to be *active*.)

References

Luenberger, D. C. 1984. *Linear and Nonlinear Programming*, 2nd ed. Addison-Wesley, Reading, MA.

Peressini, A. L. Sullivan, F. E., and Uhl, J. J., Jr. 1988. *The Mathematics of Nonlinear Programming*. Springer-Verlag, New York.

19.11 Engineering Statistics

Y. L. Tong

Introduction

In most engineering experiments, the outcomes (and hence the observed data) appear in a random and on deterministic fashion. For example, the operating time of a system before failure, the tensile strength of a certain type of material, and the number of defective items in a batch of items produced are all subject to random variations from one experiment to another. In engineering statistics, we apply the theory and methods of statistics to develop procedures for summarizing the data and making statistical inferences, thus obtaining useful information with the presence of randomness and uncertainty.

Elementary Probability

Random Variables and Probability Distributions

Intuitively speaking, a random variable (denoted by X, Y, Z, etc.) takes a numerical value that depends on the outcome of the experiment. Since the outcome of an experiment is subject to random variation, the resulting numerical value is also random. In order to provide a stochastic model for describing the probability distribution of a random variable X, we generally classify random variables into two groups: the discrete type and the continuous type. The discrete random variables are those which, technically speaking, take a finite number or a countably infinite number of possible numerical values. (In most engineering applications they take nonnegative integer values.) Continuous random variables involve outcome variables such as time, length or distance, area, and volume. We specify a function $f(x)$, called the probability density function (p.d.f.) of a random variable X, such that the random variable X takes a value in a set A (or real numbers) as given by

$$P[X \in A] = \begin{cases} \sum_{x \in A} f(x) & \text{for all sets } A \text{ if } X \text{ is discrete} \\ \int_A f(x)\,dx & \text{for all intervals } A \text{ if } X \text{ is continuous} \end{cases} \tag{9.11.1}$$

By letting A be the set of all values that are less than or equal to a fixed number t, i.e., $A = (-\infty, t)$, the probability function $P[X \le t]$, denoted by $F(t)$, is called the distribution function of X. We note that, by calculus, if X is a continuous random variable and if $F(x)$ is differentiable, then $f(x) = \frac{d}{dx} F(x)$.

Expectations

In many applications the "payoff" or "reward" of an experiment with a numerical outcome X is a specific function of X ($u(X)$, say). Since X is a random variable, $u(X)$ is also a random variable. We define the expected value of $u(X)$ by

$$Eu(X) = \begin{cases} \sum_x u(x) f(x) & \text{if } X \text{ is discrete} \\ \int_{-\infty}^{\infty} u(x) f(x)\,dx & \text{if } X \text{ is continuous} \end{cases} \tag{9.11.12}$$

provided of course, that, the sum or the integral exists. In particular, if $u(x) = x$, the $EX \equiv \mu$ is called the mean of X (of the distribution) and $E(X - \mu)^2 \equiv \sigma^2$ is called the variance of X (of the distribution). The mean is a measurement of the central tendency, and the variance is a measurement of the dispersion of the distribution.

Some Commonly Used Distributions

Many well-known distributions are useful in engineering statistics. Among the discrete distributions, the hypergeometric and binomial distributions have applications in acceptance sampling problems and quality control, and the Poisson distribution is useful for studying queuing theory and other related problems. Among the continuous distributions, the uniform distribution concerns random numbers and can be applied in simulation studies, the exponential and gamma distributions are closely related to the Poisson distribution, and they, together with the Weibull distribution, have important applications in life testing and reliability studies. All of these distributions involve some unknown parameter(s), hence their means and variances also depend on the parameter(s). The reader is referred to textbooks in this area for details. For example, Hahn and Shapiro (1967, pp. 163–169 and pp. 120–134) contains a comprehensive listing of these and other distributions on their p.d.f.'s and the graphs, parameter(s), means, variances, with discussions and examples of their applications.

The Normal Distribution

Perhaps *the* most important distribution in statistics and probability is the normal distribution (also known as the Gaussian distribution). This distribution involves two parameters: μ and σ^2, and its p.d.f. is given by

$$f(x) = f\left(x;\mu,\sigma^2\right) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} \tag{9.11.3}$$

for $-\infty < \mu < \infty$, $\sigma^2 > 0$, and $-\infty < \chi < \infty$. It can be shown analytically that, for a p.d.f. of this form, the values of μ and σ^2 are, respectively, that of the mean and the variance of the distribution. Further, the quantity, $\sigma = \sqrt{\sigma^2}$ is called the standard deviation of the distribution. We shall use the symbol $X \sim N(\mu, \sigma^2)$ to denote that X has a normal distribution with mean μ and variance σ^2. When plottting the p.d.f. $f(x; \mu, \sigma^2)$ given in Equation (19.11.3) we see that the resulting graph represents a bell-shaped curve symmetric about μ, as shown in Figure 19.11.1.

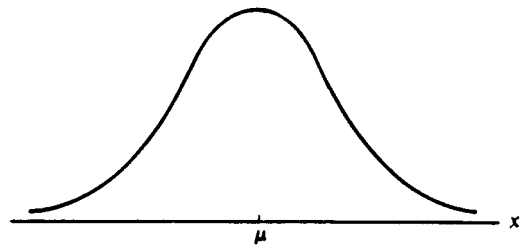


FIGURE 19.11.1 The normal curve with mean μ and variance σ^2.

If a random variable Z has an $N(0,1)$ distribution, then the p.d.f. of Z is given by (from Equation (19.11.3))

$$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} \qquad -\infty < z < \infty \tag{9.11.4}$$

The distribution function of Z,

$$\Phi(z) = \int_{-\infty}^{x} \phi(u)\, du \qquad -\infty < z < \infty \tag{9.11.5}$$

cannot be given in a closed form, hence it has been tabulated. The table of $\Phi(z)$ can be found in most textbooks in statistics and probability, including those listed in the references at the end of this section. (We note in passing that, by the symmetry property, $\Phi(z) + \Phi(-z) = 1$ holds for all z.)

Random Sample and Sampling Distributions

Random Sample and Related Statistics

As noted in Box et al., (1978), the design and analysis of engineering experiments usually involves the following steps:

1. The choice of a suitable stochastic model by assuming that the observations follow a certain distribution. The functional form of the distribution (or the p.d.f.) is assumed to be known, except the value(s) of the parameters(s).
2. Design of experiments and collection of data.
3. Summarization of data and computation of certain statistics.
4. Statistical inference (including the estimation of the parameters of the underlying distribution and the hypothesis-testing problems).

In order to make statistical inference concerning the parameter(s) of a distribution, it is essential to first study the sampling distributions. We say that $X_1, X_2, \ldots, X_n$ represent a random sample of size n if they are independent random variables and each of them has the same p.d.f., $f(x)$. (Due to space limitations, the notion of independence will not be carefully discussed here. Nevertheless, we say that $X_1, X_2, \ldots, X_n$ are independent if

$$P\left[X_1 \in A_1, X_2 \in A_2, \ldots, X_n \in A_n\right] = \prod_{i=1}^{n} P\left[X_i \in A_i\right] \tag{19.11.6}$$

holds for all sets $A_1, A_2, \ldots, A_n$.) Since the parameter(s) of the population is (are) unknown, the population mean μ and the population variance σ^2 are unknown. In most commonly used distributions μ and σ^2 can be estimated by the sample mean $\bar{X}$ and the sample variance S^2, respectively, which are given by

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \qquad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} X_i^2 - n\bar{X}^2\right] \tag{19.11.7}$$

(The second equality in the formula for S^2 can be verified algebraically.) Now, since $X_1, X_2, \ldots, X_n$ are random variables $\bar{X}$ and S^2 are also random variables. Each of them is called a statistic and has a probability distribution which also involves the unknown parameter(s). In probability theory there are two fundamental results concerning their distributional properties.

Theorem 1. (Weak Law of Large Numbers). As the sample size n becomes large, $\bar{X}$ converges to μ in probability and S^2 converges to σ^2 in probability. More precisely, for every fixed positive number $\varepsilon > 0$ we have

$$P\left[\left|\bar{X} - \mu\right| \le \varepsilon\right] \to 1, \qquad P\left[\left|S^2 - \sigma^2\right| \le \varepsilon\right] \to 1 \tag{19.11.8}$$

as $n \to \infty$.

Theorem 2. (Central Limit Theorem). As n becomes large, the distribution of the random variable

$$Z = \frac{\bar{X} - \mu}{\sigma\ \sqrt{n}} = \frac{\sqrt{n}\left(\bar{X} - \mu\right)}{\sigma} \tag{19.11.9}$$

has approximately an $N(0,1)$ distribution. More precisely,

$$P[Z \le z] \to \Phi(z) \text{ for every fixed } z \text{ as } n \to \infty \tag{19.11.10}$$

Normal Distribution-Related Sampling Distributions

One-Sample Case

Additional results exist when the observations come from a normal population. If $X_1, X_2, \ldots, X_n$ represent a random sample of size n from an $N(\mu,\sigma^2)$ population, then the following sample distributions are useful:

Fact 3. For every fixed n the distribution of Z given in Equation (19.11.9) has *exactly* an N(0,1) distribution.

Fact 4. The distribution of the statistic $T = \sqrt{n}(\bar{X} - \mu)/S$, where $S = \sqrt{S^2}$ is the sample standard deviation, is called a *Student's t distribution* with $\nu = n - 1$ degrees of freedom, in symbols, $t(n - 1)$.

This distribution is useful for making inference on μ when σ^2 is unknown; a table of the percentiles can be found in most statistics textbooks.

Fact 5. The distribution of the statistic $W = (n - 1)S^2/\sigma^2$ is called a *chi-squared distribution* with $\nu = n - 1$ degrees of freedom, in symbols $\chi^2(\nu)$.
Such a distribution is useful in making inference on σ^2; a table of the percentiles can also be found in most statistics books.

Two-Sample Case

In certain applications we may be interested in the comparisons of two different treatments. Suppose that independent samples from treatments T_1 and T_2 are to be observed as shown in Table 19.11.1.

TABLE 19.11.1 Summarization of Data for a Two-Sample Problem

Treatment	Observations	Distribution	Sample Size	Sample Mean	Sample Variance
T_1	$X_{11}, X_{12}, \ldots, X_{1n_1}$	$\mathcal{N}(\mu_1, \sigma_1^2)$	n_1	$\bar{X}_1$	S_1^2
T_2	$X_{21}, X_{22}, \ldots, X_{2n_2}$	$\mathcal{N}(\mu_2, \sigma_2^2)$	n_2	$\bar{X}_2$	S_2^2

The difference of the population means $(\mu_1 - \mu_2)$ and the ratio of the population variances can be estimated, respectively, by $(\bar{X}_1 - \bar{X}_2)$ and S_1^2/S_2^2. The following facts summarize the distributions of these statistics:

Fact 6. Under the assumption of normality, $(\bar{X}_1 - \bar{X}_2)$ has an $N(\mu_1 - \mu_2, (\sigma_1^2/n_1) + (\sigma_2^2/n_2))$ distribution; or equivalently, for all n_1, n_2 the statistic.

$$Z = \left[\left(\bar{X}_1 - \bar{X}_2\right) - \left(\mu_1 - \mu_2\right)\right] \Big/ \left(\sigma_1^2/n_1 + \sigma_2^2/n_2\right)^{1/2} \tag{19.11.11}$$

has an $N(0,1)$ distribution.

Fact 7. When $\sigma_1^2 = \sigma_2^2 \equiv \sigma^2$, the common population variance is estimated by

$$S_p^2 = \left(n_1 + n_2 - 2\right)^{-1}\left[\left(n_1 - 1\right)S_1^2 + \left(n_2 - 1\right)S_2^2\right] \tag{19.11.12}$$

and $(n_1 + n_2 - 2)S_p^2/\sigma^2$ has a $\chi^2(n_1 + n_2 - 2)$ distribution.

Fact 8. When $\sigma_1^2 = \sigma_2^2$, the statistic

$$T = \left[\left(\bar{X}_1 - \bar{X}_2\right) - \left(\mu_1 - \mu_2\right)\right] \Big/ S_p\left(1/n_1 + 1/n_2\right)^{1/2} \tag{19.11.13}$$

has a $t(n_1 + n_2 - 2)$ distribution, where $S_p = \sqrt{S_p^2}$.

Fact 9. The distribution of $F = (S_1^2/\sigma_1^2)/(S_2^2/\sigma_2^2)$ is called an F distribution with degrees of freedom $(n_1 - 1, n_2 - 1)$, in symbols, $F(n_1 - 1, n_2 - 1)$,

The percentiles of this distribution have also been tabulated and can be found in statistics books.

In the following two examples we illustrate numerically how to find probabilities and percentiles using the existing tables for the normal, Student's *t*, chi-squared, and *F* distributions.

Example 10. Suppose that in an experiment four observations are taken, and that the population is assumed to have a normal distribution with mean μ and variance σ^2. Let $\bar{X}$ and S^2 be the sample mean and sample variance as given in Equation (19.11.7).

(a) If, based on certain similar experiments conducted in the past, we know that $\sigma^2 = 1.8^2 \times 10^{-6}$ ($\sigma = 1.8 \times 10^{-3}$), then from $\Phi(-1.645) = 0.05$ and $\Phi(1.96) = 0.975$ we have

$$P\left[-1.645 \le \frac{\bar{X} - \mu}{1.8 \times 10^{-3}\ \sqrt{4}} \le 1.96\right] = 0.975 - 0.05 = 0.925$$

or equivalently,

$$P\left[-1.645 \times 0.9 \times 10^{-3} \le \bar{X} - \mu \le 1.96 \times 0.9 \times 10^{-3}\right] = 0.925$$

(b) The statistic $T = 2(\bar{X} - \mu)/S$ has a Student's *t* distribution with 3 degrees of freedom (in symbols, $t(3)$). From the *t* table we have

$$P\left[-3.182 \le 2\left(\bar{X} - \mu\right)/S \le 3.182\right] = 0.95$$

which yields

$$P\left[-3.182 \times \frac{S}{2} \le \bar{X} - \mu \le 3.182 \times \frac{S}{2}\right] = 0.95$$

or equivalently,

$$P\left[\bar{X} - 3.182 \times \frac{S}{2} \le \mu \le \bar{X} + 3.182 \times \frac{S}{2}\right] = 0.95$$

This is, in fact, the basis for obtaining the confidence interval for μ given in Equation (19.11.17) when σ^2 is unknown.

(c) The statistic $3S^2/\sigma^2$ has a chi-squared distribution with 3 degrees of freedom (in symbols, $\chi^2(3)$). Thus from the chi-squared table we have $P[0.216 \le 3S^2/\sigma^2 \le 9.348] = 0.95$, which yields

$$P\left[\frac{3S^2}{9.348} \le \sigma^2 \le \frac{3S^2}{0.216}\right] = 0.95$$

and it forms the basis for obtaining a confidence interval for σ^2 as given in Equation (19.11.18).

Example 11. Suppose that in Table 19.11.1 (with two treatments) we have $n_1 = 4$ and $n_2 = 5$, and we let $\bar{X}_1, \bar{X}_2$ and S_1^2, S_2^2 denote the corresponding sample means and sample variances, respectively.

(a) Assume that $\sigma_1^2 = \sigma_2^2$ where the common variance is unknown and is estimated by S_p^2 given in Equation (19.11.12). Then the statistic

$$T = \left[\left(\bar{X}_1 - \bar{X}_2\right) - \left(\mu_1 - \mu_2\right)\right] \Big/ S_p\left({}^1\!/_4 + {}^1\!/_5\right)^{1/2}$$

has a $t(7)$ distribution. Thus from the t table we have

$$P = \left[-2.998 \le \left[\left(\bar{X}_1 - \bar{X}_2\right) - \left(\mu_1 - \mu_2\right)\right] \Big/ S_p\left({}^1\!/_4 + {}^1\!/_5\right)^{1/2} \le 2.998\right] = 0.98$$

which is equivalent to saying that

$$P\left[-2.998 S_p\left({}^1\!/_4 + {}^1\!/_5\right)^{1/2} \le \mu_1 - \mu_2 \le \left(\bar{X}_1 - \bar{X}_2\right) + 2.998 S_p\left({}^1\!/_4 + {}^1\!/_5\right)^{1/2}\right] = 0.98$$

(b) The statistic $F = (S_1^2/\sigma_1^2)/(S_2^2/\sigma_2^2)$ has an $F(3,4)$ distribution. Thus from the F-table we have

$$P\left[\left(\frac{\sigma_2^2}{\sigma_1^2}\right)\left(\frac{S_1^2}{S_2^2}\right) \le 6.59\right] = 0.95$$

or equivalently,

$$P\left[\frac{\sigma_2^2}{\sigma_1^2} \le 6.59 \frac{S_2^2}{S_1^2}\right] = 0.95$$

The distributions listed above (normal, Student's t, chi-squared, and F) form an integral part of the classical statistical inference theory, and they are developed under the assumption that the observations follow a normal distribution. When the distribution of the population is not normal and inference on the populations means is to be made, we conclude that (1) if the sample sizes n_1, n_2 are large, then the statistic Z in Equation (19.11.11) has an approximate $N(0,1)$ distribution and (2) in the small-sample case, the exact distribution of $\bar{X}$ (of $(\bar{X}_1 - \bar{X}_2)$)) depends on the population p.d.f. There are several analytical methods for obtaining it, and those methods can be found in statistics textbooks.

Confidence Intervals

A method for estimating the population parameters based on the sample mean(s) and sample variance(s) involves the confidence intervals for the parameters.

One-Sample Case

1. Confidence Interval for μ When σ^2 is Known. Consider the situation in which a random sample of size n is taken from an $N(\mu,\sigma^2)$ population and σ^2 is known. An interval, I_1, of the form $I_1 = (\bar{X} - d, \bar{X} + d)$ (with width $2d$) is to be constructed as a "confidence interval or μ." If we make the assertion that μ is in this interval (i.e., μ is bounded below by $\bar{X} - d$ and bounded above by $\bar{X} + d$), then sometimes this assertion is correct and sometimes it is wrong, depending on the value of $\bar{X}$ in a given experiment. If for a fixed α value we would like to have a confidence probability (called confidence coefficient) such that

$$P[\mu \in I_1] = P[\bar{X} - d < \mu < \bar{X} + d] = 1 - \alpha \tag{19.11.14}$$

then we need to choose the value of d to satisfy $d = z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$, i.e.,

$$I_1 = \left(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) \tag{19.11.15}$$

where $z_{\alpha/2}$ is the $(1 - \alpha/2)$th percentile of the $N(0,1)$ distribution such that $\Phi(z_{\alpha/2}) = 1 - \alpha/2$. To see this, we note that from the sampling distribution of $\bar{X}$ (Fact 3) we have

$$\begin{aligned} P\left[\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right] &= P\left[\frac{|\bar{X} - \mu|}{\sigma\ \sqrt{n}} \le z_{\alpha/2}\right] \\ &= \Phi\left(z_{\alpha/2}\right) - \Phi\left(-z_{\alpha/2}\right) = 1 - \alpha \end{aligned} \tag{19.11.16}$$

We further note that, even when the original population is not normal, by Theorem 2 the confidence probability is approximately $(1 - \alpha)$ when the sample size is reasonably large.

2. Confidence Interval for μ When σ^2 is Unknown. Assume that the observations are from an $N(\mu,\sigma^2)$ population. When σ^2 is unknown, by Fact 4 and a similar argument we see that

$$I_2 = \left(\bar{X} - t_{\alpha/2}(n-1)\frac{S}{\sqrt{n}}, \bar{X} + t_{\alpha/2}(n-1)\frac{S}{\sqrt{n}}\right) \tag{19.11.17}$$

is a confidence interval for μ with confidence probability $1 - \alpha$, where $t_{\alpha/2}(n - 1)$ is the $(1 - \alpha/2)$th percentile of the $t(n - 1)$ distribution.

3. Confidence Interval for σ^2. If, under the same assumption of normality, a confidence interval for σ^2 is needed when μ is unknown, then

$$I_3 = \left((n-1)S^2/\chi^2_{1-\alpha/2}(n-1), (n-1)S^2/\chi^2_{\alpha/2}(n-1)\right) \tag{19.11.18}$$

has a confidence probability $1 - \alpha$, when $\chi^2_{1-\alpha/2}(n - 1)$ and $\chi^2_{\alpha/2}(n - 1)$ are the $(\alpha/2)$th and $(1 - \alpha/2)$th percentiles, respectively, of the $\chi^2(n - 1)$ distribution.

Two-Sample Case

1. Confidence Intervals for $\mu_1 - \mu_2$ When $\sigma_1^2 = \sigma_2^2$ are Known. Consider an experiment that involves the comparison of two treatments, T_1 and T_2, as indicated in Table 19.11.1. If a confidence interval for $\delta = \mu_1 - \mu_2$ is needed when σ_1^2 and σ_2^2 are unknown, then by Fact 6 and a similar argument, the confidence interval

$$I_4 = \left(\left(\bar{X}_1 - \bar{X}_2\right) - z_{\alpha/2}\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}, \left(\bar{X}_1 - \bar{X}_2\right) + z_{\alpha/2}\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}\right) \tag{19.11.19}$$

has a confidence probability $1 - \alpha$.

2. Confidence Interval for $\mu_1 - \mu_2$ when σ_1^2, σ_2^2 are Unknown but Equal. Under the additional assumption that $\sigma_1^2 = \sigma_2^2$, but the common variance is unknown, then by Fact 8 the confidence interval

$$I_5 = \left(\left(\bar{X}_1 - \bar{X}_2\right) - d, \left(\bar{X}_1 - \bar{X}_2\right) + d\right) \tag{19.11.20}$$

has a confidence probability $1 - \alpha$, where

$$d = t_{\alpha/2}\left(n_1 + n_2 - 2\right) S_p \left(1/n_1 + 1/n_2\right)^{1/2} \tag{19.11.21}$$

3. Confidence Interval for σ_2^2/σ_1^2. A confidence interval for the ratio of the variances σ_2^2/σ_1^2 can be obtained from the F distribution (see Fact 9), and the confidence interval

$$I_6 = \left(F_{1-\alpha/2}\left(n_1 - 1, n2_1 - 1\right)\frac{S_2^2}{S_1^2}, F_{\alpha/2}\left(n_1 - 1, n2_1 - 1\right)\frac{S_2^2}{S_1^2}\right) \tag{19.11.22}$$

has a confidence probability $1 - \alpha$, where $F_{1-\alpha/2}(n_1 - 1, n_2 - 1)$ and $F_{\alpha/2}(n_1 - 1, n_2 - 1)$ are, respectively, the $(\alpha/2)$th and $(1 - \alpha/2)$th percentiles of the $F(n_1 - 1, n_2 - 1)$ distribution.

Testing Statistical Hypotheses

A statistical hypothesis concerns a statement or assertion about the true value of the parameter in a given distribution. In the two-hypothesis problems, we deal with a null hypothesis and an alternative hypothesis, denoted by H_0 and H_1, respectively. A decision is to be made, based on the data of the experiment, to either accept H_0 (hence reject H_1) or reject H_0 (hence accept H_1). In such a two-action problem, we may commit two types of errors: the type I error is to reject H_0 when it is true, and the type II error is to accept H_0 when it is false. As a standard practice, we do not reject H_0 unless there is significant evidence indicating that it may be false. (In doing so, the burden of proof that H_0 is false is on the experimenter.) Thus we usually choose a small fixed number, α (such as 0.05 or 0.01), such that the probability of committing a type I error is at most (or equal to) α. With such a given α, we can then determine the region in the data space for the rejection of H_0 (called the critical region).

One-Sample Case

Suppose that $X_1, X_2, \ldots, X_n$ represent a random sample of size n from an $N(\mu,\sigma^2)$ population, and $\bar{X}$ and S^2 are, respectively, the sample mean and sample variance.

1. Test for Mean. In testing

$$H_0 : \mu = \mu_0 \text{ vs. } H_1 : \mu = \mu_1\left(\mu_1 > \mu_0\right) \text{ or } H_1 : \mu > \mu_0$$

when σ^2 is known, we reject H_0 when $\bar{X}$ is large. To determine the cut-off point, we note (by Fact 3) that the statistic $Z_0 = (\bar{X} - \mu_0)/(\sigma/\sqrt{n})$ has an $N(0,1)$ distribution under H_0. Thus, if we decide to reject H_0 when $Z_0 > z_\alpha$, then the probability of committing a type I error is α. As a consequence, we apply the decision rule

$$d_1 : \text{reject } H_0 \text{ if and only if } \bar{X} > \mu_0 + z_\alpha \frac{\sigma}{\sqrt{n}}$$

Similarly, from the distribution of Z_0 under H_0 we can obtain the critical region for the other types of hypotheses. When σ^2 is unknown, then by Fact 4 $T_0 = \sqrt{n}(\bar{X} - \mu_0)/S$ has a $t(n - 1)$ distribution under H_0. Thus the corresponding tests can be obtained by substituting $t_\alpha(n - 1)$ for z_α and S for σ. The tests for the various one-sided and two-sided hypotheses are summarized in Table 19.11.2 below. For each set of hypotheses, the critical region given on the first line is for the case when σ^2 is known, and that

TABLE 19.11.2 One-Sample Tests for Mean

Null Hypothesis H_0	Alternative Hypothesis H_1	Critical Region
$\mu = \mu_0$ or $\mu \le \mu_0$	$\mu = \mu_1 > \mu_0$ or $\mu > \mu_0$	$\bar{X} > \mu_0 + z_\alpha \frac{\sigma}{\sqrt{n}}$
		$\bar{X} > \mu_0 + t_\alpha \frac{S}{\sqrt{n}}$
$\mu = \mu_0$ or $\mu \ge \mu_0$	$\mu = \mu_1 < \mu_0$ or $\mu < \mu_0$	$\bar{X} < \mu_0 - z_\alpha \frac{\sigma}{\sqrt{n}}$
		$\bar{X} < \mu_0 - t_\alpha \frac{S}{\sqrt{n}}$
$\mu = \mu_0$	$\mu \ne \mu_0$	$\lvert \bar{X} - \mu_0 \rvert > z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$
		$\lvert \bar{X} - \mu_0 \rvert > t_{\alpha/2} \frac{S}{\sqrt{n}}$

given on the second line is for the case when σ^2 is unknown. Furthermore, t_α and $t_{\alpha/2}$ stand for $t_\alpha(n-1)$ and $t_{\alpha/2}(n-1)$, respectively.

2. Test for Variance. In testing hypotheses concerning the variance σ^2 of a normal distribution, use Fact 5 to assert that, under H_0: $\sigma^2 = \sigma_0^2$, the distribution of $w_0 = (n-1)\,S^2/\sigma_0^2$ is $\chi^2(n-1)$. The corresponding tests and critical regions are summarized in the following table (χ^2_α and $\chi^2_{\alpha/2}$ stand for $\chi^2_\alpha(n-1)$ and $\chi^2_{\alpha/2}(n-1)$, respectively):

TABLE 19.11.3 One-Sample Tests for Variance

Null Hypothesis H_0	Alternative Hypothesis H_1	Critical Region
$\sigma^2 = \sigma_0^2$ or $\sigma^2 \le \sigma_0^2$	$\sigma^2 = \sigma_1^2 > \sigma_0^2$ or $\sigma^2 > \sigma_0^2$	$\left(S^2/\sigma_0^2\right) > \frac{1}{n-1}\chi^2_\alpha$
$\sigma^2 = \sigma_0^2$ or $\sigma^2 \ge \sigma_0^2$	$\sigma^2 = \sigma_1^2 < \sigma_0^2$ or $\sigma^2 < \sigma_0^2$	$\left(S^2/\sigma_0^2\right) < \frac{1}{n-1}\chi^2_{1-\alpha}$
$\sigma^2 = \sigma_0^2$	$\sigma^2 \ne \sigma_0^2$	$\left(S^2/\sigma_0^2\right) > \frac{1}{n-1}\chi^2_{\alpha/2}$
		or $\left(S^2/\sigma_0^2\right) < \frac{1}{n-1}\chi^2_{1-\alpha/2}$

Two-Sample Case

In comparing the means and variances of two normal populations, we once again refer to Table 19.11.1 for notation and assumptions.

1. Test for Difference of Two Means. Let $\delta = \mu_1 - \mu_2$ be the difference of the two population means. In testing H_0: $\delta = \delta_0$ vs. a one-sided or two-sided alternative hypothesis, we note that, for

$$\tau = \left(\sigma_1^2/n_1 + \sigma_2^2/n_2\right)^{1/2} \tag{19.11.23}$$

and

$$\nu = S_p\left(1/n_1 + 1/n_2\right)^{1/2} \tag{19.11.24}$$

$Z_0 = [(\bar{X}_1 - \bar{X}_2) - \delta_0]/\tau$ has an $N(0,1)$ distribution under H_0 and $T_0 = [(\bar{X}_1 - \bar{X}_2) - \delta_0]/\nu$ has a $t(n_1 + n_2 - 2)$ distribution under H_0 when $\sigma_1^2 = \sigma_2^2$. Using these results, the corresponding critical regions for one-sided and two-sided tests can be obtained, and they are listed below. Note that, as in the one-sample case, the critical region given on the first line for each set of hypotheses is for the case of known variances, and that given on the second line is for the case in which the variances are equal but unknown. Further, t_α and $t_{\alpha/2}$ stand for $t_\alpha(n_1 + n_2 - 2)$ and $t_{\alpha/2}(n_1 + n_2 - 2)$, respectively.

TABLE 19.11.4 Two-Sample Tests for Difference of Two Means

Null Hypothesis H_0	Alternative Hypothesis H_1	Critical Region
$\delta = \delta$ or $\delta \le \delta_0$	$\delta = \delta_1 > \delta_0$ or $\delta > \delta_0$	$(\bar{X}_1 - \bar{X}_2) > \delta_0 + z_\alpha \tau$
		$(\bar{X}_1 - \bar{X}_2) > \delta_0 + t_\alpha \nu$
$\delta = \delta_0$ or $\delta \ge \delta_0$	$\delta = \delta_1 < \delta_0$ or $\delta < \delta_0$	$(\bar{X}_1 - \bar{X}_2) < \delta_0 - z_\alpha \tau$
		$(\bar{X}_1 - \bar{X}_2) < \delta_0 - t_\alpha \nu$
$\delta = \delta_0$	$\delta \ne \delta_0$	$\lvert(\bar{X}_1 - \bar{X}_2) - \delta_0\rvert > z_{\alpha/2} \tau$
		$\lvert(\bar{X}_1 - \bar{X}_2) - \delta_0\rvert > t_{\alpha/2} \nu$

A Numerical Example

In the following we provide a numerical example for illustrating the construction of confidence intervals and hypothesis-testing procedures. The example is given along the line of applications in Wadsworth (1990, p. 4.21) with artificial data.

Suppose that two processes (T_1 and T_2) manufacturing steel pins are in operation, and that a random sample of 4 pins (or 5 pins) was taken from the process T_1 (the process T_2) with the following results (in units of inches):

$$T_1 : 0.7608,\ 0.7596,\ 0.7622,\ 0.7638$$

$$T_2 : 0.7546,\ 0.7561,\ 0.7526,\ 0.7572,\ 0.7565$$

Simple calculation shows that the observed values of sample means sample variances, and sample standard deviations are:

$$\bar{X}_1 = 0.7616, \quad S_1^2 = 3.280 \times 10^{-6}, \quad S_1 = 1.811 \times 10^{-3}$$

$$\bar{X}_2 = 0.7554, \quad S_2^2 = 3.355 \times 10^{-6}, \quad S_2 = 1.832 \times 10^{-3}$$

One-Sample Case

Let us first consider confidence intervals for the parameters of the first process, T_1, only.

1. Assume that, based on previous knowledge of processes of this type, the variance is known to be $\sigma_1^2 = 1.80^2 \times 10^{-6}$ ($\sigma_1 = 0.0018$). Then from the normal table (see, e.g., Ross (1987, p. 482) we have $z_{0.025} = 1.96$. Thus a 95% confidence interval for μ_1 is

$$\left(0.7616 - 1.96 \times 0.0018/\sqrt{4},\ 0.7616 + 1.96 \times 0.0018/\sqrt{4}\right)$$

or (0.7598, 0.7634) (after rounding off to the 4th decimal place).

2. If σ_1^2 is unknown and a 95% confidence interval for μ_1 is needed then, for $t_{0.023}(3) = 3.182$ (see, e.g., Ross, 1987, p. 484) the confidence interval is

$$\left(0.7616 - 3.182 \times 0.001811/\sqrt{4},\ 0.7616 + 3.182 \times 0.001811/\sqrt{4}\right)$$

or (0.7587, 0.7645)

3. From the chi-squared table with 4 – 1 = 3 degrees of freedom, we have (see, e.g., Ross, 1987, p. 483) $\chi^2_{0.975} = 0.216$, $\chi^2_{0.025} = 9.348$. Thus a 95% confidence interval for σ_1^2 is $(3 \times 3.280 \times 10^{-6}/9.348,\ 3 \times 3.280 \times 10^{-6}/0.216)$, or $(1.0526 \times 10^{-6},\ 45{,}5556 \times 10^{-6})$.
4. In testing the hypotheses

$$H_0 : \mu_1 = 0.76 \text{ vs. } H_1 : \mu_1 > 0.76$$

with a = 0.01 when σ_1^2 is unknown, the critical region is $\bar{x}_1 > 0.76 + 4.541 \times 0.001811/\sqrt{4} = 0.7641$. Since the observed value $\bar{x}_1$ is 0.7616, H_0 is accepted. That is, we assert that there is no significant evidence to call for the rejection of H_0.

Two-Sample Case

If we assume that the two populations have a common unknown variance, we can use the Student's *t* distribution (with degree of freedom $\nu = 4 + 5 - 2 = 7$) to obtain confidence intervals and to test hypotheses for $\mu_1 - \mu_2$. We first note that the data given above yield

$$S_p^2 = \frac{1}{7}(3 \times 3.280 + 4 \times 3.355) \times 10^{-6}$$

$$= 3.3229 \times 10^{-6}$$

$$S_p = 1.8229 \times 10^{-3} \qquad \nu = S_p\sqrt{1/4 + 1/5} = 1.2228 \times 10^{-3}$$

and $\bar{X}_1 - \bar{X}_2 = 0.0062$.

1. A 98% confidence interval for $\mu_1 - \mu_2$ is $(0.0062 - 2.998\nu,\ 0.0062 + 2.998\nu)$ or (0.0025, 0.0099).
2. In testing the hypotheses H_0: $\mu_1 = \mu_2$ (i.e., $\mu_1 - \mu_2 = 0$) vs. H_1: $\mu_1 > \mu_2$ with $\alpha = 0.05$, the critical region is $(\bar{X}_1 - \bar{X}_2) > 1.895\nu = 2.3172 \times 10^{-3}$. Thus H_0 is rejected; i.e., we conclude that there is significant evidence to indicate that $\mu_1 > \mu_2$ may be true.
3. In testing the hypotheses H_0: $\mu_1 = \mu_2$ vs. $\mu_1 \neq \mu_2$ with $\alpha = 0.02$, the critical region is $|\bar{X}_1 - \bar{X}_2| > 2.998\nu = 3.6660 \times 10^{-3}$. Thus H_0 is rejected. We note that the conclusion here is consistent with the result that, with confidence probability $1 - \alpha = 0.98$, the confidence interval for $(\mu_1 - \mu_2)$ does not contain the origin.

Concluding Remarks

The history of probability and statistics goes back to the days of the celebrated mathematicians K. F. Gauss and P. S. Laplace. (The normal distribution, in fact, is also called the Gaussian distribution.) The theory and methods of classical statistical analysis began its developments in the late 1800s and early 1900s when F. Galton and R.A. Fisher applied statistics to their research in genetics, when Karl Pearson developed the chi-square goodness-of-fit method for stochastic modeling, and when E.S. Pearson and J. Neyman developed the theory of hypotheses testing. Today statistical methods have been found useful in analyzing experimental data in biological science and medicine, engineering, social sciences, and many other fields. A non-technical review on some of the applications is Hacking (1984).

Applications of statistics in engineering include many topics. In addition to those treated in this section, other important ones include sampling inspection and quality (process) control, reliability, regression analysis and prediction, design of engineering experiments, and analysis of variance. Due to space limitations, these topics are not treated here. The reader is referred to textbooks in this area for further information. There are many well-written books that cover most of these topics, the following short list consists of a small sample of them.

References

Box, G.E.P., Hunter, W.G., and Hunter, J.S. 1978. *Statistics for Experimenters.* John Wiley & Sons, New York.

Bowker, A.H. and Lieberman, G.J. 1972. *Engineering Statistics*, 2nd ed. Prentice-Hall, Englewood Cliffs, NJ.

Hacking, I. 1984. Trial by number, *Science*, 84(5), 69–70.

Hahn, G.J. and Shapiro, S.S. 1967. *Statistical Models in Engineering.* John Wiley & Sons, New York.

Hines, W.W. and Montgomery, D.G. 1980. *Probability and Statistics in Engineering and Management Science.* John Wiley & Sons, New York.

Hogg, R.V. and Ledolter, J. 1992. *Engineering Statistics.* Macmillan, New York.

Ross, S.M. 1987. *Introduction to Probability and Statistics for Engineers and Scientists.* John Wiley & Sons, New York.

Wadsworth, H.M., Ed. 1990. *Handbook of Statistical Methods for Engineers and Scientists.* John Wiley & Sons, New York.

19.12 Numerical Methods

William F. Ames

Introduction

Since many mathematical models of physical phenomena are not solvable by available mathematical methods one must often resort to approximate or numerical methods. These procedures do not yield exact results in the mathematical sense. This inexact nature of numerical results means we must pay attention to the errors. The two errors that concern us here are *round-off errors* and *truncation errors.*

Round-off errors arise as a consequence of using a number specified by m correct digits to approximate a number which requires more than m digits for its exact specification. For example, using 3.14159 to approximate the irrational number π. Such errors may be especially serious in matrix inversion or in any area where a very large number of numerical operations are required. Some attempts at handling these errors are called *enclosure methods.* (Adams and Kulisch, 1993).

Truncation errors arise from the substitution of a finite number of steps for an infinite sequence of steps (usually an iteration) which would yield the exact result. For example, the iteration $y_n(x) = 1 + \int_0^x xty_{n-1}(t)dt$, $y(0) = 1$ is only carried out for a *few steps*, but it converges in *infinitely* many steps.

The study of some errors in a computation is related to the theory of probability. In what follows, a relation for the error will be given in certain instances.

Linear Algebra Equations

A problem often met is the determination of the solution vector $u = (u_1, u_2, \ldots, u_n)^T$ for the set of linear equations $Au = v$ where A is the $n \times n$ square matrix with coefficients, a_{ij} $(i, j = 1, \ldots, n)$, $v = (v_1, \ldots, v_n)^T$ and i denotes the row index and j the column index.

There are many numerical methods for finding the solution, u, of $Au = v$. The direct inversion of A is usually too expensive and is not often carried out unless it is needed elsewhere. We shall only list a few methods. One can check the literature for the many methods and computer software available. Some of the software is listed in the References section at the end of this chapter. The methods are usually subdivided into *direct* (once through) or *iterative* (repeated) procedures.

In what follows, it will often be convenient to partition the matrix A into the form $A = U + D + L$, where U, D, and L are matrices having the same elements as A, respectively, above the main diagonal, on the main diagonal, and below the main diagonal, and zeros elsewhere. Thus,

$$U = \begin{bmatrix} 0 & a_{12} & & \cdots & a_{1n} \\ 0 & 0 & a_{23} & \cdots & a_{2n} \\ \vdots & \cdots & & \cdots & \\ 0 & 0 & \cdots & \cdots & 0 \end{bmatrix}$$

We also assume the u_js are not all zero and $\det A \neq 0$ so the solution is unique.

Direct Methods

Gauss Reduction. This classical method has spawned many variations. It consists of dividing the first equation by a_{11} (if $a_{11} = 0$, reorder the equations to find an $a_{11} \neq 0$) and using the result to eliminate the terms in u_1 from each of the succeeding equations. Next, the modified second equation is divided by a'_{22} (if $a'_{22} = 0$, a reordering of the modified equations may be necessary) and the resulting equation is used to eliminate all terms in u_2 in the succeeding modified equations. This elimination is done n times resulting in a triangular system:

$$
\begin{aligned}
u_1 + a'_{12}u_2 + \cdots + a'_{1n}u_n &= v'_1 \\
0 + u_2 + \cdots + a'_{2n}u_n &= v'_2 \\
&\cdots \\
0 + \cdots + u_{n-1} + a'_{n-1,n}u_n &= v'_{n-1} \\
u_n &= v'_n
\end{aligned}
$$

where a'_{ij} and v'_j represent the specific numerical values obtained by this process. The solution is obtained by working backward from the last equation. Various modifications, such as the Gauss-Jordan reduction, the Gauss-Doolittle reduction, and the Crout reduction, are described in the classical reference authored by Bodewig (1956). Direct methods prove very useful for sparse matrices and banded matrices that often arise in numerical calculation for differential equations. Many of these are available in computer packages such as IMSL, Maple, Matlab, and Mathematica.

The Tridiagonal Algorithm. When the linear equations are tridiagonal, the system

$$
\begin{aligned}
b_1u_1 + c_1u_2 &= d_1 \\
a_iu_{i-1} + b_iu_i + c_iu_{i+1} &= d_i \\
a_nu_{n-1} + b_nu_n &= d_n, \quad i = 2, 3, \ldots, n-1
\end{aligned}
$$

can be solved explicitly for the unknown, thereby eliminating any matrix operations.

The Gaussian elimination process transforms the system into a simpler one of *upper bidiagonal* form. We designate the coefficients of this new system by a'_i, b'_i c'_i and d'_i, and we note that

$$
\begin{aligned}
a'_i &= 0, \quad i = 2, 3, \ldots, n \\
b'_i &= 1, \quad i = 1, 2, \ldots, n
\end{aligned}
$$

The coefficients c'_i and d'_i are calculated successively from the relations

$$
c'_1 = \frac{c_1}{b_1} \qquad d'_1 = \frac{d_1}{b_1}
$$

$$
c'_{i+1} = \frac{c_{i+1}}{b_{i+1} - a_{i+1}c'_i}
$$

$$
d'_{i+1} = \frac{d_{i+1} - a_{i+1}d'_i}{b_{i+1} - a_{i+1}c'_i}, \quad i = 1, 2, \ldots, n-1
$$

and, of course, $c_n = 0$.

Having completed the elimination we examine the new system and see that the nth equation is now

$$
u_n = d'_n
$$

Substituting this value into the $(n - 1)$st equation,

$$
u_{n-1} + c'_{n-1}u_n = d'_{n-1}
$$

we have

$$u_{n-1} = d'_{n-1} - c'_{n-1}u_n$$

Thus, starting with u_n, we have successively the solution for u_i as

$$u_i = d'_i - c'_i u_{i+1}, \quad i = n-1,\ n-2,\ \ldots,\ 1$$

Algorithm for Pentadiagonal Matrix. The equations to be solved are

$$a_i u_{i-2} + b_i u_{i-1} + c_i u_i + d_i u_{i+1} + e_i u_{i+2} = f_i$$

for $1 \le i \le R$ with $a_1 = b_1 = a_2 = e_{R-1} = d_R = e_R = 0$.

The algorithm is as follows. First, compute

$$\delta_1 = d_1/c_1$$
$$\lambda_1 = e_1/c_1$$
$$\gamma_1 = f_1/c_1$$

and

$$\mu_2 = c_2 - b_2\delta_1$$
$$\delta_2 = \left(d_2 - b_2\lambda_1\right)/\mu_2$$
$$\lambda_2 = e_2/\mu_2$$
$$\gamma_2 = \left(f - b_2\gamma_1\right)/\mu_2$$

Then, for $3 \le i \le R-2$, compute

$$\beta_i = b_i - a_i\delta_{i-2}$$
$$\mu_i = c_i - \beta_i\delta_{i-1} - a_i\lambda_{i-2}$$
$$\delta_i = \left(d_i - \beta_i\lambda_{i-1}\right)/\mu_i$$
$$\lambda_i = e_i/\mu_i$$
$$\gamma_i = \left(f_i - \beta_i\gamma_{i-1} - a_i\gamma_{i-2}\right)/\mu_i$$

Next, compute

$$\beta_{R-1} = b_{R-1} - a_{R-1}\delta_{R-3}$$
$$\mu_{R-1} = c_{R-1} - \beta_{R-1}\delta_{R-2} - a_{R-1}\lambda_{R-3}$$
$$\delta_{R-1} = \left(d_{R-1} - \beta_{R-1}\lambda_{R-2}\right)/\mu_{R-1}$$
$$\gamma_{R-1} = \left(f_{R-1} - \beta_{R-1}\gamma_{R-2} - a_{R-1}\gamma_{R-3}\right)/\mu_{R-1}$$

and

$$\beta_R = b_R - a_R\delta_{R-2}$$
$$\mu_R = c_R - \beta_R\delta_{R-1} - a_R\lambda_{R-2}$$
$$\gamma_R = \left(f_R - \beta_R\gamma_{R-1} - a_R\gamma_{R-2}\right)/\mu_R$$

The β_i and μ_i are used only to compute δ_i, λ_i, and γ_i, and need not be stored after they are computed. The δ_i, λ_i, and γ_i, must be stored, as they are used in the back solution. This is

$$u_R = \gamma_R$$
$$u_{R-1} = \gamma_{R-1} - \delta_{R-1}u_R$$

and

$$u_i = \gamma_i - \delta_i u_{i+1} - \lambda_i u_{i+2}$$

for $R - 2 \geq i \geq 1$.

General Band Algorithm. The equations are of the form

$$A_j^{(M)}X_{j-M} + A_j^{(M-1)}X_{j-M+1} + \cdots + A_j^{(2)}X_{j-2} + A_j^{(1)}X_{j-1} + B_jX_j$$
$$+ C_j^{(1)}X_{j+1} + C_j^{(2)}X_{j+2} + \cdots + C_j^{(M-1)}X_{j+M-1} + C_j^{(M)}X_{j+M} = D_j$$

for $1 \leq j \leq N$, $N \geq M$. The algorithm used is as follows:

$$\alpha_j^{(k)} = A_j^{(k)} = 0, \quad \text{for } k \geq j$$
$$C_j^{(k)} = 0, \quad \text{for } k \geq N + 1 - j$$

The forward solution ($j = 1, \ldots, N$) is

$$\alpha_j^{(k)} = A_j^{(k)} - \sum_{p=k+1}^{p=M} \alpha_j^{(p)} W_{j-p}^{(p-k)}, \quad k = M, \ldots, 1$$
$$\beta_j = B_j - \sum_{p=1}^{M} \alpha_j^{(p)} W_{j-p}^{(p)}$$
$$W_j^{(k)} = \left(C_j^{(k)} - \sum_{p=k+1}^{p=M} \alpha_j^{(p-k)} W_{j-(p-k)}^{(p)}\right)\Big/\beta_j, \quad k = 1, \ldots, M$$
$$\gamma_j = \left(D_j - \sum_{p=1}^{M} \alpha_j^{(p)}\gamma_{j-p}\right)\Big/\beta_j$$

The back solution ($j = N, \ldots, 1$) is

$$X_j = \gamma_j - \sum_{p=1}^{M} W_j^{(p)} X_{j+p}$$

Cholesky Decomposition. When the matrix A is a symmetric and positive definite, as it is for many discretizations of self-adjoint positive definite boundary value problems, one can improve considerably on the band procedures by using the Cholesky decomposition. For the system $Au = v$, the Matrix A can be written in the form

$$A = (I + L)D(I + U)$$

where L is lower triangular, U is upper triangular, and D is diagonal. If $A = A'$ (A' represents the transpose of A), then

$$A = A' = (I + U)' D(I + L)'$$

Hence, because of the uniqueness of the decomposition.

$$I + L = (I + U)' = I + U'$$

and therefore,

$$A = (I + U)' D(I + U)$$

that is,

$$A = B'B, \text{ where } B = \sqrt{D}(I + U)$$

The system $Au = v$ is then solved by solving the two triangular system

$$B'w = v$$

followed by

$$Bu = w$$

To carry out the decomposition $A = B'B$, all elements of the first row of A, and of the derived system, are divided by the square root of the (positive) leading coefficient. This yields smaller rounding errors than the banded methods because the relative error of $\sqrt{a}$ is only half as large as that of a itself. Also, taking the square root brings numbers nearer to each other (i.e., the new coefficients do not differ as widely as the original ones do). The actual computation of $B = (b_{ij})$, $j > i$, is given in the following:

$$b_{11} = \left(a_{11}\right)^{1/2}, \qquad b_{1j} = a_{ij}/b_{11}, \quad j \geq 2$$

$$b_{22} = \left(a_{22} - b_{12}^2\right)^{1/2}, \qquad b_{2j} = \left(a_{2j} - b_{12}b_{1j}\right)\Big/b_{22}$$

$$b_{33} = \left(a_{33} - b_{13}^2 - b_{23}^2\right)^{1/2}, \qquad b_{3j} = \left(a_{3j} - b_{13}b_{1j} - b_{23}b_{2j}\right)\Big/b_{33}$$

$$\vdots$$

$$b_{ii} = \left(a_{ii} - \sum_{k=1}^{i-1} b_{ki}^2\right)^{1/2}, \qquad b_{ij} = \left(a_{ij} - \sum_{k=1}^{i-1} b_{ki}b_{kj}\right)\Big/b_{ii}, \qquad i \geq 2,\ j \geq 2$$

Iterative Methods

Iterative methods consist of repeated application of an often simple algorithm. They yield the exact answer only as the limit of a sequence. They can be programmed to take care of zeros in A and are self-correcting. Their structure permits the use of convergence accelerators, such as overrelaxation, Aitkins acceleration, or Chebyshev acceleration.

Let $a_{ii} > 0$ for all i and $\det A \neq 0$. With $A = U + D + L$ as previously described, several iteration methods are described for $(U + D + L)u = v$.

Jacobi Method (Iteration by total steps). Since $u = -D^{-1}[U + L]u + D^{-1}v$, the iteration $u^{(k)}$ is $u^{(k)} = -D^{-1}[U + L]u^{(k-1)} + D^{-1}v$. This procedure has a slow convergent rate designated by R, $0 < R \ll 1$.

Gauss-Seidel Method (Iteration by single steps). $u^{(k)} = -(L + D)^{-1}Uu^{(k-1)} + (L + D)^{-1}v$. Convergence rate is $2R$, twice as fast as that of the Jacobi method.

Gauss-Seidel with Successive Overrelaxation (SOR). Let $\bar{u}_i^{(k)}$ be the ith components of the Gauss-Seidel iteration. The SOR technique is defined by

$$u_i^{(k)} = (1 - \omega)u_i^{(k-1)} + \omega\bar{u}_i^{(k)}$$

where $1 < \omega < 2$ is the overrelaxation parameter. The full iteration is $u^{(k)} = (D + \omega L)^{-1}\{[(1 - \omega)D - \omega U]u^{(k-1)} + \omega v\}$. Optimal values of ω can be computed and depend upon the properties of A (Ames, 1993). With optimal values of ω, the convergence rate of this method is $2R\sqrt{2}$ which is much larger than that for Gauss-Seidel (R is usually much less than one).

For other acceleration techniques, see the literature (Ames, 1993).

Nonlinear Equations in One Variable

Special Methods for Polynomials

The polynomial $P(x) = a_0x^n + a_1x^{n-1} + \cdots + a_{n-1}x + a_n = 0$, with real coefficients a_j, $j = 0, \ldots, n$, has exactly n roots which may be real or complex.

If all the coefficients of $P(x)$ are integers, then any rational roots, say r/s (r and s are integers with no common factors), of $P(x) = 0$ must be such that r is an integral divisor of a_n and s is an integral division of a_0. Any polynomial with rational coefficients may be converted into one with integral coefficients by multiplying the polynomial by the lowest common multiple of the denominators of the coefficients.

Example. $x^4 - 5x^2/3 + x/5 + 3 = 0$. The lowest common multiple of the denominators is 15. Multiplying by 15, which does not change the roots, gives $15x^4 - 25x^2 + 3x + 45 = 0$. The only possible rational roots r/s are such that r may have the value ±45, ±15, ±5, ±3, and ±1, while s may have the values ±15,

±5, ±3, and ±1. All possible rational roots, with no common factors, are formed using all possible quotients.

If $a_0 > 0$, the first negative coefficient is preceded by k coefficients which are positive or zero, and G is the largest of the absolute values of the negative coefficients, then each real root is less than $1 + \sqrt[k]{G/a_0}$ (upper bound on the real roots). For a lower bound to the real roots, apply the criterion to $P(-x) = 0$.

Example. $P(x) = x^5 + 3x^4 - 2x^3 - 12x + 2 = 0$. Here $a_0 = 1$, $G = 12$, and $k = 2$. Thus, the upper bound for the real roots is $1 + \sqrt[2]{12} \approx 4.464$. For the lower bound, $P(-x) = -x^5 + 3x^4 + 2x^3 + 12x + 2 = 0$, which is equivalent to $x^5 - 3x^4 - 2x^3 - 12x - 2 = 0$. Here $k = 1$, $G = 12$, and $a_0 = 1$. A lower bound is $-(1 + 12) = 13$. Hence all real roots lie in $-13 < x < 1 + \sqrt[2]{12}$.

A useful *Descartes rule of signs* for the number of positive or negative real roots is available by observation for polynomials with real coefficients. The number of positive real roots is either equal to the number of sign changes, n, or is less than n by a positive *even* integer. The number of negative real roots is either equal to the number of sign changes, n, of $P(-x)$, or is less than n by a positive even integer.

Example. $P(x) = x^5 - 3x^3 - 2x^2 + x - 1 = 0$. There are three sign changes, so $P(x)$ has either three or one positive roots. Since $P(-x) = -x^5 + 3x^3 - 2x^2 - 1 = 0$, there are either two or zero negative roots.

The Graeffe Root-Squaring Technique

This is an iterative method for finding the roots of the algebraic equation

$$f(x) = a_0 x^p + a_1 x^{p-1} + \cdots + a_{p-1} x + a_p = 0$$

If the roots are r_1, r_2, r_3, …, then one can write

$$S_p = r_1^p \left(1 + \frac{r_2^p}{r_1^p} + \frac{r_3^p}{r_1^p} + \cdots \right)$$

and if one root is larger than all the others, say r_1, then for large enough p all terms (other than 1) would become negligible. Thus,

$$S_p \approx r_1^p$$

or

$$\lim_{p \to \infty} S_p^{1/p} = r_1$$

The Graeffe procedure provides an efficient way for computing S_p via a sequence of equations such that the roots of each equation are the squares of the roots of the preceding equations in the sequence. This serves the purpose of ultimately obtaining an equation whose roots are so widely separated in magnitude that they may be read approximately from the equation by inspection. The basic procedure is illustrated for a polynomial of degree 4:

$$f(x) = a_0 x^4 + a_1 x^3 + a_2 x^2 + a_3 x + a_4 = 0$$

Rewrite this as

$$a_0 x^4 + a_2 x^2 + a_4 = -a_1 x^3 - a_3 x$$

and square both sides so that upon grouping

$$a_0^2x^8 + \left(2a_0a_2 - a_1^2\right)x^6 + \left(2a_0a_4 - 2a_1a_3 + a_2^2\right)x^4 + \left(2a_2a_4 - a_3^2\right)x^2 + a_4^2 = 0$$

Because this involves only even powers of x, we may set $y = x^2$ and rewrite it as

$$a_0^2y^4 + \left(2a_0a_2 - a_1^2\right)y^3 + \left(2a_0a_4 - 2a_1a_3 + a_2^2\right)y^2 + \left(2a_2a_4 - a_3^2\right)y + a_4^2 = 0$$

whose roots are the squares of the original equation. If we repeat this process again, the new equation has roots which are the fourth power, and so on. After p such operations, the roots are 2^p (original roots). If at any stage we write the coefficients of the unknown in sequence

$$a_0^{(p)} \quad a_1^{(p)} \quad a_2^{(p)} \quad a_3^{(p)} \quad a_4^{(p)}$$

then, to get the new sequence $a_i^{(p+1)}$, write $a_i^{(p+1)} = 2a_0^{(p)}$ (times the symmetric coefficient) with respect to $a_i^{(p)} - 2a_1^{(p)}$ (times the symmetric coefficient) $- \cdots (-1)^i a_i^{(p)2}$. Now if the roots are r_1, r_2, r_3, and r_4, then $a_1/a_0 = -\Sigma_{i=1}^4 r_i$, $a_i^{(1)}/a_0^{(1)} = -\Sigma r_i^2$, …, $a_i^{(p)}/a_0^{(p)} = -\Sigma r_i^{2p}$. If the roots are all distinct and r_1 is the largest in magnitude, then eventually

$$r_1^{2p} \approx -\frac{a_1^{(p)}}{a_0^{(p)}}$$

And if r_2 is the next largest in magnitude, then

$$r_2^{2p} \approx -\frac{a_2^{(p)}}{a_1^{(p)}}$$

And, in general $a_n^{(p)}/a_{n-1}^{(p)} \approx -r_n^{2p}$. This procedure is easily generalized to polynomials of arbitrary degree and specialized to the case of multiple and complex roots.

Other methods include Bernoulli iteration, Bairstow iteration, and Lin iteration. These may be found in the cited literature. In addition, the methods given below may be used for the numerical solution of polynomials.

General Methods for Nonlinear Equations in One Variable

Successive Substitutions

Let $f(x) = 0$ be the nonlinear equation to be solved. If this is rewritten as $x = F(x)$, then an iterative scheme can be set up in the form $x_{k+1} = F(x_k)$. To start the iteration, an initial guess must be obtained graphically or otherwise. The convergence or divergence of the procedure depends upon the method of writing $x = F(x)$, of which there will usually be several forms. A general rule to ensure convergence cannot be given. However, if a is a root of $f(x) = 0$, a necessary condition for convergence is that $|F'(x)| < 1$ in that interval about a in which the iteration proceeds (this means the iteration cannot converge unless $|F'(x)| < 1$, but it does not ensure convergence). This process is called *first order* because the error in x_{k+1} is proportional to the first power of the error in x_k.

Example. $f(x) = x^3 - x - 1 = 0$. A rough plot shows a real root of approximately 1.3. The equation can be written in the form $x = F(x)$ in several ways, such as $x = x^3 - 1$, $x = 1/(x^2 - 1)$, and $x = (1 + x)^{1/3}$. In the first case, $F'(x) = 3x^2 = 5.07$ at $x = 1.3$; in the second, $F'(1.3) = 5.46$; only in the third case

is $F'(1.3) < 1$. Hence, only the third iterative process has a chance to converge. This is illustrated in the iteration table below.

Step k	$x = \dfrac{1}{x^2 - 1}$	$x = x^3 - 1$	$x = (1 + x)^{1/3}$
0	1.3	1.3	1.3
1	1.4493	1.197	1.32
2	0.9087	0.7150	1.3238
3	–5.737	–0.6345	1.3247
4	⋯	⋯	1.3247

Numerical Solution of Simultaneous Nonlinear Equations

The techniques illustrated here will be demonstrated for two simultaneous equations — $f(x, y) = 0$ and $g(x, y) = 0$. They immediately generalize to more than two simultaneous equations.

The Method of Successive Substitutions

The two simultaneous equations can be written in various ways in equivalent forms

$$x = F(x,y)$$

$$y = G(x,y)$$

and the method of successive substitutions can be based on

$$x_{k+1} = F(x_k, y_k)$$

$$y_{k+1} = G(x_k, y_k)$$

Again, the procedure is of the first order and a necessary condition for convergence is

$$\left|\frac{\partial F}{\partial x}\right| + \left|\frac{\partial F}{\partial y}\right| < 1 \qquad \left|\frac{\partial G}{\partial x}\right| + \left|\frac{\partial G}{\partial y}\right| < 1$$

in the iteration neighborhood of the true solution.

The Newton-Raphson Procedure

Using the two simultaneous equation, start from an approximate, say (x_0, y_0), obtained graphically or from a two-way table. Then, solve successively the linear equations

$$\Delta x_k \frac{\partial f}{\partial x}(x_k, y_k) + \Delta y_k \frac{\partial f}{\partial y}(x_k, y_k) = -f(x_k, y_k)$$

$$\Delta x_k \frac{\partial g}{\partial x}(x_k, y_k) + \Delta y_k \frac{\partial g}{\partial y}(x_k, y_k) = -g(x_k, y_k)$$

for Δx_k and Δy_k. Then, the $k + 1$ approximation is given from $x_{k+1} = x_k + \Delta x_k, y_{k+1} = y_k + \Delta y_k$. A modification consists in solving the equations with (x_k, y_k) replaced by (x_0, y_0) (or another suitable pair later on in the iteration) in the derivatives. This means the derivatives (and therefore the coefficients of Δx_k, Δy_k) are independent of k. Hence, the results become

$$\Delta x_k = \frac{-f(x_k,y_k)(\partial g/\partial y)(x_0,y_0) + g(x_k,y_k)(\partial f/\partial y)(x_0,y_0)}{(\partial f/\partial x)(x_0,y_0)(\partial g/\partial y)(x_0,y_0) - (\partial f/\partial y)(x_0,y_0)(\partial g/\partial x)(x_0,y_0)}$$

$$\Delta y_k = \frac{-g(x_k,y_k)(\partial f/\partial x)(x_0,y_0) + f(x_k,y_k)(\partial g/\partial x)(x_0,y_0)}{(\partial f/\partial x)(x_0,y_0)(\partial g/\partial y)(x_0,y_0) - (\partial f/\partial y)(x_0,y_0)(\partial g/\partial x)(x_0,y_0)}$$

and $x_{k+1} = \Delta x_k + x_k$, $y_{k+1} = \Delta y_k + y_k$. Such an alteration of the basic technique reduces the rapidity of convergence.

Example

$$f(x,y) = 4x^2 + 6x - 4xy + 2y^2 - 3$$

$$g(x,y) = 2x^2 - 4xy + y^2$$

By plotting, one of the approximate roots is found to be $x_0 = 0.4$, $y_0 = 0.3$. At this point, there results $\partial f/\partial x = 8$, $\partial f/\partial y = -0.4$, $\partial g/\partial x = 0.4$, and $\partial g/\partial y = -1$. Hence,

$$x_{k+1} = x_k + \Delta x_k = x_k + \frac{-f(x_k,y_k) - 0.4g(x_k,y_k)}{8(-1) - (-0.4)(0.4)}$$

$$= x_k - 0.127\,55 f(x_k,y_k) - 0.051\,02 g(x_k,y_k)$$

and

$$y_{k+1} = y_k - 0.051\,02 f(x_k,y_k) + 1.020\,41 g(x_k,y_k)$$

The first few iteration steps are shown in the following table.

Step k	x_k	y_k	$f(x_k, y_k)$	$g(x_k, y_k)$
0	0.4	0.3	–0.26	0.07
1	0.43673	0.24184	0.078	0.0175
2	0.42672	0.25573	–0.0170	–0.007
3	0.42925	0.24943	0.0077	0.0010

Methods of Perturbation

Let $f(x) = 0$ be the equation. In general, the iterative relation is

$$x_{k+1} = x_k - \frac{f(x_k)}{\alpha_k}$$

where the iteration begins with x_0 as an initial approximation and α_k is some functional.

The Newton-Raphson Procedure. This variant chooses $\alpha_k = f'(x_k)$ where $f' = df/dx$ and geometrically consists of replacing the graph of $f(x)$ by the tangent line at $x = x_k$ in each successive step. If $f'(x)$ and $f''(x)$ have the same sign throughout an interval $a \le x \le b$ containing the solution, with $f(a)$ and $f(b)$ of opposite signs, then the process converges starting from any x_0 in the interval $a \le x \le b$. The process is second order.

Example

$$f(x) = x - 1 + \frac{(0.5)^x - 0.5}{0.3}$$

$$f'(x) = 1 - 2.3105[0.5]^x$$

An approximate root (obtained graphically) is 2.

Step k	x_k	$f(x_k)$	$f'(x_k)$
0	2	0.1667	0.4224
1	1.605	–0.002	0.2655
2	1.6125	–0.0005	⋯

The Method of False Position. This variant is commenced by finding x_0 and x_1 such that $f(x_0)$ and $f(x_1)$ are of opposite signs. Then, α_1 = slope of secant line joining $[x_0, f(x_0)]$ and $[x_1, f(x_1)]$ so that

$$x_2 = x_1 - \frac{x_1 - x_0}{f(x_1) - f(x_0)} f(x_1)$$

In each following step, α_k is the slope of the line joining $[x_k, f(x_k)]$ to the most recently determined point where $f(x_j)$ has the opposite sign from that of $f(x_k)$. This method is of first order.

The Method of Wegstein

This is a variant of the method of successive substitutions which forces or accelerates convergence. The iterative procedure $x_{k+1} = F(x_k)$ is revised by setting $\hat{x}_{k+1} = F(x_k)$ and then taking $x_{k+1} = qx_k + (1 - q)\ \hat{x}_{k+1}$. Wegstein found that suitably chosen qs are related to the basic process as follows:

Behavior of Successive Substitution Process	Range of Optimum q
Oscillatory convergence	$0 < q < 1/2$
Oscillatory divergence	$1/2 < q < 1$
Monotonic convergence	$q < 0$
Monotonic divergence	$1 < q$

At each step, q may be calculated to give a locally optimum value by setting

$$q = \frac{x_{k+1} - x_k}{x_{k+1} - 2x_k + x_{k-1}}$$

The Method of Continuity

In the case of n equations in n unknowns, when n is large, determining the approximate solution may involve considerable effort. In such a case, the method of continuity is admirably suited for use on either digital or analog computers. It consists basically of the introduction of an extra variable into the n equations

$$f_i(x_1, x_2, \ldots, x_n) = 0, \qquad i = 1, \ldots, n$$

and replacing them by

$$f_i(x_1, x_2, \ldots, x_n, \lambda) = 0, \qquad i = 1, \ldots, n$$

where λ is introduced in such a way that the functions depend in a simple way upon λ and reduce to an easily solvable system for λ = 0 and to the original equations for λ = 1. A system of ordinary differential equations, with independent variable λ, is then constructed by differentiating with respect to λ. There results

$$\sum_{j=1}^{n} \frac{\partial f_i}{\partial x_j} \frac{dx_j}{d\lambda} + \frac{\partial f_i}{\partial \lambda} = 0$$

where x_1, …, x_n are considered as functions of λ. The equations are integrated, with initial conditions obtained with λ = 0, from λ = 0 to λ = 1. If the solution can be continued to λ = 1, the values of x_1, …, x_n for λ = 1 will be a solution of the original equations. If the integration becomes infinite, the parameter λ must be introduced in a different fashion. Integration of the differential equations (which are usually nonlinear in λ) may be accomplished on an analog computer or by digital means using techniques described in a later section entitled “Numerical Solution of Ordinary Differential Equations.”

Example

$$f(x, y) = 2 + x + y - x^2 + 8xy + y^3 = 0$$

$$g(x, y) = 1 + 2x + 3y + x^2 + xy - ye^x = 0$$

Introduce λ as

$$f(x, y, \lambda) = (2 + x + y) + \lambda(-x^2 + 8xy + y^3) = 0$$

$$g(x, y, \lambda) = (1 + 2x - 3y) + \lambda(x^2 + xy - ye^x) = 0$$

For λ = 1, these reduce to the original equations, but, for λ = 0, they are the linear systems

$$x + y = -2$$

$$2x - 3y = -1$$

which has the unique solution $x = -1.4$, $y = -0.6$. The differential equations in this case become

$$\frac{\partial f}{\partial x}\frac{dx}{d\lambda} + \frac{\partial f}{\partial y}\frac{dy}{d\lambda} = -\frac{\partial f}{\partial \lambda}$$

$$\frac{\partial g}{\partial x}\frac{dx}{d\lambda} + \frac{\partial g}{\partial y}\frac{dy}{d\lambda} = -\frac{\partial g}{\partial \lambda}$$

or

$$\frac{dx}{d\lambda} = \frac{\dfrac{\partial f}{\partial y}\dfrac{\partial g}{\partial \lambda} - \dfrac{\partial f}{\partial \lambda}\dfrac{\partial g}{\partial x}}{\dfrac{\partial f}{\partial x}\dfrac{\partial g}{\partial y} - \dfrac{\partial f}{\partial y}\dfrac{\partial g}{\partial x}}$$

$$\frac{dy}{d\lambda} = \frac{\dfrac{\partial f}{\partial \lambda}\dfrac{\partial g}{\partial x} - \dfrac{\partial f}{\partial x}\dfrac{\partial g}{\partial \lambda}}{\dfrac{\partial f}{\partial x}\dfrac{\partial g}{\partial y} - \dfrac{\partial f}{\partial y}\dfrac{\partial g}{\partial x}}$$

Integrating in λ, with initial values $x = -1.4$ and $y = -0.6$ at $\lambda = 0$, from $\lambda = 0$ to $\lambda = 1$ gives the solution.

Interpolation and Finite Differences

The practicing engineer constantly finds it necessary to refer to tables as sources of information. Consequently, interpolation, or that procedure of "reading between the lines of the table," is a necessary topic in numerical analysis.

Linear Interpolation

If a function $f(x)$ is approximately linear in a certain range, then the ratio $[f(x_1) - f(x_0)]/(x_1 - x_0) = f[x_0, x_1]$ is approximately independent of x_0 and x_1 in the range. The linear approximation to the function $f(x)$, $x_0 < x < x_1$, then leads to the interpolation formula

$$f(x) \approx f(x_0) + (x - x_0) f[x_0, x_1] \approx f(x_0) + \frac{x - x_0}{x_1 - x_0}\left[f(x_1) - f(x_0)\right]$$

$$\approx \frac{1}{x_1 - x_0}\left[(x_1 - x) f(x_0) - (x_0 - x) f(x_1)\right]$$

Divided Differences of Higher Order and Higher-Order Interpolation

The first-order divided difference $f[x_0, x_1]$ was defined above. Divided differences of second and higher order are defined iteratively by

$$f[x_0, x_1, x_2] = \frac{f[x_1, x_2] - f[x_0, x_1]}{x_2 - x_0}$$

$$\vdots$$

$$f[x_0, x_1, \ldots, x_k] = \frac{f[x_1, \ldots, x_k] - f[x_0, x_1, \ldots, x_{k-1}]}{x_k - x_0}$$

and a convenient form for computational purposes is

$$f[x_0, x_1, \ldots, x_k] = \sum_{j=0}^{k}{}' \frac{f(x_j)}{(x_j - x_0)(x_j - x_1)\cdots(x_j - x_k)}$$

for any $k \geq 0$, where the $'$ means the term $(x_j - x_j)$ is omitted in the denominator. For example,

$$f[x_0, x_1, x_2] = \frac{f(x_0)}{(x_0 - x_1)(x_0 - x_2)} + \frac{f(x_1)}{(x_1 - x_0)(x_1 - x_2)} + \frac{f(x_2)}{(x_2 - x_0)(x_2 - x_1)}$$

If the accuracy afforded by a linear approximation is inadequate, a generally more accurate result may be based upon the assumption that $f(x)$ may be approximated by a polynomial of degree 2 or higher over certain ranges. This assumptions leads to *Newton's fundamental interpolation formula* with divided differences:

$$f(x) \approx f(x_0) + (x - x_0) f[x_0, x_1] + (x - x_0)(x - x_1) f[x_0, x_1, x_2]$$
$$+ (x - x_0)(x - x_1) \cdots (x - x_{n-1}) f[x_0, x_1, \ldots, x_n] + E_n(x)$$

where $E_n(x)$ = error = $[1/(n + 1)!] f^{(n-1)}(\varepsilon)\pi(x)$ where $\min(x_0, \ldots, x) < \varepsilon < \max(x_0, x_1, \ldots, x_n, x)$ and $\pi(x) = (x - x_0)(x - x_1)\ldots(x - x_n)$. In order to use this most effectively, one may first form a divided-difference table. For example, for third-order interpolation, the difference table is

$$\begin{array}{c|c|cc|c} x_0 & f(x_0) & & & \\ & & f[x_0, x_1] & & \\ x_1 & f(x_1) & & f[x_0, x_1, x_2] & \\ & & f[x_1, x_2] & & f[x_0, x_1, x_2, x_3] \\ x_2 & f(x_2) & & f[x_1, x_2, x_3] & \\ & & f[x_2, x_3] & & \\ x_3 & f(x_3) & & & \end{array}$$

where each entry is given by taking the difference between diagonally adjacent entries to the left, divided by the abscissas corresponding to the ordinates intercepted by the diagonals passing through the calculated entry.

Example. Calculate by third-order interpolation the value of cosh 0.83 given cosh 0.60, cosh 0.80, cosh 0.90, and cosh 1.10.

$$\begin{array}{c|c|cc|c} x_0 = 0.60 & 1.185\ 47 & & & \\ & & 0.7598 & & \\ x_1 = 0.80 & 1.337\ 43 & & 0.6560 & \\ & & 0.9566 & & 0.1586 \\ x_2 = 0.90 & 1.433\ 09 & & 0.7353 & \\ & & 1.1772 & & \\ x_3 = 1.10 & 1.668\ 52 & & & \end{array}$$

With $n = 3$, we have

$$\cosh 0.83 \approx 1.185\ 47 + (0.23)(0.7598) + (0.23)(0.03)(0.6560)$$
$$+ (0.23)(0.03)(-0.07)(0.1586) = 1.364\ 64$$

which varies from the true value by 0.000 04.

Lagrange Interpolation Formulas

The Newton formulas are expressed in terms of divided differences. It is often useful to have interpolation formulas expressed explicitly in terms of the ordinates involved. This is accomplished by the Lagrange interpolation polynomial of degree n:

$$y(x) = \sum_{j=0}^{n} \frac{\pi(x)}{(x - x_j)\pi'(x_j)} f(x_j)$$

where

$$\pi(x) = (x - x_0)(x - x_1)\cdots(x - x_n)$$

$$\pi'(x_j) = (x_j - x_0)(x_j - x_1)\cdots(x_j - x_n)$$

where $(x_j - x_j)$ is the omitted factor. Thus,

$$f(x) = y(x) + E_n(x)$$

$$E_n(x) = \frac{1}{(n+1)!}\pi(x) f^{(n+1)}(\varepsilon)$$

Example. The interpolation polynomial of degree 3 is

$$y(x) = \frac{(x - x_1)(x - x_2)(x - x_3)}{(x_0 - x_1)(x_0 - x_2)(x_0 - x_3)} f(x_0) + \frac{(x - x_0)(x - x_2)(x - x_3)}{(x_1 - x_0)(x_1 - x_2)(x_1 - x_3)} f(x_1)$$
$$+ \frac{(x - x_0)(x - x_1)(x - x_3)}{(x_2 - x_0)(x_2 - x_1)(x_2 - x_3)} f(x_2) + \frac{(x - x_0)(x - x_1)(x - x_2)}{(x_3 - x_0)(x_3 - x_1)(x_3 - x_2)} f(x_3)$$

Thus, directly from the data

x	0	1	3	4
$f(x)$	1	1	–1	2

we have as an interpolation polynomial $y(x)$ for (x):

$$y(x) = 1 \cdot \frac{(x-1)(x-3)(x-4)}{(0-1)(0-3)(0-4)} + 1 \cdot \frac{x(x-3)(x-4)}{(1-0)(1-3)(1-4)}$$
$$-1 \cdot \frac{x(x-1)(x-4)}{(3-0)(3-1)(3-4)} + 2 \cdot \frac{(x-0)(x-1)(x-3)}{(4-0)(4-1)(4-3)}$$

Other Difference Methods (Equally Spaced Ordinates)

Backward Differences. The backward differences denoted by

$$\nabla f(x) = f(x) - f(x-h)$$

$$\nabla^2 f(x) = \nabla f(x) - \nabla f(x-h)$$

$$\cdots$$

$$\nabla f^n(x) = \nabla^{n-1} f(x) - \nabla^{n-1} f(x-h)$$

are useful for calculation near the end of tabulated data.

Central Differences. The central differences denoted by

$$\delta f(x) = f\left(x + \frac{h}{2}\right) - f\left(x - \frac{h}{2}\right)$$

$$\delta^n f(x) = \delta^{n-1} f\left(x + \frac{h}{2}\right) - \delta^{n-1} f\left(x - \frac{h}{2}\right)$$

are useful for calculating at the interior points of tabulated data.
Also to be found in the literature are Gaussian, Stirling, Bessel, Everett, Comrie differences, and so forth.

Inverse Interpolation

This is the process of finding the value of the independent variable or abscissa corresponding to a given value of the function when the latter is between two tabulated values of the abscissa. One method of accomplishing this is to use Lagrange's interpolation formula in the form

$$x = \psi(y) = \sum_{j=0}^{n} \frac{\pi(y)}{(y - y_j)\pi'(y_j)} x_j$$

where x is expressed as a function of y. Other methods revolve about methods of iteration.

Numerical Differentiation

Numerical differentiation should be avoided wherever possible, particularly when data are empirical and subject to appreciable observation errors. Errors in data can affect numerical derivatives quite strongly (i.e., differentiation is a roughening process). When such a calculation must be made, it is usually desirable first to *smooth* the data to a certain extent.

The Use of Interpolation Formulas

If the data are given over equidistant values of the independent variable x, an interpolation formula, such as the Newton formula, may be used, and the resulting formula differentiated analytically. If the independent variable is not at equidistant values, then Lagrange's formulas must be used. By differentiating three- and five-point Lagrange interpolation formulas, the following differentiation formulas result for equally spaced tabular points.

Three-point Formulas. Let x_0, x_1, and x_2 be the three points

$$f'(x_0) = \frac{1}{2h}\left[-3f(x_0) + 4f(x_1) - f(x_2)\right] + \frac{h^2}{3} f'''(\varepsilon)$$

$$f'(x_1) = \frac{1}{2h}\left[-f(x_0) + f(x_2)\right] + \frac{h^2}{6} f'''(\varepsilon)$$

$$f'(x_2) = \frac{1}{2h}\left[f(x_0) - 4f(x_1) + 3f(x_2)\right] + \frac{h^2}{3} f'''(\varepsilon)$$

where the last term is an error term and $\min_j x_j < \varepsilon < \max_j x_j$.

Five-point Formulas. Let x_0, x_1, x_2, x_3, and x_4 be the five values of the equally spaced independent variable and $f_j = f(x_j)$.

$$f'(x_0) = \frac{1}{12h}\left[-25f_0 + 48f_1 - 36f_2 + 16f_3 - 3f_4\right] + \frac{h^4}{5} f^{(v)}(\varepsilon)$$

$$f'(x_1) = \frac{1}{12h}\left[-3f_0 - 10f_1 + 18f_2 - 6f_3 + f_4\right] - \frac{h^4}{20} f^{(v)}(\varepsilon)$$

$$f'(x_2) = \frac{1}{12h}\left[f_0 - 8f_1 + 8f_3 - f_4\right] + \frac{h^4}{30} f^{(v)}(\varepsilon)$$

$$f'(x_3) = \frac{1}{12h}\left[-f_0 + 6f_1 - 18f_2 + 10f_3 + 3f_4\right] - \frac{h^4}{20} f^{(v)}(\varepsilon)$$

$$f'(x_4) = \frac{1}{12h}\left[3f_0 - 16f_1 + 36f_2 - 48f_3 + 25f_4\right] + \frac{h^4}{5} f^{(v)}(\varepsilon)$$

and the last term is again an error term.

Smoothing Techniques

These techniques involve the approximation of the tabular data by a least squares fit of the data using some known functional form, usually a polynomial. In place of approximating $f(x)$ by a single least squares polynomial of degree n over the entire range of the tabulation, it is often desirable to replace each tabulated value by the value taken on by a last squares polynomial of degree n relevant to a subrange of $2M + 1$ points centered, where possible, at the point for which the entry is to be modified. Thus, each smoothed value replaces a tabulated value. Let $f_i = f(x_i)$ be the tabular points and y_j = smoothed values. A first-degree least squares with three points would be

$$y_0 = \tfrac{1}{6}\left[5f_0 + 2f_1 - f_2\right]$$

$$y_1 = \tfrac{1}{3}\left[f_0 + f_1 + f_2\right]$$

$$y_2 = \tfrac{1}{6}\left[-f_0 + 2f_1 + 5f_2\right]$$

A first-degree least squares with five points would be

$$y_0 = \tfrac{1}{5}\left[3f_0 + 2f_1 + f_2 - f_4\right]$$

$$y_1 = \tfrac{1}{10}\left[4f_0 + 3f_1 + 2f_2 + f_3\right]$$

$$y_2 = \tfrac{1}{5}\left[f_0 + f_1 + f_2 + f_3 + f_4\right]$$

$$y_3 = \tfrac{1}{10}\left[f_0 + 2f_1 + 3f_2 + 4f_3\right]$$

$$y_4 = \tfrac{1}{5}\left[-f_0 + f_2 + 2f_3 + 3f_4\right]$$

Thus, for example, if first-degree, five-point least squares are used, the central formula is used for all values except the first two and the last two, where the off-center formulas are used. A third-degree least squares with seven points would be

$$y_0 = \tfrac{1}{42}\left[39f_0 + 8f_1 - 4f_2 - 4f_3 + f_4 + 4f_5 - 2f_6\right]$$

$$y_1 = \tfrac{1}{42}\left[8f_0 + 19f_1 + 16f_2 + 6f_3 - 4f_4 - 7f_5 + 4f_6\right]$$

$$y_2 = \tfrac{1}{42}\left[-4f_0 + 16f_1 + 19f_2 + 12f_3 + 2f_4 - 4f_5 + f_6\right]$$

$$y_3 = \tfrac{1}{21}\left[-2f_0 + 3f_1 + 6f_2 + 7f_3 + 6f_4 + 3f_5 - 2f_6\right]$$

$$y_4 = \tfrac{1}{42}\left[f_0 - 4f_1 + 2f_2 + 12f_3 + 19f_4 + 16f_5 - 4f_6\right]$$

$$y_5 = \tfrac{1}{42}\left[4f_0 - 7f_1 - 4f_2 + 6f_3 + 16f_4 + 19f_5 + 8f_6\right]$$

$$y_6 = \tfrac{1}{42}\left[-2f_0 + 4f_1 + f_2 - 4f_3 - 4f_4 + 8f_5 + 39f_6\right]$$

Additional smoothing formulas may be found in the references. After the data are smoothed, any of the interpolation polynomials, or an appropriate least squares polynomial, may be fitted and the results used to obtain the derivative.

Least Squares Method

Parabolic. For five evenly spaced neighboring abscissas labeled x_{-2}, x_{-1}, x_0, x_1, and x_2, and their ordinates f_{-2}, f_{-1}, f_0, f_1, and f_2, assume a parabola is fit by least squares. There results for all interior points, except the first and last two points of the data, the formula for the numerical derivative:

$$f_0' = \frac{1}{10h}\left[-2f_{-2} - f_{-1} + f_1 + 2f_2\right]$$

For the first two data points designated by 0 and h:

$$f'(0) = \frac{1}{20h}\left[-21f(0) + 13f(h) + 17f(2h) - 9f(3h)\right]$$

$$f'(h) = \frac{1}{20h}\left[-11f(0) + 3f(h) + 7f(2h) + f(3h)\right]$$

and for the last two given by $\alpha - h$ and α:

$$f'(\alpha - h) = \frac{1}{20h}\left[-11f(\alpha) + 3f(\alpha - h) + 7f(\alpha - 2h) + f(\alpha - 3h)\right]$$

$$f'(\alpha) = \frac{1}{20h}\left[-21f(\alpha) + 13f(\alpha - h) + 17f(\alpha - 2h) - 9(\alpha - 3h)\right]$$

Quartic (Douglas-Avakian). A fourth-degree polynomial $y = a + bx + cx^2 + dx^3 + ex^4$ is fitted to seven adjacent equidistant points (spacing h) after a translation of coordinates has been made so that $x = 0$ corresponds to the central point of the seven. Thus, these may be called $-3h$, $-2h$, $-h$, 0, h, $2h$, and $3h$. Let k = coefficient h for the seven points. This is, in $-3h$, $k = -3$. Then, the coefficients for the polynomial are

$$a = \frac{524\sum f(kh) - 245\sum k^2 f(kh) + 21\sum k^4 f(kh)}{924}$$

$$b = \frac{397\sum kf(kh)}{1512h} - \frac{7\sum k^3 f(kh)}{216h}$$

$$c = \frac{-840\sum f(kh) + 679\sum k^2 f(kh) - 67\sum k^4 f(kh)}{3168h^2}$$

$$d = \frac{-7\sum kf(kh) + \sum k^3 f(kh)}{216h^3}$$

$$e = \frac{72\sum f(kh) - 67\sum k^2 f(kh) + 7\sum k^4 f(kh)}{3168h^4}$$

where all summations run from $k = -3$ to $k = +3$ and $f(kh)$ = tabular value at kh. The slope of the polynomial at $x = 0$ is $dy/dx = b$.

Numerical Integration

Numerical evaluation of the finite integral $\int_a^b f(x)\,dx$ is carried out by a variety of methods. A few are given here.

Newton-Cotes Formulas (Equally Spaced Ordinates)

Trapezoidal Rule. This formula consists of subdividing the interval $a \le x \le b$ into n subintervals a to $a + h$, $a + h$ to $a + 2h$, …, and replacing the graph of $f(x)$ by the result of joining the ends of adjacent ordinates by line segments. If $f_j = f(x_j) = f(a + jh)$, $f_0 = f(a)$, and $f_n = f(b)$, the integration formula is

$$\int_a^b f(x)\,dx = \frac{h}{2}\left[f_0 + 2f_1 + 2f_2 + \cdots + 2f_{n-1} + f_n\right] + E_n$$

where $|E_n| = (nh^3/12)|f''(\varepsilon)| = [(b-a)^3/12n^2]|f''(\varepsilon)|$, $a < \varepsilon < b$. This procedure is not of high accuracy. However, if $f''(x)$ is continuous in $a < x < b$, the error goes to zero as $1/n^2$, $n \to \infty$.

Parabolic Rule (Simpson's Rule). This procedure consists of subdividing the interval $a < x < b$ into $n/2$ subintervals, each of length $2h$, where n is an even integer. Using the notation as above the integration formula is

$$\int_a^b f(x)\,dx = \frac{h}{3}\left[f_0 + 4f_1 + 2f_2 + 4f_3 + \cdots + 4f_{n-3} + 2f_{n-2} + 4f_{n-1} + f_n\right] + E_n$$

where

$$|E_n| = \frac{nh^5}{180}\left|f^{(iv)}(\varepsilon)\right| = \frac{(b-a)^3}{180n^4}\left|f^{(iv)}(\varepsilon)\right| \qquad a < \varepsilon < b$$

This method approximates $f(x)$ by a parabola on each subinterval. This rule is generally more accurate than the trapezoidal rule. It is the most widely used integration formula.

Weddle's Rule. This procedure consists of subdividing the integral $a < x < b$ into $n/6$ subintervals, each of length $6h$, where n is a multiple of 6. Using the notation from the trapezoidal rule, there results

$$\int_a^b f(x)\,dx = \frac{3h}{10}\left[f_0 + 5f_1 + f_2 + 6f_3 + f_4 + 5f_5 + 2f_6 + 5f_7 + f_8 + \cdots + 6f_{n-3} + f_{n-2} + 5f_{n-1} + f_n\right] + E_n$$

Note that the coefficients of f_j follow the rule 1, 5, 1, 6, 1, 5, 2, 5, 1, 6, 1, 5, 2, 5, etc.… This procedure consists of approximately $f(x)$ by a polynomial of degree 6 on each subinterval. Here,

$$E_n = \frac{nh^7}{1400}\left[10f^{(6)}(\varepsilon_1) + 9h^2 f^{(8)}(\varepsilon_2)\right]$$

Gaussian Integration Formulas (Unequally Spaced Abscissas)

These formulas are capable of yielding comparable accuracy with fewer ordinates than the equally spaced formulas. The ordinates are obtained by optimizing the distribution of the abscissas rather than by arbitrary choice. For the details of these formulas, Hildebrand (1956) is an excellent reference.

Two-Dimensional Formula

Formulas for two-way integration over a rectangle, circle, ellipse, and so forth, may be developed by a double application of one-dimensional integration formulas. The two-dimensional generalization of the parabolic rule is given here. Consider the iterated integral $\int_a^b \int_c^d f(x, y)\,dx\,dy$. Subdivide $c < x < d$ into m (even) subintervals of length $h = (d - c)/m$, and $a < y < b$ into n (even) subintervals of length $k = (b - a)/n$. This gives a subdivision of the rectangle $a \le y \le b$ and $c \le x \le d$ into subrectangles. Let $x_j = c + jh$, $y_j = a + jk$, and $f_{i,j} = f(x_i, y_j)$. Then,

$$\int_a^b \int_c^d f(x,y)\,dx\,dy = \frac{hk}{9}\left[\left(f_{0,0} + 4f_{1,0} + 2f_{2,0} + \cdots + f_{m,0}\right) + 4\left(f_{0,1} + 4f_{1,1} + 2f_{2,1} + \cdots + f_{m,1}\right)\right.$$
$$\left. + 2\left(f_{0,2} + 4f_{1,2} + 2f_{2,2} + \cdots + f_{m,2}\right) + \cdots + \left(f_{0,n} + 4f_{1,n} + 2f_{2,n} + \cdots + f_{m,n}\right)\right] + E_{m,n}$$

where

$$E_{m,n} = -\frac{hk}{90}\left[mh^4 \frac{\partial^4 f(\varepsilon_1, \eta_1)}{\partial x^4} + nk^4 \frac{\partial^4 f(\varepsilon_2, \eta_2)}{\partial y^4}\right]$$

where ε_1 and ε_2 lie in $c < x < d$, and η_1 and η_2 lie in $a < y < b$.

Numerical Solution of Ordinary Differential Equations

A number of methods have been devised to solve ordinary differential equations numerically. The general references contain some information. A numerical solution of a differential equation means a table of values of the function y and its derivatives over only a limited part of the range of the independent variable. Every differential equation of order n can be rewritten as n first-order differential equations. Therefore, the methods given below will be for first-order equations, and the generalization to simultaneous systems will be developed later.

The Modified Euler Method

This method is simple and yields modest accuracy. If extreme accuracy is desired, a more sophisticated method should be selected. Let the first-order differential equation be $dy/dx = f(x, y)$ with the initial condition (x_0, y_0) (i.e., $y = y_0$ when $x = x_0$). The procedure is as follows.

Step 1. From the given initial conditions (x_0, y_0) compute $y_0' = f(x_0, y_0)$ and $y_0'' = [\partial f(x_0, y_0)/\partial x] + [\partial f(x_0, y_0)/\partial y]\, y_0'$. Then, determine $y_1 = y_0 + h\, y_0' + (h^2/2)\, y_0''$, where h = subdivision of the independent variable.

Step 2. Determine $y_1' = f(x_1, y_1)$ where $x_1 = x_0 + h$. These prepare us for the following.

Predictor Steps.

Step 3. For $n \geq 1$, calculate $(y_{n+1})_1 = y_{n-1} + 2h\, y_n'$.

Step 4. Calculate $(y_{n+1}')_1 = f[x_{n+1}, (y_{n+1})_1]$.

Corrector Steps.

Step 5. Calculate $(y_{n+1})_2 = y_n + (h/2)\,[(y_{n+1}')_1 + y_n']$, where y_n and y_n' without the subscripts are the previous values obtained by this process (or by steps 1 and 2).

Step 6. $(y_{n+1}')_2 = f[x_{n+1}, (y_{n+1})_2]$.

Step 7. Repeat the corrector steps 5 and 6 if necessary until the desired accuracy is produced in y_{n+1}, y_{n+1}'.

Example. Consider the equation $y' = 2y^2 + x$ with the initial conditions $y_0 = 1$ when $x_0 = 0$. Let $h = 0.1$. A few steps of the computation are illustrated.

Step	
1	$y_0' = 2y_0^2 + x_0 = 2$
	$y_0'' = 1 + 4y_0y_0' = 1 + 8 = 9$
	$y_1 = 1 + (0.1)(2) + \left[(0.1)^2/2\right]9 = 1.245$
2	$y_1' = 2y_1^2 + x_1 = 3.100 + 0.1 = 3.200$
3	$(y_2)_1 = y_0 + 2hy_1' = 1 + 2(0.1)3.200 = 1.640$
4	$(y_2')_1 = 2(y_2)_1^2 + x_2 = 5.592$
5	$(y_2)_2 = y_1 + (0.1/2)\left[(y_2')_1 + y_1'\right] = 1.685$
6	$(y_2')_2 = 2(y_2)_2^2 + x_2 = 5.878$
5 (repeat)	$(y_2)_3 = y_1 + (0.05)\left[(y_2')_2 + y_1'\right] = 1.699$
6 (repeat)	$(y_2')_3 = 2(y_2)_3^2 + x_2 = 5.974$

and so forth. This procedure. may be programmed for a computer. A discussion of the truncation error of this process may be found in Milne (1953).

Modified Adam's Method

The procedure given here was developed retaining third differences. It can then be considered as a more exact predictor-corrector method than the Euler method. The procedure is as follows for $dy/dx = f(x, y)$ and h = interval size.

Steps 1 and 2 are the same as in Euler method.

Predictor Steps.

Step 3. $(y_{n+1})_1 = y_n + (h/24)\,[55y_n' - 59y_{n-1}' + 37y_{n-2}' - 9y_{n-3}']$, where y_n', y_{n-1}', etc…, are calculated in step 1.

Step 4. $(y_{n+1}')_1 = f[x_{n+1}, (y_{n+1})_1]$.

Corrector Steps.

Step 5. $(y_{n+1})_2 = y_n + (h/24)\,[9(y'_{n+1})_1 + 19y'_n - 5y'_{n-1} + y'_{n-2}]$.
Step 6. $(y'_{n+1})_2 = f[x_{n+1}, (y_{n+1})_2]$.
Step 7. Iterate steps 5 and 6 if necessary.

Runge-Kutta Methods

These methods are self-starting and are inherently stable. Kopal (1955) is a good reference for their derivation and discussion. Third- and fourth-order procedures are given below for $dy/dx = f(x, y)$ and h = interval size.

For third-order (error $\approx h^4$).

$$k_0 = hf\left(x_n, y_n\right)$$

$$k_1 = hf\left(x_n + \tfrac{1}{2}h, y_n + \tfrac{1}{2}k_0\right)$$

$$k_2 = hf\left(x_n + h, y_n + 2k_1 - k_0\right)$$

and

$$y_{n+1} = y_n + \tfrac{1}{6}\left(k_0 + 4k_1 + k_2\right)$$

for all $n \geq 0$, with initial condition (x_0, y_0).

For fourth-order (error $\approx h^5$),

$$k_0 = hf\left(x_n, y_n\right)$$

$$k_1 = hf\left(x_n + \tfrac{1}{2}h, y_n + \tfrac{1}{2}k_0\right)$$

$$k_2 = hf\left(x_n + \tfrac{1}{2}h, y_n + \tfrac{1}{2}k_1\right)$$

$$k_3 = hf\left(x_n + h, y_n + k_2\right)$$

and

$$y_{n+1} = y_n + \tfrac{1}{6}\left(k_0 + 2k_1 + 2k_2 + k_3\right)$$

Example. (Third-order) Let $dy/dx = x - 2y$, with initial condition $y_0 = 1$ when $x_0 = 0$, and let $h = 0.1$. Clearly, $x_n = nh$. To calculate y_1, proceed as follows:

$$k_0 = 0.1\left[x_0 - 2y_0\right] = -0.2$$

$$k_1 = 0.1\left[0.05 - 2(1 - 0.1)\right] = -0.175$$

$$k_2 = 0.1\left[0.1 - 2(1 - 0.35 + 0.2)\right] = -0.16$$

$$y_1 = 1 + \tfrac{1}{6}(-0.2 - 0.7 - 0.16) = 0.8234$$

Equations of Higher Order and Simultaneous Differential Equations

Any differential equation of second- or higher order can be reduced to a simultaneous system of first-order equations by the introduction of auxiliary variables. Consider the following equations:

$$\frac{d^2x}{dt^2} + xy\frac{dx}{dt} + z = e^x$$

$$\frac{d^2y}{dt^2} + xy\frac{dy}{dt} = 7 + t^2$$

$$\frac{d^2z}{dt^2} + xz\frac{dz}{dt} + x = e^x$$

In the new variables $x_1 = x$, $x_2 = y$, $x_3 = z$, $x_4 = dx_1/dt$, $x_5 = dx_2/dt$, and $x_6 = dx_3/dt$, the equations become

$$\frac{dx_1}{dt} = x_4$$

$$\frac{dx_2}{dt} = x_5$$

$$\frac{dx_3}{dt} = x_6$$

$$\frac{dx_4}{dt} = -x_1x_2x_4 - x_3 + e^{x_3}$$

$$\frac{dx_5}{dt} = -x_3x_2x_5 + 7 + t^2$$

$$\frac{dx_6}{dt} = -x_1x_3x_6 - x_1 + e^{x_1}$$

which is a system of the general form

$$\frac{dx_i}{dt} = f_i\left(t, x_1, x_2, x_3, \ldots, x_n\right)$$

where $i = 1, 2, \ldots, n$. Such systems may be solved by simultaneous application of any of the above numerical techniques. A Runge-Kutta method for

$$\frac{dx}{dt} = f(t, x, y)$$

$$\frac{dy}{dt} = g(t, x, y)$$

is given below. The fourth-order procedure is shown.

Starting at the initial conditions x_0, y_0, and t_0, the next values x_1 and y_1 are computed via the equations below (where $\Delta t = h$, $t_j = h + t_{j-1}$):

$$k_0 = hf(t_0, x_0, y_0) \qquad l_0 = hg(t_0, x_0, y_0)$$

$$k_1 = hf\left(t_0 + \frac{h}{2}, x_0 + \frac{k_0}{2}, y_0 + \frac{l_0}{2}\right) \qquad l_1 = hg\left(t_0 + \frac{h}{2}, x_0 + \frac{k_0}{2}, y_0 + \frac{l_0}{2}\right)$$

$$k_2 = hf\left(t_0 + \frac{h}{2}, x_0 + \frac{k_1}{2}, y_0 + \frac{l_1}{2}\right) \qquad l_2 = hg\left(t_0 + \frac{h}{2}, x_0 + \frac{k_1}{2}, y_0 + \frac{l_1}{2}\right)$$

$$k_3 = hf(t_0 + h, x_0 + k_2, y_0 + l_2) \qquad l_3 = hg(t_0 + h, x_0 + k_2, y_0 + l_2)$$

and

$$x_1 = x_0 + \tfrac{1}{6}(k_0 + 2k_1 + 2k_2 + k_3)$$

$$y_1 = y_0 + \tfrac{1}{6}(l_0 + 2l_1 + 2l_2 + l_3)$$

To continue the computation, replace t_0, x_0, and y_0 in the above formulas by $t_1 = t_0 + h$, x_1, and y_1 just calculated. Extension of this method to more than two equations follows precisely this same pattern.

Numerical Solution of Integral Equations

This section considers a method of numerically solving the Fredholm integral equation of the second kind:

$$u(x) = f(x) + \lambda \int_a^b k(x,t)u(t)\,dt \qquad \text{for } u(x)$$

The method discussed arises because a definite integral can be closely approximated by any of several numerical integration formulas (each of which arises by approximating the function by some polynomial over an interval). Thus, the definite integral can be replaced by an integration formula which becomes

$$u(x) = f(x) + \lambda(b-a)\left[\sum_{i=1}^{n} c_i k(x, t_i)u(t_i)\right]$$

where $t_1, \ldots, t_n$ are points of subdivision of the t axis, $a \le t \le b$, and the cs are coefficients whose values depend upon the type of numerical integration formula used. Now, this must hold for all values of x, where $a \le x \le b$; so it must hold for $x = t_1$, $x = t_2, \ldots$, $x = t_n$. Substituting for x successively t_1, $t_2, \ldots$, t_n, and setting $u(t_i) = u_i$ and $f(t_i) = f_i$, we get n linear algebraic equations for the n unknowns $u_1, \ldots, u_n$. That is,

$$u_i = f_i + (b-a)\left[c_1 k(t_i, t_1)u_1 + c_2 k(t_i, t_2)u_2 + \cdots + c_n k(t_i, t_n)u_n\right], \qquad i = 1, 2, \ldots, n$$

These u_j may be solved for by the methods under the section entitled "Numerical Solution of Linear Equations."

Numerical Methods for Partial Differential Equations

The ultimate goal of numerical (discrete) methods for partial differential equations (PDEs) is the reduction of continuous systems (projections) to discrete systems that are suitable for high-speed computer solutions. The user must be cautioned that the seeming elementary nature of the techniques holds pitfalls that can be seriously misleading. These approximations often lead to difficult mathematical questions of adequacy, accuracy, convergence, stability, and consistency. Convergence is concerned with

the approach of the approximate numerical solution to the exact solution as the number of mesh units increase indefinitely in some sense. Unless the numerical method can be shown to converge to the exact solution, the chosen method is unsatisfactory.

Stability deals in general with error growth in the calculation. As stated before, any numerical method involves truncation and round-off errors. These errors are not serious unless they grow as the computation proceeds (i.e., the method is unstable).

Finite Difference Methods

In these methods, the derivatives are replaced by various finite differences. The methods will be illustrated for problems in two space dimensions (x, y) or (x, t) where t is timelike. Using subdivisions $\Delta x = h$ and $\Delta y = k$ with $u(i\,h, jk) = u_{i,j}$, approximate $u_x|_{i,j} = [(u_{i+1,j} - u_{i,j})/h] + O(h)$ (forward difference), a first-order $[O(h)]$ method, or $u_x|_{i,j} = [(u_{i+1,j} - u_{i-1,j})/2h] + O(h^2)$ (central difference), a second-order method. The second derivative is usually approximated with the second-order method $[u_{xx}|_{i,j} = [(u_{i+1,j} - 2u_{i,j} + u_{i-1,j})/h^2] + O(h^2)]$.

Example. Using second-order differences for u_{xx} and u_{yy}, the five-point difference equation (with $h = k$) for Laplace's equation $u_{xx} + u_{yy} = 0$ is $u_{i,j} = 1/4[u_{i+1,j} + u_{i-1,j} + u_{i,j+1} + u_{i,j-1}]$. The accuracy is $O(h^2)$. This model is called *implicit* because one must solve for the total number of unknowns at the unknown grid points (i, j) in terms of the given boundary data. In this case, the system of equations is a linear system.

Example. Using a forward-difference approximation for u_t and a second-order approximation for u_{xx}, the diffusion equation $u_t = u_{xx}$ is approximated by the *explicit* formula $u_{i,j+1} = ru_{i-1,j} + (1 - 2r)u_{i,j} + ru_{i+1,j}$. This classic result permits step-by-step advancement in the t direction beginning with the initial data at $t = 0$ $(j = 0)$ and guided by the boundary data. Here, the term $r = \Delta t/(\Delta x)^2 = k/h^2$ is restricted to be less than or equal to 1/2 for stability and the truncation error is $O(k^2 + kh^2)$.

The Crank-Nicolson implicit formula which approximates the diffusion equation $u_t = u_{xx}$ is

$$-r\lambda u_{i-1,j+1} + (1 + 2r\lambda)u_{i,j+1} - r\lambda u_{i+1,j+1} = r(1-\lambda)u_{i-1,j} + [1 - 2r(1-\lambda)]u_{i,j} + r(1-\lambda)u_{i+1,j}$$

The stability of this numerical method was analyzed by Crandall (Ames, 1993) where the λ, r stability diagram is given.

Approximation of the time derivative in $u_t = u_{xx}$ by a central difference leads to an always unstable approximation — the useless approximation

$$u_{i,j+1} = u_{i,j-1} + 2r(u_{i+1,j} - 2u_{i,j} + u_{i-1,j})$$

which is a warning to be careful.

The foregoing method is *symmetric* with respect to the point (i, j), where the method is centered. Asymmetric methods have some computational advantages, so the Saul'yev method is described (Ames, 1993). The algorithms $(r = k/h^2)$

$$(1 + r)u_{i,j+1} = u_{i,j} + r(u_{i-1,j+1} - u_{i,j} + u_{i+1,j}) \qquad \text{(Saul' yev A)}$$

$$(1 + r)u_{i,j+1} = u_{i,j} + r(u_{i+1,j+1} - u_{i,j} + u_{i-1,j}) \qquad \text{(Saul' yev B)}$$

are used as in any one of the following options:

1. Use Saul'yev A only and proceed line-by-line in the $t(j)$ direction, but *always* from the left boundary on a line.
2. Use Saul'yev B only and proceed line-by-line in the $t(j)$ direction, but *always* from the right boundary to the left on a line.

3. Alternate from line to line by first using Saul'yev A and then B, or the reverse. This is related to *alternating direction methods.*
4. Use Saul'yev A and Saul'yev B on the same line and average the results for the final answer (A first, and then B). This is equivalent to introducing the dummy variables $P_{i,j}$ and $Q_{i,j}$ such that

$$(1+r)P_{i,j+1} = U_{i,j} + r\left(P_{i-1,j+1} - U_{i,j} + U_{i+1,j}\right)$$

$$(1+r)Q_{i,j+1} = U_{i,j} + r\left(Q_{i+1,j+1} - U_{i,j} + U_{i-1,j}\right)$$

and

$$U_{i,j+1} = \tfrac{1}{2}\left(P_{i,j+1} + Q_{i,j+1}\right)$$

This averaging method has some computational advantage because of the possibility of truncation error cancellation. As an alternative, one can retain the $P_{i,j}$ and $Q_{i,j}$ from the previous step and replace $U_{i,j}$ and $U_{i+1,j}$ by $P_{i,j}$ and $P_{i+1,j}$, respectively, and $U_{i,j}$ and $U_{i-1,j}$ by $Q_{i,j}$ and $Q_{i-1,j}$, respectively.

Weighted Residual Methods (WRMs)

To set the stage for the method of finite elements, we briefly describe the WRMs, which have several variations — the interior, boundary, and mixed methods. Suppose the equation is $Lu = f$, where L is the partial differential operator and f is a known function, of say x and y. The first step in WRM is to select a class of known basis functions b_i (e.g., trigonometric, Bessel, Legendre) to approximate $u(x, y)$ as $\sim \Sigma\, a_i\, b_i\, (x, y) = U(x, y, a)$. Often, the b_i are selected so that $U(x, y, a)$ satisfy the boundary conditions. This is essentially the *interior method*. If the b_i, in $U(x, y, a)$ are selected to satisfy the differential equations, but not the boundary conditions, the variant is called the *boundary method*. When neither the equation nor the boundary conditions are satisfied, the method is said to be *mixed*. The least ingenuity is required here. The usual method of choice is the interior method.

The second step is to select an optimal set of constants a_i, $i = 1,2, \ldots, n$, by using the residual $R_I(U) = LU - f$. This is done here for the interior method. In the boundary method, there are a set of boundary residual R_B, and, in the mixed method. Both R_I and R_B. Using the spatial average $(w, v) = \int_v wvdV$, the criterion for selecting the values of a_i is the requirement that the n spatial averages

$$\left(b_i, R_E(U)\right) = 0, \qquad i = 1,2,\ldots,n$$

These represent n equations (linear if the operator L is linear and nonlinear otherwise) for the a_i.

Particular WRMs differ because of the choice of the b_js. The most common follow.

1. *Subdomain* The domain V is divided into n smaller, not necessarily disjoint, subdomains V_j with $w_j(x, y) = 1$ if (x, y) is in V_j, and 0 if (x, y) is not in V_j.
2. *Collocation* Select n points $P_j = (x_j, y_j)$ in V with $w_j(P_j) = \delta(P - P_j)$, where $\int_v \phi(P)\delta(P - P_j)dP = \phi(P_j)$ for all test functions $\phi(P)$ which vanish outside the compact set V. Thus, $(w_j, R_E) = \int_v \delta(P - P_j)R_E dV = R_E[U(P_j) \equiv 0$ (i.e., the residual is set equal to zero at the n points P_j).
3. *Least squares* Here, the functional $I(a) = \int_V R_E^2\, dV$, where $a = (a_1, \ldots, a_n)$, is to be made stationary with respect to the a_j. Thus, $0 = \partial I/\partial a_j = 2\int_v R_E(\partial R_E/\partial a_j)dV$, with $j = 1, 2, \ldots, n$. The w_j in this case are $\partial R_E/\partial a_j$.
4. *Bubnov-Galerkin* Choose $w_j(P) = b_j(P)$. This is perhaps the best-known method.
5. *Stationary Functional (Variational) Method* With ϕ a variational integral (or other functional), set $\partial\phi[U]/\partial a_j = 0$, where $j = 1, \ldots, n$, to generate the n algebraic equations.

Example. $u_{xx} + u_{yy} = -2$, with $u = 0$ on the boundaries of the square $x = \pm 1$, $y = \pm 1$. Select an interior method with $U = a_1(1 - x^2)(1 - y^2) + a_2x^2y^2(1 - x^2)(1 - y^2)$, whereupon the residual $R_E(U) = 2a_1(2 - x^2 - y^2) + 2a_2[(1 - 6x^2)y^2(1 - y^2) + (1 - 6y^2) - (1 - x^2)] + 2$. Collocating at (1/3, 1/3) and (2/3, 2/3) gives the two linear equations $-32a_1/9 + 32a_2/243x^2 + 2 = 0$ and $-20a_1/9 - 400a_2/243 + 2 = 0$ for a_1 and a_2.

WRM methods can obviously be used as approximate methods. We have now set the stage for *finite elements.*

Finite Elements

The WRM methods are more general than the *finite elements* (FE) methods. FE methods require, in addition, that the basis functions be finite elements (i.e., functions that are zero except on a small part of the domain under consideration). A typical example of an often used basis is that of triangular elements. For a triangular element with Cartesian coordinates (x_1, y_1), (x_2, y_2), and (x_3, y_3), define natural coordinates L_1, L_2, and L_3 (L_i, $\leftrightarrow (x_i, y_i)$) so that $L_i = A_i/A$ where

$$A = \tfrac{1}{2}\det\begin{bmatrix} 1 & x_1 & y_1 \\ 1 & x_2 & y_2 \\ 1 & x_3 & y_3 \end{bmatrix}$$

is the area of the triangle and

$$A_1 = \tfrac{1}{2}\det\begin{bmatrix} 1 & x & y \\ 1 & x_2 & y_2 \\ 1 & x_3 & y_3 \end{bmatrix}$$

$$A_2 = \tfrac{1}{2}\det\begin{bmatrix} 1 & x_1 & y_1 \\ 1 & x & y \\ 1 & x_3 & y_3 \end{bmatrix}$$

$$A_3 = \tfrac{1}{2}\det\begin{bmatrix} 1 & x_1 & y_1 \\ 1 & x_2 & y_2 \\ 1 & x & y \end{bmatrix}$$

Clearly $L_1 + L_2 + L_3 = 1$, and the L_i are one at node i and zero at the other nodes. In terms of the Cartesian coordinates,

$$\begin{bmatrix} L_1 \\ L_2 \\ L_3 \end{bmatrix} = \frac{1}{2A}\begin{bmatrix} x_2y_3 - x_3y_2, & y_2 - y_3, & x_3 - x_2 \\ x_3y_1 - x_1y_3, & y_3 - y_1, & x_1 - x_3 \\ x_1y_2 - x_2y_1, & y_1 - y_2, & x_2 - x_1 \end{bmatrix}\begin{bmatrix} 1 \\ x \\ y \end{bmatrix}$$

is the linear triangular element relation.

Tables of linear, quadratic, and cubic basis functions are given in the literature. Notice that while the linear basis needs three nodes, the quadratic requires six and the cubic basis ten. Various modifications, such as the Hermite basis, are described in the literature. Triangular elements are useful in approximating irregular domains.

For rectangular elements, the *chapeau* functions are often used. Let us illustrate with an example. Let $u_{xx} + u_{yy} = Q$, $0 < x < 2$, $0 < y < 2$, $u(x, 2) = 1$, $u(0, y) = 1$, $u_y(x, 0) = 0$, $u_x(2, y) = 0$, and $Q(x, y) = Qw\delta(x - 1)\delta(y - 1)$,

$$\delta(x) = \begin{cases} 0 & x \neq 0 \\ 1 & x = 0 \end{cases}$$

Using four equal rectangular elements, map the element I with vertices at (0, 0) (0, 1), (1, 1), and (1, 0) into the local (canonical) coordinates (ξ, η), $-1 \le \xi \le 1$, $-1 \le \eta \le 1$, by means of $x = 1/2(\xi + 1)$, $y = 1/2(\eta + 1)$. This mapping permits one to develop software that standardizes the treatment of all elements. Converting to (ξ, η) coordinates, our problem becomes $u_{\xi\xi} + u_{\eta\eta} = 1/4Q$, $-1 \le \xi \le 1$, $-1 \le \eta \le 1$, $Q = Qw\delta(\xi - 1)\delta(\eta - 1)$.

First, a trial function $\bar{u}(\xi, \eta)$ is defined as $u(\xi, \eta) \approx \bar{\mu}(\xi, \eta) = \Sigma_{j=1}^{4} A_j \phi_j(\xi, \eta)$ (in element I) where the ϕ_j are the two-dimensional chapeau functions

$$\phi_1 = \left[\tfrac{1}{2}(1-\xi)\tfrac{1}{2}(1-\eta)\right] \qquad \phi_2 = \left[\tfrac{1}{2}(1+\xi)\tfrac{1}{2}(1-\eta)\right]$$

$$\phi_3 = \left[\tfrac{1}{2}(1+\xi)\tfrac{1}{2}(1+\eta)\right] \qquad \phi_4 = \left[\tfrac{1}{2}(1-\xi)\tfrac{1}{2}(1+\eta)\right]$$

Clearly ϕ_i take the value one at node i, provide a bilinear approximation, and are nonzero only over elements adjacent to node i.

Second, the equation residual $R_E = \nabla^2 \bar{u} - 1/4Q$ is formed and a WRM procedure is selected to formulate the algebraic equations for the A_i. This is indicated using the Galerkin method. Thus, for element I, we have

$$\iint_{D_I} \left(\bar{u}_{\xi\xi} + \bar{u}_{\eta\eta} - Q\right)\phi_i(\xi,\eta)\, d\xi\, d\eta = 0, \qquad i = 1,\ldots,4$$

Applying Green's theorem, this result becomes

$$\iint_{D_I} \left[\bar{u}_\xi \left(\phi_i\right)_\xi + \bar{u}_\eta \left(\phi_i\right)_\eta + \tfrac{1}{4} Q \phi_i\right] d\xi\, d\eta - \int_{\partial D_I} \left(\bar{u}_\xi c_\xi + \bar{u}_\eta c_\eta\right) \phi_i\, ds = 0, \qquad i = 1,2,\ldots,4$$

Using the same procedure in all four elements and recalling the property that the ϕ_i in each element are nonzero only over elements adjacent to node i gives the following nine equations:

$$\sum_{e=1}^{4} \left\{ \iint_{D_e} \sum_{j=1}^{9} A_j \left[\left(\phi_j\right)_\xi \left(\phi_i\right)_\xi + \left(\phi_j\right)_\eta \left(\phi_i\right)_\eta\right] + \tfrac{1}{4} Q \phi_i \right\} d\xi\, d\eta$$

$$- \sum_{e=1}^{4} \int_{\partial D_e} \left(\bar{u}_\xi c_\xi + \bar{u}_\eta c_\eta\right) \phi\, ds = 0, \qquad n = 1,2,\ldots,9$$

where the c_ξ and c_η are the direction cosines of the appropriate element (e) boundary.

Method of Lines

The *method of lines*, when used on PDEs in two dimensions, reduces the PDE to a system of ordinary differential equations (ODEs), usually by finite difference or finite element techniques. If the original problem is an initial value (boundary value) problem, then the resulting ODEs form an initial value (boundary value) problem. These ODEs are solved by ODE numerical methods.

Example. $u_t = u_{xx} + u^2$, $0 < x < 1$, $0 < t$, with the initial value $u(x, 0) = x$, and boundary data $u(0, t) = 0$, $u(1, t) = \sin t$. A discretization of the space variable (x) is introduced and the time variable is left continuous. The approximation is $\dot{u}_i = (u_{i+1} - 2u_i + u_{i-1})/h^2 + u_i^2$. With $h = 1/5$, the equations become

$$u_0(t) = 0$$

$$\dot{u}_1 = \tfrac{1}{25}\left[u_2 - 2u_1\right] + u_1^2$$

$$\dot{u}_2 = \tfrac{1}{25}\left[u_3 - 2u_2 + u_1\right] + u_2^2$$

$$\dot{u}_3 = \tfrac{1}{25}\left[u_4 - 2u_3 + u_2\right] + u_3^2$$

$$\dot{u}_4 = \tfrac{1}{25}\left[\sin t - 2u_4 + u_3\right] + u_4^2$$

$$u_5 = \sin t$$

and $u_1(0) = 0.2$, $u_2(0) = 0.4$, $u_3(0) = 0.6$, and $u_4(0) = 0.8$.

Discrete and Fast Fourier Transforms

Let $x(n)$ be a sequence that is nonzero only for a finite number of samples in the interval $0 \le n \le N - 1$. The quantity

$$X(k) = \sum_{n=0}^{N-1} x(n) e^{-i(2\pi/N)nk}, \quad k = 0,1,\ldots,N-1$$

is called the *discrete Fourier transform* (DFT) of the sequence $x(n)$. Its inverse (IDFT) is given by

$$x(n) = \frac{1}{N}\sum_{k=0}^{N-1} X(k) e^{i(2\pi/N)nk}, \quad n = 0,1,\ldots,N-1 \quad \left(i^2 = -1\right)$$

Clearly, DFT and IDFT are finite sums and there are N frequency values. Also, $X(k)$ is periodic in k with period N.

Example. $x(0) = 1$, $x(1) = 2$, $x(2) = 3$, $x(3) = 4$

$$X(k) = \sum_{n=0}^{3} x(n) e^{-i(2\pi/4)nk}, \quad k = 0,\ 1,\ 2,\ 3,\ 4$$

Thus,

$$X(0) = \sum_{n=0}^{3} x(n) = 10$$

and $X(1) = x(0) + x(1)e^{-i\pi/2} + x(2)e^{-i\pi} + x(3)e^{-i3\pi/2} = 1 - 2i - 3 + 4i = -2 + 2i$; $X(2) = -2$; $X(3) = -2 - 2i$.

DFT Properties

1. Linearity: If $x_3(n) = ax_1(n) + bx_2(n)$, then $X_3(k) = aX_1(k) + bX_2(k)$.
2. Symmetry: For $x(n)$ real, $\mathrm{Re}[X(k)] = \mathrm{Re}[X(N - k)]$, $\mathrm{Im}[(k)] = -\mathrm{Im}[X(N - k)]$.

3. Circular shift: By a circular shift of a sequence defined in the interval $0 \le n \le N-1$, we mean that, as values *fall off* from one end of the sequence, they are appended to the other end. Denoting this by $x(n \oplus m)$, we see that positive m means shift left and negative m means shift right. Thus, $x_2(n) = x_1(n \oplus m) \Leftrightarrow X_2(k) = X_1(k)e^{i(2\pi/N)km}$.
4. Duality: $x(n) \Leftrightarrow X(k)$ implies $(1/N)X(n) \Leftrightarrow x(-k)$.
5. Z-transform relation: $X(k) = X(z)|_{z=e^{i(2\pi k/N)}}$, $k = 0,1, \ldots, N-1$.
6. Circular convolution: $x_3(n) = \Sigma_{m=0}^{N-1} x_1(m)x_2(n \ominus m) = \Sigma_{\ell=0}^{N-1} x_1(n \ominus \ell)x_2(\ell)$ where $x_2(n \ominus m)$ corresponds to a circular shift to the right for positive m.

One fast algorithm for calculating DFTs is the radix-2 *fast Fourier transform* developed by J. W. Cooley and J. W. Tucker. Consider the two-point DFT $X(k) = \Sigma_{n=0}^{1} x(n)e^{-i(2\pi/2)nk}$, $k = 0, 1$. Clearly, $X(k) = x(0) + x(1)e^{-i\pi k}$. So, $X(0) = x(0) + x(1)$ and $X(1) = x(0) - x(1)$. This process can be extended to DFTs of length $N = 2^r$, where r is a positive integer. For $N = 2^r$, decompose the N-point DFT into *two* $N/2$-point DFTs. Then, decompose each $N/2$-point DFT into *two* $N/4$-point DFTs, and so on until eventually we have $N/2$ *two*-point DFTs. Computing these as indicated above, we combine them into $N/4$ four-point DFTs and then $N/8$ eight-point DFTs, and so on, until the DFT is computed. The total number of DFT operations (for large N) is $O(N^2)$, and that of the FFT is $O(N \log_2 N)$, quite a saving for large N.

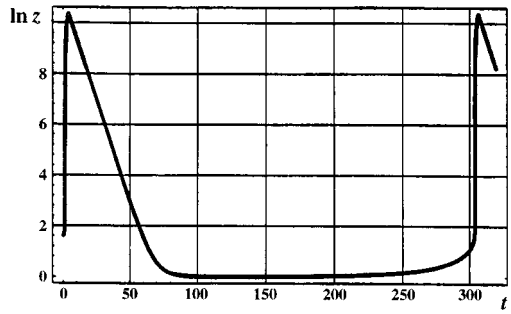


Figure 19.12.1

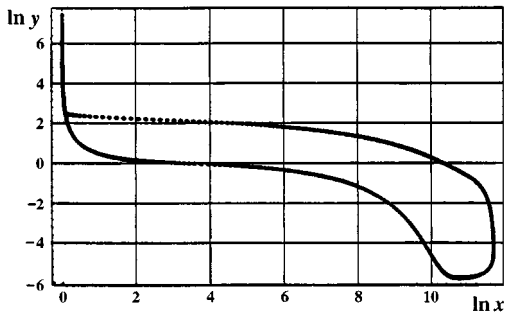


Figure 19.12.2

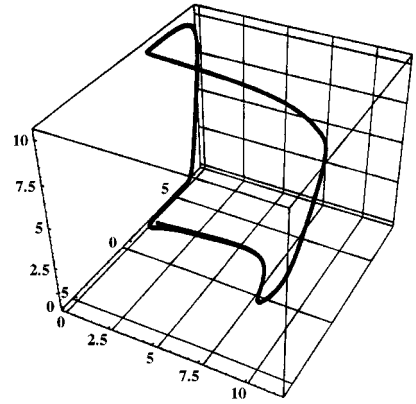


Figure 19.12.3

FIGURES 19.12.1 to 19.12.3 The "Oregonator" is a periodic chemical reaction describable by three nonlinear first-order differential equations. The results (Figure 19.12.1) illustrate the periodic nature of the major chemical versus time. Figure 19.12.2 shows the phase diagram of two of the reactants, and Figure 19.12.3 is the three-dimensional phase diagram of all reactants. The numerical computation was done using a fourth-order Runge-Kuta method on Mathematica by Waltraud Rufeger at the Georgia Institute of Technology.

Software

Some available software is listed here.

General Packages

General software packages include Maple, Mathematica, and Matlab. All contain algorithms for handling a large variety of both numerical and symbolic computations.

Special Packages for Linear Systems

In the IMSL Library, there are three complementary linear system packages of note.

LINPACK is a collection of programs concerned with *direct* methods for general (or full) symmetric, symmetric positive definite, triangular, and tridiagonal matrices. There are also programs for least squares problems, along with the QR algorithm for eigensystems and the singular value decompositions of rectangular matrices. The programs are intended to be completely machine independent, fully portable, and run with good efficiency in most computing environments. The LINPACK User's Guide by Dongarra *et al.* is the basic reference.

ITPACK is a modular set of programs for iterative methods. The package is oriented toward the sparse matrices that arise in the solution of PDEs and other applications. While the programs apply to full matrices, that is rarely profitable. Four basic iteration methods and two convergence acceleration methods are in the package. There is a Jacobi, SOR (with optimum relaxation parameter estimated), symmetric SOR, and reduced system (red-black ordering) iteration, each with semi-iteration and conjugate gradient acceleration. All parameters for these iterations are automatically estimated. The practical and theoretical background for ITPACK is found in Hagemen and Young (1981).

YALEPACK is a substantial collection of programs for sparse matrix computations.

Ordinary Differential Equations Packages

Also in IMSL, one finds such sophisticated software as DVERK, DGEAR, or DREBS for initial value problems. For two-point boundary value problems, one finds DTPTB (use of DVERK and multiple shooting) or DVCPR.

Partial Differential Equations Packages

DISPL was developed and written at Argonne National Laboratory. DISPL is designed for nonlinear second-order PDEs (parabolic, elliptic, hyperbolic (some cases), and parabolic-elliptic). Boundary conditions of a general nature and material interfaces are allowed. The spatial dimension can be either one or two and in Cartesian, cylindrical, or spherical (one dimension only) geometry. The PDEs are reduced to ordinary DEs by Galerkin discretization of the spatial variables. The resulting ordinary DEs in the timelike variable are then solved by an ODE software package (such as GEAR). Software features include graphics capabilities, printed output, dump/restart/facilities, and free format input. DISPL is intended to be an engineering and scientific tool and is not a finely tuned production code for a small set of problems. DISPL makes no effort to control the spatial discretization errors. It has been used to successfully solve a variety of problems in chemical transport, heat and mass transfer, pipe flow, etc.

PDELIB was developed and written at Los Alamos Scientific Laboratory. PDELIB is a library of subroutines to support the numerical solution of evolution equations with a timelike variable and one or two space variables. The routines are grouped into a dozen independent modules according to their function (i.e., accepting initial data, approximating spatial derivatives, advancing the solution in time). Each task is isolated in a distinct module. Within a module, the basic task is further refined into general-purpose flexible lower-level routines. PDELIB can be understood and used at different levels. Within a small period of time, a large class of problems can be solved by a novice. Moreover, it can provide a wide variety of outputs.

DSS/2 is a differential systems simulator developed at Lehigh University as a transportable numerical method of lines (NMOL) code. See also LEANS.

FORSIM is designed for the automated solution of sets of implicitly coupled PDEs of the form

$$\frac{\partial u_i}{\partial t} = \phi_i\left(x, t, u_i, u_j, \ldots, (u_i)_x, \ldots, (u_i)_{xx}, (u_j)_{xx}, \ldots\right), \quad \text{for } i = 1, \ldots, N$$

The user specifies the ϕ_i in a simple FORTRAN subroutine. Finite difference formulas of any order may be selected for the spatial discretization and the spatial grid need not be equidistant. The resulting system of time-dependent ODEs is solved by the method of lines.

SLDGL is a program package for the self-adaptive solution of nonlinear systems of elliptic and parabolic PDEs in up to three space dimensions. Variable step size and variable order are permitted. The discretization error is estimated and used for the determination of the optimum grid and optimum orders. This is the most general of the codes described here (not for hyperbolic systems, of course). This package has seen extensive use in Europe.

FIDISOL (finite difference solver) is a program package for nonlinear systems of two-or three-dimensional elliptic and parabolic systems in rectangular domains or in domains that can be transformed analytically to rectangular domains. This package is actually a redesign of parts of SLDGL, primarily for the solution of large problems on vector computers. It has been tested on the CYBER 205, CRAY-IM, CRAY X-MP/22, and VP 200. The program vectorizes very well and uses the vector arithmetic efficiently. In addition to the numerical solution, a reliable error estimate is computed.

CAVE is a program package for conduction analysis via eigenvalues for three-dimensional geometries using the method of lines. In many problems, much time is saved because only a few terms suffice.

Many industrial and university computing services subscribe to the IMSL Software Library. Announcements of new software appear in *Directions*, a publication of IMSL. A brief description of some IMSL packages applicable to PDEs and associated problems is now given. In addition to those packages just described, two additional software packages bear mention. The first of these, the ELLPACK system, solves elliptic problems in two dimensions with general domains and in three dimensions with box-shaped domains. The system contains over 30 numerical methods modules, thereby providing a means of evaluating and comparing different methods for solving elliptic problems. ELLPACK has a special high-level language making it easy to use. New algorithms can be added or deleted from the system with ease.

Second, TWODEPEP is IMSL's general finite element system for two-dimensional elliptic, parabolic, and eigenvalue problems. The Galerkin finite elements available are triangles with quadratic, cubic, or quartic basic functions, with one edge curved when adjacent to a curved boundary, according to the isoparametric method. Nonlinear equatons are solved by Newton's method, with the resulting linear system solved directly by Gauss elimination. PDE/PROTRAN is also available. It uses triangular elements with piecewise polynomials of degree 2, 3, or 4 to solve quite general steady state, time-dependent, and eigenvalue problems in general two-dimensional regions. There is a simple user input. Additional information may be obtained from IMSL. NASTRAN and STRUDL are two advanced finite element computer systems available from a variety of sources. Another, UNAFEM, has been extensively used.

References

General

Adams, E. and Kulisch, U. (Eds.) 1993. *Scientific Computing with Automatic Result Verification*, Academic Press, Boton, MA.

Gerald, C. F. and Wheatley, P. O. 1984. *Applied Numerical Analysis*, Addison-Wesley, Reading, MA.

Hamming, R. W. 1962. *Numerical Methods for Scientists and Engineers*, McGraw-Hill, New York.

Hildebrand, F. B. 1956. *Introduction to Numerical Analysis*, McGraw-Hill, New York.

Isaacson, E. and Keller, H. B. 1966. *Analysis of Numerical Methods*, John Wiley & Sons, New York.

Kopal, Z. 1955. *Numerical Analysis*, John Wiley & Sons, New York.

Rice, J. R. 1993. *Numerical Methods, Software and Analysis*, 2nd ed. Academic Press, Boston, MA.

Stoer, J. and Bulirsch, R. 1976. *Introduction to Numerical Analysis*, Springer, New York.

Linear Equations

Bodewig, E. 1956. *Matrix Calculus*, Wiley (Interscience), New York.
Hageman, L. A. and Young, D. M. 1981. *Applied Iterative Methods*, Academic Press, Boston, MA.
Varga, R. S. 1962. *Matrix Iterative Numerical Analysis*, John Wiley & Sons, New York.
Young, D. M. 1971. *Iterative Solution of Large-Linear Systems*, Academic Press, Boston, MA.

Ordinary Differential Equations

Aiken, R. C. 1985. *Stiff Computation*, Oxford Unitersity Press, New York.
Gear, C. W. 1971. *Numerical Initial Value Problems in Ordinary Differential Equations*, Prentice Hall, Englewood Cliffs, NJ.
Keller, H. B. 1976. *Numerical Solutions of Two Point Boundary Value Problems*, SIAM, Philadelphia, PA.
Lambert, J. D. 1973. *Computational Methods in Ordinary Differential Equations*, Cambridge University Press, New York.
Milne, W.E. 1953. *Numerical Solution of Differential Equations*, John Wiley & Sons, New York.
Rickey, K. C., Evans, H. R., Griffiths, D. W., and Nethercot, D. A. 1983. *The Finite Element Method — A Basic Introduction for Engineers*, 2nd ed. Halstead Press, New York.
Shampine, L. and Gear, C. W. 1979. A User's View of Solving Stiff Ordinary Differential Equatons, *SIAM Rev.* 21:1–17.

Partial Differential Equations

Ames, W. F. 1993. *Numerical Methods for Partial Differential Equations*, 3d ed. Academic Press, Boston, MA.
Brebbia, C. A. 1984. *Boundary Element Techniques in Computer Aided Engineering*, Martinus Nijhoff, Boston, MA.
Burnett, D. S. 1987. *Finite Element Analysis*, Addison-Wesley, Reading, MA.
Lapidus, L. and Pinder, G. F. 1982. *Numerical Solution of Partial Differential Equations in Science and Engineering*, John Wiley & Sons, New York.
Roache, P. 1972. *Computational Fluid Dynamics*, Hermosa, Albuquerque, NM.

19.13 Experimental Uncertainty Analysis

W.G. Steele and H.W. Coleman

Introduction

The goal of an experiment is to answer a question by measuring a specific variable, X_i, or by determining a result, r, from a functional relationship among measured variables

$$r = r\left(X_1, X_2, \ldots, X_i, \ldots, X_J\right) \tag{19.13.1}$$

In all experiments there is some error that prevents the measurement of the true value of each variable, and therefore, prevents the determination of r_{true}.

Uncertainty analysis is a technique that is used to estimate the interval about a measured variable or a determined result within which the true value is thought to lie with a certain degree of confidence. As discussed by Coleman and Steele (1989), uncertainty analysis is an extremely useful tool for all phases of an experimental program from initial planning (general uncertainty analysis) to detailed design, debugging, test operation, and data analysis (detailed uncertainty analysis).

The application of uncertainty analysis in engineering has evolved considerably since the classic paper of Kline and McClintock (1953). Developments in the field have been especially rapid and significant over the past decade, with the methods formulated by Abernethy and co-workers (1985) that were incorporated into ANSI/ASME Standards in (1984) and (1986) being superseded by the more rigorous approach presented in the International Organization for Standardization (ISO) *Guide to the Expression of Uncertainty in Measurement* (1993). This guide, published in the name of ISO and six other international organizations, has in everything but name established a new international experimental uncertainty standard.

The approach in the ISO *Guide* deals with "Type A" and "Type B" categories of uncertainties, not the more traditional engineering categories of systematic (bias) and precision (random) uncertainties, and is of sufficient complexity that its application in normal engineering practice is unlikely. This issue has been addressed by AGARD Working Group 15 on Quality Assessment for Wind Tunnel Testing, by the Standards Subcommittee of the AIAA Ground Test Technical Committee, and by the ASME Committee PTC 19.1 that is revising the ANSI/ASME Standard (1986). The documents issued by two of these groups (AGARD-AR-304, 1994) and (AIAA Standard S-071-1995, 1995) and in preparation by the ASME Committee present and discuss the additional assumptions necessary to achieve a less complex "large sample" methodology that is consistent with the ISO *Guide*, that is applicable to the vast majority of engineering testing, including most single-sample tests, and that retains the use of the traditional engineering concepts of systematic and precision uncertainties. The range of validity of this "large sample" approximation has been presented by Steele et al. (1994) and by Coleman and Steele (1995). The authors of this section are also preparing a second edition of Coleman and Steele (1989), which will incorporate the ISO *Guide* methodology and will illustrate its use in all aspects of engineering experimentation.

In the following, the uncertainties of individual measured variables and of determined results are discussed. This section concludes with an overview of the use of uncertainty analysis in all phases of an experimental program.

Uncertainty of a Measured Variable

For a measured variable, X_i, the total error is caused by both precision (random) and systematic (bias) errors. This relationship is shown in Figure 19.13.1. The possible measurement values of the variable are scattered in a distribution (here assumed Gaussian) around the parent population mean, μ_i. The parent population mean differs from $(X_i)_{true}$ by an amount called the systematic (or bias) error, β_i. The quantity

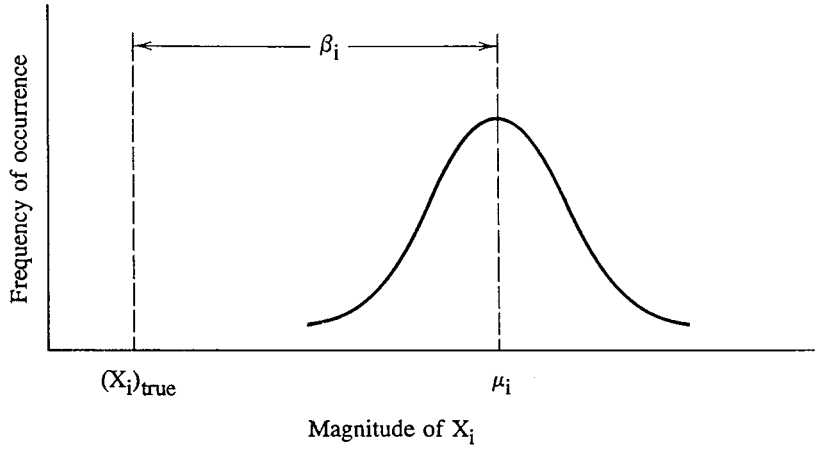


FIGURE 19.13.1 Errors in the measurement of variable X_i.

β_i is the total fixed error that remains in the measurement process after all calibration corrections have been made. In general, there will be several sources of bias error such as calibration standard errors, data acquisition errors, data reduction errors, and test technique errors. There is usually no direct way to measure these errors, so they must be estimated.

For each bias error source, $(\beta_i)_k$, the experimenter must estimate a systematic uncertainty (or bias limit), $(B_i)_k$, such that there is about a 95% confidence that $(B_i)_k \geq |(\beta_i)_k|$. Systematic uncertainties are usually estimated from previous experience, calibration data, analytical models, and the application of sound engineering judgment. For each variable, there will be a set, K_i, of elemental systematic uncertainties, $(B_i)_k$, for the significant fixed error sources. The overall systematic uncertainty for variable X_i is determined from these estimates as

$$B_i^2 = \sum_{k=1}^{K_i} \left(B_i\right)_k^2 \tag{19.13.2}$$

For a discussion on estimating systematic uncertainties (bias limits), see Coleman and Steele (1989).

The estimate of the precision error for a variable is the sample standard deviation, or the estimate of the error associated with the repeatability of a particular measurement. Unlike the systematic error, the precision error varies from reading to reading. As the number of readings, N_i, of a particular variable tends to infinity, the distribution of these readings becomes Gaussian.

The readings used to calculate the sample standard deviation for each variable must be taken over the time frame and conditions which cover the variation in the variable. For example, taking multiple samples of data as a function of time while holding all other conditions constant will identify the random variation associated with the measurement system and the unsteadiness of the test condition. If the sample standard deviation of the variable being measured is also expected to be representative of other possible variations in the measurement, e.g., repeatability of test conditions, variation in test configuration, then these additional error sources will have to be varied while the multiple data samples are taken to determine the standard deviation.

When the value of a variable is determined as the mean, $\overline{X}_i$, of N_i readings, then the sample standard deviation of the mean is

$$S_{\overline{X}_i} = \left\{ \left[\frac{1}{N_i\left(N_i - 1\right)} \right] \sum_{k=1}^{N_i} \left[\left(X_i\right)_k - \overline{X}_i \right]^2 \right\}^{1/2} \tag{19.13.3}$$

where

$$\bar{X}_i = \frac{\sum_{k=1}^{N_i} (X_i)_k}{N_i} \tag{19.13.4}$$

It must be stressed that these N_i readings have to be taken over the appropriate range of variations for X_i as described above.

When only a single reading of a variable is available so that the value used for the variable is X_i, then N_{P_i} previous readings, $(X_{Pi})_k$, must be used to find the standard deviation for the variable as

$$S_{X_i} = \left\{ \frac{1}{N_{Pi} - 1} \sum_{k=1}^{N_{P_i}} \left[\left(X_{P_i}\right)_k - \bar{X}_{P_i} \right]^2 \right\}^{1/2} \tag{19.13.5}$$

where

$$\bar{X}_{P_i} = \frac{1}{N_{P_i}} \sum_{k=1}^{N_{P_i}} \left(X_{P_i}\right)_k \tag{19.13.6}$$

Another situation where previous readings of a variable are useful is when a small current sample size, N_i, is used to calculate the mean value, $\bar{X}_i$, of a variable. If a much larger set of previous readings for the same test conditions is available, then it can be used to calculate a more appropriate standard deviation for the variable (Steele et al., 1993) as

$$S_{\bar{X}_i} = \frac{S_{X_i}}{\sqrt{N_i}} \tag{19.13.7}$$

where N_i is the number of current readings averaged to determine $\bar{X}_i$, and S_{X_i} is computed from N_{P_i} previous readings using Equation (19.13.5). Typically, these larger data sets are taken in the early "shake-down" or "debugging" phases of an experimental program.

For many engineering applications, the "large sample" approximation applies, and the uncertainty for variable i (X_i or $\bar{X}_i$) is

$$U_i = \sqrt{B_i^2 + \left(2S_i\right)^2} \tag{19.13.8}$$

where S_i is found from the applicable Equation (19.13.3), (19.13.5) or (19.13.7). The interval $X_i \pm U_i$ or $\bar{X}_i + U_i$, as appropriate, should contain $(X_i)_{\text{true}}$ 95 times out of 100. If a small number of samples (N_i or $N_{P_i} < 10$) is used to determine $S_{\bar{X}_i}$ or S_{X_i}, then the "large sample" approximation may not apply and the methods in ISO (1993) or Coleman and Steele (1995) should be used to find U_i.

Uncertainty of a Result

Consider an experimental result that is determined for J measured variables as

$$r = r\left(X_1, X_2, \ldots, X_i, \ldots, X_J\right)$$

where some variables may be single readings and others may be mean values. A typical mechanical engineering experiment would be the determination of the heat transfer in a heat exchanger as

$$q = \dot{m} c_p \left(T_o - T_i\right) \tag{19.13.9}$$

where q is the heat rate, $\dot{m}$ is the flow rate, c_p is the fluid specific heat, and T_o and T_i are the heated fluid outlet and inlet temperatures, respectively. For the "large sample" approximation, U_r is found as

$$U_r = \sqrt{B_r^2 + \left(2S_r\right)^2} \tag{19.13.10}$$

where B_r is the systematic uncertainty of the result

$$B_r^2 = \sum_{i=1}^{J} \left(\theta_i B_i\right)^2 + 2 \sum_{i=1}^{J-1} \sum_{k=i+1}^{J} \theta_i \theta_k B_{ik} \tag{19.13.11}$$

with

$$\theta_i = \frac{\partial r}{\partial X_i} \tag{19.13.12}$$

and S_r is the standard deviation of the result

$$S_r^2 = \sum_{i=1}^{J} \left(\theta_i S_i\right)^2 \tag{19.13.13}$$

The term B_{ik} in Equation (19.13.11) is the covariance of the systematic uncertainties. When the elemental systematic uncertainties for two separately measured variables are related, for instance when the transducers used to measure different variables are each calibrated against the same standard, the systematic uncertainties are said to be correlated and the covariance of the systematic errors is nonzero. The significance of correlated systematic uncertainties is that they can have the effect of either decreasing or increasing the uncertainty in the result. B_{ik} is determined by summing the products of the elemental systematic uncertainties for variables i and k that arise from the same source and are therefore perfectly correlated (Brown et al., 1996) as

$$B_{ik} = \sum_{\alpha=1}^{L} \left(B_i\right)_\alpha \left(B_k\right)_\alpha \tag{19.13.14}$$

where L is the number of elemental systematic error sources that are common for measurements X_i and X_k.

If, for example,

$$r = r\left(X_1, X_2\right) \tag{19.13.15}$$

and it is possible for portions of the systematic uncertainties B_1 and B_2 to arise from the same source(s), Equation (19.13.11) gives

$$B_r^2 = \theta_1^2 B_1^2 + \theta_2^2 B_2^2 + 2\theta_1 \theta_2 B_{12} \tag{19.13.16}$$

For a case in which the measurements of X_1 and X_2 are each influenced by four elemental systematic error sources and sources two and three are the same for both X_1 and X_2, Equation (19.13.2) gives

$$B_1^2 = \left(B_1\right)_1^2 + \left(B_1\right)_2^2 + \left(B_1\right)_3^2 + \left(B_1\right)_4^2 \tag{19.13.17}$$

and

$$B_2^2 = \left(B_2\right)_1^2 + \left(B_2\right)_2^2 + \left(B_2\right)_3^2 + \left(B_2\right)_4^2 \tag{19.13.18}$$

while Equation (19.13.14) gives

$$B_{12} = \left(B_1\right)_2 \left(B_2\right)_2 + \left(B_1\right)_3 \left(B_2\right)_3 \tag{19.13.19}$$

In the general case, there would be additional terms in the expression for the standard deviation of the result, S_r, (Equation 19.13.13) to take into account the possibility of precision errors in different variables being correlated. These terms have traditionally been neglected, although precision errors in different variables caused by the same uncontrolled factor(s) are certainly possible and can have a substantial impact on the value of S_r (Hudson et al., 1996). In such cases, one would need to acquire sufficient data to allow a valid estimate of the precision covariance terms using standard statistical techniques (ISO, 1993). Note, however, that if multiple test results over an appropriate time period are available, these can be used to directly determine S_r. This value of the standard deviation of the result implicitly includes the correlated error effect.

If a test is performed so that M multiple sets of measurements $(X_1, X_2, \ldots, X_J)_k$ at the same test condition are obtained, then M results can be determined using Equation (19.13.1) and a mean result, $\bar{r}$, can be determined using

$$\bar{r} = \frac{1}{M}\sum_{k=1}^{M} r_k \tag{19.13.20}$$

The standard deviation of the sample of M results, S_r, is calculated as

$$S_r = \left[\frac{1}{M-1}\sum_{k=1}^{M}\left(r_k - \bar{r}\right)^2\right]^{1/2} \tag{19.13.21}$$

The uncertainty associated with the mean result, $\bar{r}$, for the "large sample" approximation is then

$$U_{\bar{r}} = \sqrt{B_r^2 + \left(2S_{\bar{r}}\right)^2} \tag{19.13.22}$$

where

$$S_{\bar{r}} = \frac{S_r}{\sqrt{M}} \tag{19.13.23}$$

and where B_r is given by Equation (19.13.11).

The "large sample" approximation for the uncertainty of a determined result (Equations (19.13.10) or (19.13.22)) applies for most engineering applications even when some of the variables have fewer

than 10 samples. A detailed discussion of the applicability of this approximation is given in Steele et al. (1994) and Coleman and Steele (1995).

The determination of U_r from S_r (or $S_{\bar{r}}$) and B_r using the "large sample" approximation is called detailed uncertainty analysis (Coleman and Steele, 1989). The interval r (or $\bar{r}$) $\pm U_r$ (or $U_{\bar{r}}$) should contain r_{true} 95 times out of 100. As discussed in the next section, detailed uncertainty analysis is an extremely useful tool in an experimental program. However, in the early stages of the program, it is also useful to estimate the overall uncertainty for each variable, U_i. The overall uncertainty of the result is then determined as

$$U_r^2 = \sum_{k=1}^{J} \left(\theta_i U_i\right)^2 \tag{19.13.24}$$

This determination of U_r is called general uncertainty analysis.

Using Uncertainty Analysis in Experimentation

The first item that should be considered in any experimental program is "What question are we trying to answer?" Another key item is how accurately do we need to know the answer, or what "degree of goodness" is required? With these two items specified, general uncertainty analysis can be used in the planning phase of an experiment to evaluate the possible uncertainties from the various approaches that might be used to answer the question being addressed. Critical measurements that will contribute most to the uncertainty of the result can also be identified.

Once past the planning, or preliminary design phase of the experiment, the effects of systematic errors and precision errors are considered separately using the techniques of detailed uncertainty analysis. In the design phase of the experiment, estimates are made of the systematic and precision uncertainties, B_r and $2S_r$, expected in the experimental result. These detailed design considerations guide the decisions made during the construction phase of the experiment.

After the test is constructed, a debugging phase is required before production tests are begun. In the debugging phase, multiple tests are run and the precision uncertainty determined from them is compared with the $2S_r$ value estimated in the design phase. Also, a check is made to see if the test results plus and minus U_r compare favorably with known results for certain ranges of operation. If these checks are not successful, then further test design, construction, and debugging is required.

Once the test operation is fully understood, the execution phase can begin. In this phase, balance checks can be used to monitor the operation of the test apparatus. In a balance check, a quantity, such as flow rate, is determined by different means and the difference in the two determinations, z, is compared to the ideal value of zero. For the balance check to be satisfied, the quantity z must be less than or equal to U_z.

Uncertainty analysis will of course play a key role in the data analysis and reporting phases of an experiment. When the experimental results are reported, the uncertainties should be given along with the systematic uncertainty, B_r, the precision uncertainty, $2S_r$, and the associated confidence level, usually 95%.

References

Abernethy, R.B., Benedict, R.P., and Dowdell, R.B. 1985. ASME Measurement Uncertainty. *J. Fluids Eng.*, 107, 161–164.

AGARD-AR-304. 1994. *Quality Assessment for Wind Tunnel Testing*. AGARD, Neuilly Sur Seine, France.

AIAA Standard S-071-1995. 1995. *Assessment of Wind Tunnel Data Uncertainty.* AIAA, Washington, D.C.

ANSI/ASME MFC-2M-1983. 1984. *Measurement Uncertainty for Fluid Flow in Closed Conduits.* ASME, New York.

ANSI/ASME PTC 19.1-1985, Part 1. 1986. *Measurement Uncertainty.* ASME, New York.

Coleman, H.W. and Steele, W.G. 1989. *Experimentation and Uncertainty Analysis for Engineers*. John Wiley & Sons, New York.

Coleman, H.W. and Steele, W.G. 1995. Engineering Application of Experimental Uncertainty Analysis. *AIAA Journal*, 33(10), 1888–1896.

Brown, K.B., Coleman, H.W., Steele, W.G., and Taylor, R.P. 1996. Evaluation of Correlated Bias Approximations in Experimental Uncertainty Analysis. *AIAA J.*, 34(5), 1013–1018.

Hudson, S.T., Bordelon, Jr., W.J., and Coleman, H.W. 1996. Effect of Correlated Precision Errors on the Uncertainty of a Subsonic Venturi Calibration. *AIAA J.*, 34(9), 1862–1867.

Kline, S.J. and McClintock, F.A. 1953. Describing Uncertainties in Single-Sample Experiments. *Mech. Eng.*, 75, 3–8.

ISO. 1993. *Guide to the Expression of Uncertainty in Measurement.* ISO, Geneva, Switzerland.

Steele, W.G., Taylor, R.P., Burrell, R.E., and Coleman, H.W. 1993. Use of Previous Experience to Estimate Precision Uncertainty of Small Sample Experiments. *AIAA J.*, 31(10), 1891–1896.

Steele, W.G., Ferguson, R.A., Taylor, R.P., and Coleman, H.W. 1994. Comparison of ANSI/ASME and ISO Models for Calculation of Uncertainty. *ISA Trans.*, 33, 339–352.

19.14 Chaos

R. L. Kautz

Introduction

Since the time of Newton, the science of dynamics has provided quantitative descriptions of regular motion, from a pendulum's swing to a planet's orbit, expressed in terms of differential equations. However, the role of Newtonian mechanics has recently expanded with the realization that it can also describe chaotic motion. In elementary terms, **chaos** can be defined as **pseudorandom** behavior observed in the steady-state dynamics of a deterministic **nonlinear system**. How can motion be pseudorandom, or random according to statistical tests and yet be entirely predictable? This is just one of the paradoxes of chaotic motion, which is globally stable but locally unstable, predictable in principle but not in practice, and geometrically complex but derived from simple equations.

The strange nature of chaotic motion was first understood by Henri Poincaré, who established the mathematical foundations of chaos in a treatise published in 1890 (Holmes, 1990). However, the practical importance of chaos began to be widely appreciated only in the 1960s, beginning with the work of Edward Lorenz (1963), a meteorologist who discovered chaos in a simple model for fluid convection. Today, chaos is understood to explain a wide variety of apparently random natural phenomena, ranging from dripping faucets (Martien et al., 1985), to the flutter of a falling leaf (Tanabe and Kaneko, 1994), to the irregular rotation of a moon of Saturn (Wisdom et al., 1984).

Although chaos is used purposely to provide an element of unpredictability in some toys and carnival rides (Kautz and Huggard, 1994), it is important from an engineering point of view primarily as a phenomenon to be avoided. Perhaps the simplest scenario arises when a nonlinear mechanism is used to achieve a desired effect, such as the synchronization of two oscillators. In many such cases, the degree of nonlinearity must be chosen carefully: strong enough to ensure the desired effect but not so strong that chaos results. In another scenario, an engineer might be required to deal with an intrinsically chaotic system. In this case, if the system can be modeled mathematically, then a small feedback signal can often be applied to eliminate the chaos (Ott et al., 1990). For example, low-energy feedback has been used to suppress chaotic behavior in a thermal convection loop (Singer et al., 1991). As such considerations suggest, chaos in rapidly becoming an important topic for engineers.

Flows, Attractors, and Liapunov Exponents

Dynamic systems can generally be described mathematically in terms of a set of differential equations of the form.

$$d\mathbf{x}(t)/dt = \mathbf{F}\left[\mathbf{x}(t)\right] \tag{19.14.1}$$

where $x = (x_1, \ldots, x_N)$ is an N-dimensional vector called the **state vector** and the vector function $\mathbf{F} = F_1(\mathrm{x}), \ldots, F_N(\mathbf{x}))$ defines how the state vector changes with time. In mechanics, the state variables x_i are typically the positions and velocities associated with the potential and kinetic energies of the system. Because the state vector at times $t > 0$ depends only on the initial state vector $\mathbf{x}(0)$, the system defined by Equation (19.14.1) is deterministic, and its motion is in principle exactly predictable.

The properties of a dynamic system are often visualized most readily in terms of trajectories $\mathbf{x}(t)$ plotted in **state space**, where points are defined by the coordinates $(x_1, \ldots, x_N)$. As an example, consider the motion of a damped pendulum defined by the normalized equation

$$d^2\theta/dt^2 = -\sin\theta - \rho d\theta/dt \tag{19.14.2}$$

which expresses the angular acceleration $d^2\theta/dt^2$ in terms of the gravitational torque – sin θ and a damping torque – $\rho d\theta/dt$ proportional to the angular velocity $\nu = d\theta/dt$. If we define the state vector as $x = (x_1,x_2) = (\theta,\nu)$, then Equation (19.14.2) can be written in the form of Equation (19.14.1) with $\mathbf{F} = (x_2, -\sin x_1 - \rho x_2)$. In this case, the state space is two dimensional, and a typical trajectory is a spiral, as shown in Figure 19.14.1 for the initial condition x(0) = (0,1). If additional trajectories, corresponding to other initial conditions, were plotted in Figure 19.14.1, we would obtain a set of interleaved spirals, all converging on the point **x** = (0,0). Because the direction of a trajectory passing through a given point is uniquely defined by Equation (19.14.1), state–space trajectories can never cross, and, by analogy with the motion of a fluid, the set of all of trajectories is called a flow.

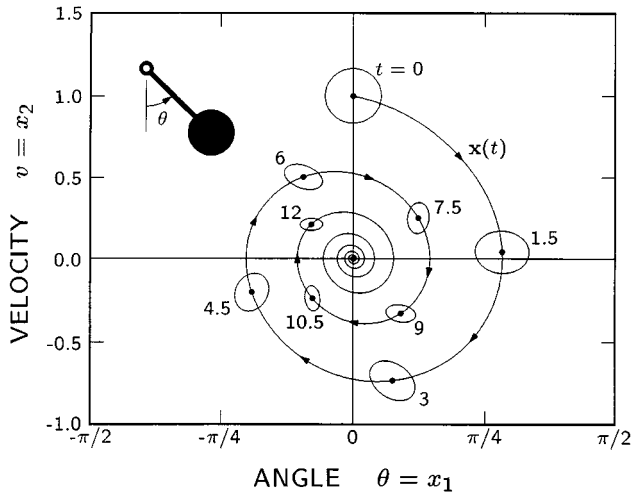


FIGURE 19.14.1 The state-space trajectory **x**(*t*) for a pendulum with a damping coefficient ρ = 0.2 for the initial condition **x**(0) = (0,1). The evolution of trajectories initialized in a small circle surrounding **x** = (0,1) is indicated by the ellipses plotted at time intervals of $\Delta t = 1.5$.

The tendency of a flow to converge toward a single point or other restricted subset of state space is characteristic of dissipative systems like the damped pendulum. Such an asymptotic set, called an attracting set or **attractor**, can be a fixed point (for which **F(x)** = 0) as in Figure 19.14.1, but might also be a periodic or chaotic trajectory. The convergence of neighboring trajectories is suggested in Figure 19.14.1 by a series of ellipses spaced at time intervals $\Delta t = 1.5$ that track the flow of all trajectories originating within the circle specified at $t = 0$. In general, the contraction of an infinitesimal state–space volume V as it moves with the flow is given by

$$V^{-1}\,\partial V/\partial t = \nabla \cdot \mathbf{F} \tag{19.14.3}$$

where $\nabla \cdot \mathbf{F} = \Sigma_{i=1}^{N} \partial F_i/x_i$ is the divergence of **F**. For the damped pendulum, $\nabla \cdot \mathbf{F} = \rho$, so the area of the ellipse shown in Figure 19.14.1 shrinks exponentially as $V(t) = V(0)\exp(-\rho t)$. The contraction of state–space volumes explains the existence of attractors in dissipative systems, but in conservative systems such as the pendulum with ρ = 0, state–space volumes are preserved, and trajectories are instead confined to constant-energy surfaces.

While the existence of chaotic behavior is generally difficult to predict, two essential conditions are easily stated. First, the complex topology of a chaotic trajectory can exist only in a state–space of dimension $N \geq 3$. Thus, the pendulum defined by Equation (19.14.2) cannot be chaotic because $N = 2$ for this system. Second, a system must be nonlinear to exhibit chaotic behavior. Linear systems, for which any linear combination $c_1x_a(t) + c_2x_b(t)$ of two solutions $x_a(t)$ and $x_b(t)$ is also a solution, are mathematically simple and amenable to analysis. In contrast, nonlinear systems are noted for their

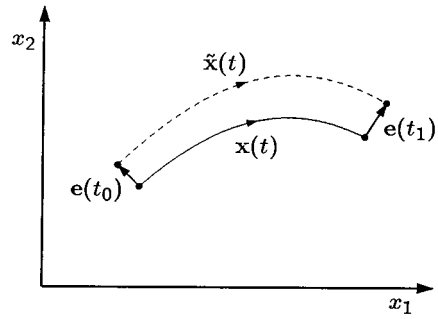


FIGURE 19.14.2 A trajectory $\mathbf{x}(t)$ and a neighboring trajectory $\tilde{\mathbf{x}}(t)$ plotted in state space from time t_0 to t_1. The vectors $\mathbf{e}(t_0)$ and $\mathbf{e}(t_1)$ indicate the deviation of $\tilde{\mathbf{x}}(t)$ from $\mathbf{x}(t)$ at times t_0 and t_1.

intractability. Thus, chaotic behavior is of necessity explored more frequently by numerical simulation than mathematical analysis, a fact that helps explain why the prevalence of chaos was discovered only after the advent of efficient computation.

A useful criterion for the existence of chaos can be developed from an analysis of a trajectory's local stability. As sketched in Figure 19.14.2, the local stability of a trajectory x(t) is determined by considering a neighboring trajectory $\tilde{x}(t)$ initiated by an infinitesimal deviation e(t_0) from x(t) at time t_0. The deviation vector e(t) = $\tilde{x}(t)$ – x(t) at times $t_1 > t_0$ can be expressed in terms of the Jacobian matrix

$$J_{ij}(t_1,t_0) = \partial x_i(t_1)/\partial x_j(t_0) \tag{19.14.4}$$

which measures the change in state variable x_i at time t_1 due to a change in x_j at time t_0. From the Jacobian's definition, we have $\mathbf{e}(t_1) = \mathbf{J}(t_1,t_0)\mathbf{e}(t_0)$. Although the local stability of $\mathbf{x}(t)$ is determined simply by whether deviations grow or decay in time, the analysis is complicated by the fact that deviation vectors can also rotate, as suggested in Figure 19.14.2. Fortunately, an arbitrary deviation can be written in terms of the eigenvectors $\mathbf{e}^{(i)}$ of the Jacobian, defined by

$$\mathbf{J}(t_1,t_0)\mathbf{e}^{(i)} = \mu_i(t_1,t_0)\mathbf{e}^{(i)} \tag{19.14.5}$$

which are simply scaled by the eigenvalues $\mu_i(t_1,t_0)$ without rotation. Thus, the N eigenvalues of the Jacobian matrix provide complete information about the growth of deviations. Anticipating that the asymptotic growth will be exponential in time, we define the **Liapunov exponents**,

$$\lambda_i = \lim_{t_1\to\infty} \frac{\ln|\mu_i(t_1,t_0)|}{t_1 - t_0} \tag{19.14.6}$$

Because any deviation can be broken into components that grow or decay asymptotically as $\exp(\lambda_i t)$, the N exponents associated with a trajectory determine its local stability.

In dissipative systems, chaos can be defined as motion on an attractor for which one or more Liapunov exponents are positive. Chaotic motion thus combines global stability with local instability in that motion is confined to the attractor, generally a bounded region of state space, but small deviations grow exponentially in time. This mixture of stability and instability in chaotic motion is evident in the behavior of an infinitesimal deviation ellipsoid similar to the finite ellipse shown in Figure 19.14.1. Because some λ_i are positive, an ellipsoid centered on a chaotic trajectory will expand exponentially in some directions. On the other hand, because state-space volumes always contract in dissipative systems and the asymptotic volume of the ellipsoid scales as $\exp(\Lambda t)$, where $\Lambda = \Sigma_{i=1}^{N} \lambda_i$, the sum of the negative exponents must be greater in magnitude than the sum of the positive exponents. Thus, a deviation ellipsoid tracking a

chaotic trajectory expands in some directions while contracting in others. However, because an arbitrary deviation almost always includes a component in a direction of expansion, nearly all trajectories neighboring a chaotic trajectory diverge exponentially.

According to our definition of chaos, neighboring trajectories must diverge exponentially and yet remain on the attractor. How is this possible? Given that the attractor is confined to a bounded region of state space, perpetual divergence can occur only for trajectories that differ infinitesimally. Finite deviations grow exponentially at first but are limited by the bounds of the chaotic attractor and eventually shrink again. The full picture can be understood by following the evolution of a small state-space volume selected in the neighborhood of the chaotic attractor. Initially, the volume expands in some directions and contracts in others. When the expansion becomes too great, however, the volume begins to fold back on itself so that trajectories initially separated by the expansion are brought close together again. As time passes, this stretching and folding is repeated over and over in a process that is often likened to kneading bread or pulling taffy.

Because all neighboring volumes approach the attractor, the stretching and folding process leads to an attracting set that is an infinitely complex filigree of interleaved surfaces. Thus, while the differential equation that defines chaotic motion can be very simple, the resulting attractor is highly complex. Chaotic attractors fall into a class of geometric objects called **fractals**, which are characterized by the presence of structure at arbitrarily small scales and by a dimension that is generally fractional. While the existence of objects with dimensions falling between those of a point and a line, a line and a surface, or a surface and a volume may seem mysterious, fractional dimensions result when dimension is defined by how much of an object is apparent at various scales of resolution. For the dynamical systems encompassed by Equation (19.14.1), the fractal dimension D of a chaotic attractor falls in the range of $2 < D < N$ where N is the dimension of the state space. Thus, the dimension of a chaotic attractor is large enough that trajectories can continually explore new territory within a bounded region of state space but small enough that the attractor occupies no volume of the space.

Synchronous Motor

As an example of a system that exhibits chaos, we consider a simple model for a synchronous motor that might be used in a clock. As shown in Figure 19.14.3, the motor consists of a permanent-magnet rotor subjected to a uniform oscillatory magnetic field $B \sin t$ provided by the stator. In dimensionless notation, its equation of motion is

$$d^2\theta/dt^2 = -f \sin t \sin\theta - \rho d\theta/dt \tag{19.14.7}$$

where $d^2\theta/dt^2$ is the angular acceleration of the rotor, $-f \sin t \sin\theta$ is the torque due to the interaction of the rotor's magnetic moment with the stator field, and $-\rho d\theta/dt$ is a viscous damping torque. Although Equation (19.14.7) is explicitly time dependent, it can be case in the form of Equation (19.14.1) by

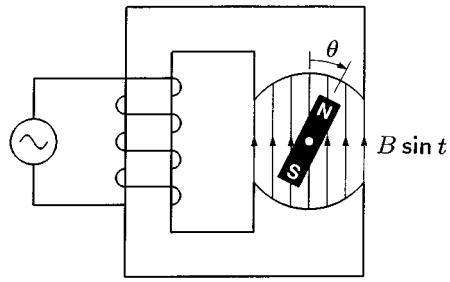


FIGURE 19.14.3 A synchronous motor, consisting of a permanent magnet free to rotate in a uniform magnetic field $B \sin t$ with an amplitude that varies sinusoidally in time.

defining the state vector as $\mathbf{x} = (x_1,x_2,x_3) = (\theta,\nu,t)$, where $\nu = d\theta/dt$ is the angular velocity, and by defining the flow as $\mathbf{F} = (x_2, -f \sin x_3 \sin x_1 - \rho x_2, 1)$. The state space is thus three dimensional and large enough to allow chaotic motion. Equation (19.14.7) is also nonlinear due to the term $-f \sin t \sin \theta$, since $\sin(\theta_a + \theta_b)$ is not generally equal to $\sin \theta_a + \sin \theta_b$. Chaos in this system has been investigated by several authors (Ballico et al., 1990).

By intent, the motor uses nonlinearity to synchronize the motion of the rotor with the oscillatory stator field, so it evolves exactly once during each field oscillation. Although synchronization can occur over a range of system parameters, proper operation requires that the drive amplitude f, which measures the strength of the nonlinearity, be chosen large enough to produce the desired rotation but not so large that chaos results. Calculating the motor's dynamics for $\rho = 0.2$, we find that the rotor oscillates without rotating for f less than 0.40 and that the intended rotation is obtained for $0.40 < \rho < 1.87$. The periodic attractor corresponding to synchronized rotation is shown for $f = 1$ in Figure 19.14.4(a). Here the three-dimensional state-space trajectory is projected onto the (x_1,x_2) or (θ,ν) plane, and a dot marks the point in the rotation cycle at which $t = 0$ modulo 2π. As Figure 19.14.4(a) indicates, the rotor advances by exactly 2π during each drive cycle.

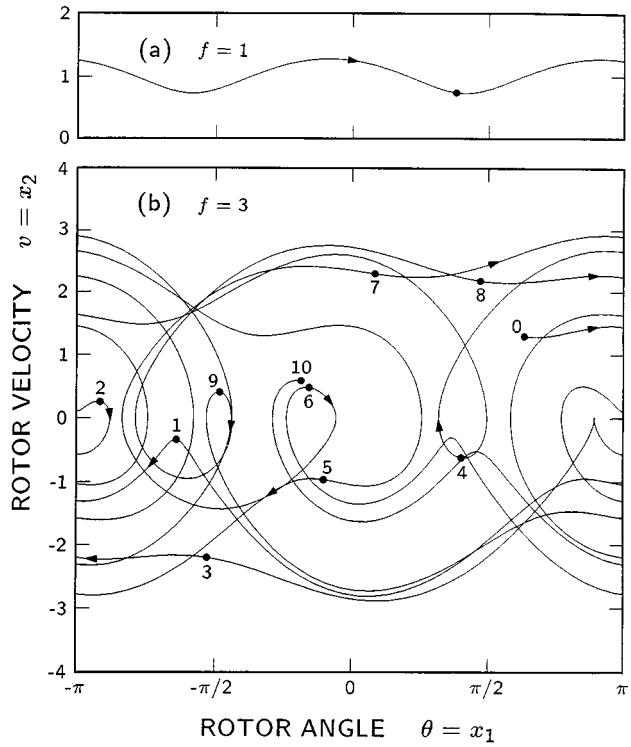


FIGURE 19.14.4 State-space trajectories projected onto the (x_1,x_2) or (θ,ν) plane, showing attractors of the synchronous motor for $\rho = 0.2$ and two drive amplitudes, $f = 1$ and 3. Dots mark the state of the system at the beginning of each drive cycle ($t = 0$ modulo 2π). The angles $\theta = \pi$ and $-\pi$ are equivalent.

The utility of the motor hinges on the stability of the synchronous rotation pattern shown in Figure 19.14.4(a). This periodic pattern is the steady–state motion that develops after initial transients decay and represents the final asymptotic trajectory resulting for initial conditions chosen from a wide area of state space. Because the flow approaches this attracting set from all neighboring points, the effect of a perturbation that displaces the system from the attractor is short lived. This stability is reflected in the Liapunov exponents of the attractor: $\lambda_1 = 0$ and $\lambda_2 = \lambda_3 = -0.100$. The zero exponent is associated with deviations coincident with the direction of the trajectory and is a feature common to all bounded attractors

other than fixed points. The zero exponent results because the system is neutrally stable with respect to offsets in the time coordinate. The exponents of –0.100 are associated with deviations transverse to the trajectory and indicate that these deviations decay exponentially with a characteristic time of 1.6 drive cycles. The negative exponents imply that the synchrony between the rotor and the field is maintained in spite of noise or small variations in system parameters, as required of a clock motor.

For drive amplitudes greater than $f = 1.87$, the rotor generally does not advance by precisely 2π during every drive cycle, and its motion is commonly chaotic. An example of chaotic behavior is illustrated for $f = 3$ by the trajectory plotted in Figure 19.14.4(b) over an interval of 10 drive cycles. In this figure, sequentially numbered dots mark he beginning of each drive cycle. When considered cycle by cycle, the trajectory proves to be a haphazard sequence of oscillations, forward rotations, and reverse rotations. Although we might suppose that this motion is just an initial transient, it is instead characteristic of the steady–state behavior of the motor. If extended, the trajectory continued with an apparently random mixture of oscillation and rotation, without approaching a repetitive cycle. The motion is aptly described as chaotic.

The geometry of the chaotic attractor sampled in Figure 19.14.4(b) is revealed more fully in Figure 19.14.5. Here we plot points (θ,ν) recording the instantaneous angle and velocity of the rotor at the beginning of each drive cycle for 100,000 successive cycles, Figure 19.14.5 displays the three-dimensional attractor called a **Poincaré section**, at its intersection with the planes $t = x_3 = 0$ modulo 2π, corresponding to equivalent times in the drive cycle. For the periodic attractor of Figure 19.14.4(a), the rotor returns to the same position and velocity at the beginning of each drive cycle, so its Poincaré section is a single point, the dot in this figure. For chaotic motion, in contrast, we obtain the complex swirl of points shown in Figure 19.14.5. If the system is initialized at a point far from the swirl, the motion quickly converges to this attracting set. On succeeding drive cycles, the state of the system jumps from one part of the swirl to another in an apparently random fashion that continues indefinitely. As the number of plotted points approaches infinity, the swirl becomes a cross section of the chaotic attractor. Thus, Figure 19.14.5 approximates a slice through the infinite filigree of interleaved surfaces that compose the attracting set. In this case, the fractal dimension of the attractor is 2.52 and that of its Poincaré section is 1.52.

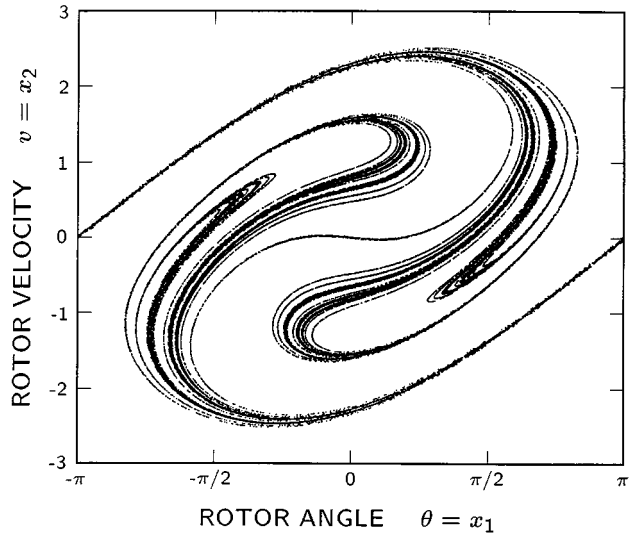


FIGURE 19.14.5 Poincaré section of a chaotic attractor of the synchronous motor with $\rho = 0.2$ and $f = 3$, obtained by plotting points $(x_1,x_2) = (\theta,\nu)$ corresponding to the position and velocity of the rotor at the beginning of 100,000 successive drive cycles.

The computed Liapunov exponents of the chaotic solution at $\rho = 0.2$ and $f = 3$ are $\lambda_1 = 0$, $\lambda_2 = 0.213$, and $\lambda_3 = -0.413$. As for the periodic solution, the zero exponent implies neutral stability associated with deviations directed along a given trajectory. The positive exponent, which signifies the presence of chaos, is associated with deviations transverse to the given trajectory but tangent to the surface of the attracting set in which it is embedded. The positive exponent implies that such deviations grow exponentially in time and that neighboring trajectories on the chaotic attractor diverge exponentially, a property characteristic of chaotic motion. The negative exponent is associated with deviations transverse to the surface of the attractor and assures the exponential decay of displacements from the attracting set. Thus, the Liapunov exponents reflect both the stability of the chaotic attractor and the instability of a given chaotic trajectory with respect to neighboring trajectories.

One sequence of a positive Liapunov exponent is a practical limitation on our ability to predict the future state of a chaotic system. This limitation is illustrated in Figure 19.14.6, where we plot a given chaotic trajectory (solid line) and three perturbed trajectories (dashed lines) that result by offsetting the initial phase of the given solution by various deviations $e_1(0)$. When the initial angular offset is $e_1(0) = 10^{-3}$ radian, the perturbed trajectory (short dash) closely tracks the given trajectory for about seven drive cycles before the deviation become significant. After seven drive cycles, the perturbed trajectory is virtually independent of the given trajectory, even though it is confined to the same attractor. Similarly, initial offsets of 10^{-6} and 10^{-9} radian lead to perturbed trajectories (medium and long dash) that track the given trajectory for about 12 and 17 drive cycles, respectively, before deviations become significant. These results reflect the fact that small deviations grow exponentially and, in the present case, increase on average by a factor of 10 every 1.7 drive cycles. If the position of the rotor is to be predicted with an accuracy of 10^{-1} radian after 20 drive cycles, its initial angle must be known to better than 10^{-13} radian, and the calculation must be carried out with at least 14 significant digits. If a similar prediction is to be made over 40 drive cycles, then 25 significant digits are required. Thus, even though chaotic motion is predictable in principle, the state of a chaotic system can be accurately predicted in practice for only a short time into the future. According to Lorenz (1993), this effect explains why weather forecasts are of limited significance beyond a few days.

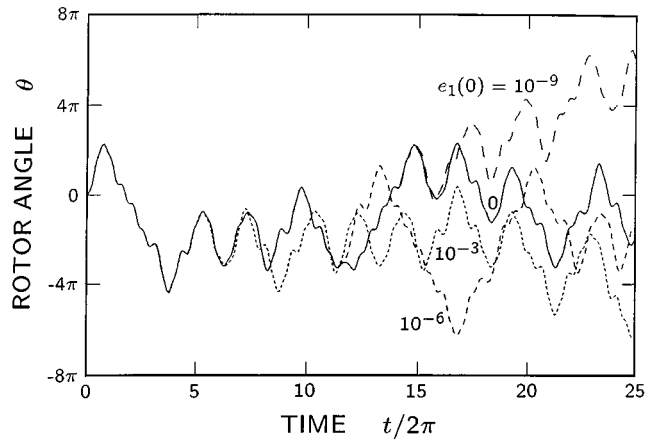


FIGURE 19.14.6 Rotor angle as a function of time for chaotic trajectories of the synchronous motor with $\rho = 0.2$ and $f = 3$. Solid line shows a given trajectory and dashed lines show perturbed trajectories resulting from initial angular deviations of $e_1(0) = 10^{-3}$ (short dash), 10^{-6} (medium dash), and 10^{-9} (long dash).

This pseudorandom nature of chaotic motion is illustrated in Figure 19.14.7 for the synchronous motor by a plot of the net rotation during each of 100 successive drive cycles. Although this sequence of rotations results from solving a deterministic equation, it is apparently random, jumping erratically between forward and reverse rotations of various magnitudes up to about 1.3 revolutions. The situation

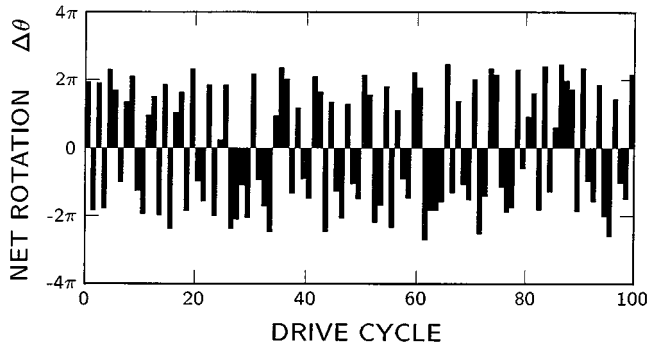


FIGURE 19.14.7 Net rotation of a synchronous motor during each of 100 successive drive cycles, illustrating chaotic motion for $\rho = 0.2$ and $f = 3$. By definition, $\Delta\theta = \theta(2\pi n) - \theta(2\pi(n-1))$ on the nth drive cycle.

is similar to that of a digital random number generator, in which a deterministic algorithm is used to produce a sequence of pseudorandom numbers. In fact, the similarity is not coincidental since chaotic processes often underlie such algorithms (Li, 1978). For the synchronous motor, statistical analysis reveals almost no correlation between rotations separated by more than a few drive cycles. This statistical independence is a result of the motor's positive Liapunov exponent. Because neighboring trajectories diverge exponentially, a small region of the attractor can quickly expand to cover the entire attractor, and a small range of rotations on one drive cycle can lead to almost any possible rotation a few cycles later. Thus, there is little correlation between rotations separated by a few drive cycles, and on this time scale the motor appears to select randomly between the possible rotations.

From an engineering point of view, the problem of chaotic behavior in the synchronous motor can be solved simply by selecting a drive amplitude in the range of $0.40 < f < 1.87$. Within this range, the strength of the nonlinearity is large enough to produce synchronization but not so large as to produce chaos. As this example suggests, it is important to recognize that erratic, apparently random motion can be an intrinsic property of a dynamic system and is not necessarily a product of external noise. Searching a real motor for a source of noise to explain the behavior shown in Figure 19.14.7 would be wasted effort since the cause is hidden in a noise-free differential equation. Clearly, chaotic motion is a possibility that every engineer should understand.

Defining Terms

Attractor: A set of points in state space to which neighboring trajectories converge in the limit of large time.

Chaos: Pseudorandom behavior observed in the steady–state dynamics of a deterministic nonlinear system.

Fractal: A geometric object characterized by the presence of structure at arbitrarily small scales and by a dimension that is generally fractional.

Liapunov exponent: One of N constants λ_i that characterize the asymptotic exponential growth of infinitesimal deviations from a trajectory in an N-dimensional state space. Various components of a deviation grow or decay on average in proportion to $\exp(\lambda_i t)$.

Nonlinear system: A system of equations for which a linear combination of two solutions is not generally a solution.

Poincaré section: A cross section of a state-space trajectory formed by the intersection of the trajectory with a plane defined by a specified value of one state variable.

Pseudorandom: Random according to statistical tests but derived from a deterministic process.

State space: The space spanned by state vectors.

State vector: A vector **x** whose components are the variables, generally positions and velocities, that define the time evolution of a dynamical system through an equation of the form $\dot{x}/dt = \mathbf{F}(x)$, where **F** is a vector function.

References

Ballico, M.J., Sawley, M.L., and Skiff, F. 1990. The bipolar motor: A simple demonstration of deterministic chaos. *Am. J. Phys.*, 58, 58–61.

Holmes, P. 1990. Poincaré, celestial mechanics, dynamical-systems theory and "chaos." *Phys. Reports*, 193, 137–163.

Kautz, R.L. and Huggard, B.M. 1994. Chaos at the amusement park: dynamics of the Tilt-A-Whirl. *Am. J. Phys.*, 62, 69–66.

Li, T.Y. and Yorke, J.A. 1978. Ergodic maps on [0,1] and nonlinear pseudo-random number generators. *Nonlinear Anal. Theory Methods Appl.*, 2, 473–481.

Lorenz, E.N. 1963. Deterministic nonperiodic flow. *J. Atmos. Sci.*, 20, 130–141.

Lorenz, E.N. 1993. *The Essence of Chaos*, University of Washington Press, Seattle, WA.

Martien, P., Pope, S.C., Scott, P.L., and Shaw, R.S. 1985. The chaotic behavior of a dripping faucet. *Phys. Lett. A.*, 110, 399–404.

Ott, E., Grebogi, C., and Yorke, J.A. 1990. Controlling chaos. *Phys. Rev. Lett.*, 64, 1196–1199.

Singer, J., Wang, Y.Z., and Bau, H.H. 1991. Controlling a chaotic system. *Phys. Rev. Lett.*, 66, 1123–1125.

Tanabe, Y. and Kaneko, K. 1994. Behavior of falling paper. *Phys. Rev. Lett.*, 73, 1372–1375.

Wisdom, J., Peale, S.J., and Mignard, F. 1984. The chaotic rotation of Hyperion. *Icarus*, 58, 137–152.

For Further Information

A good introduction to deterministic chaos for undergraduates is provided by *Chaotic and Fractal Dynamics: An Introduction for Applied Scientists and Engineers* by Francis C. Moon. This book presents numerous examples drawn from mechanical and electrical engineering.

Chaos in Dynamical Systems by Edward Ott provides a more rigorous introduction to chaotic dynamics at the graduate level.

Practical methods for experimental analysis and control of chaotic systems are presented in *Coping with Chaos: Analysis of Chaotic Data and the Exploitation of Chaotic Systems*, a reprint volume edited by Edward Ott, Tim Sauer, and James A. York.

19.15 Fuzzy Sets and Fuzzy Logic

Dan M. Frangopol

Introduction

In the sixties, Zaheh (1965) introduced the concept of fuzzy sets. Since its inception more than 30 years ago, the theory and methods of fuzzy sets have developed considerably. The demands for treating situations in engineering, social sciences, and medicine, among other applications that are complex and not crisp have been strong driving forces behind these developments.

The concept of the fuzzy set is a generalization of the concept of the ordinary (or crisp) set. It introduces vagueness by eliminating the clear boundary, defined by the ordinary set theory, between full nonmembers (i.e., grade of membership equals zero) and full members (i.e., grade of membership equals one). According to Zaheh (1965) a fuzzy set A, defined as a collection of elements (also called objects) $x \in X$, where X denotes the universal set (also called universe of discourse) and the symbol $\in$ denotes that the element x is a member of X, is characterized by a membership (also called characteristic) function $\mu_A(x)$ which associates each point in X a real member in the unit interval [0,1]. The value of $\mu_A(x)$ at x represents the grade of membership of x in A. Larger values of $\mu_A(x)$ denote higher grades of membership of x in A. For example, a fuzzy set representing the concept of control might assign a degree of membership of 0.0 for no control, 0.1 for weak control, 0.5 for moderate control, 0.9 for strong control, and 1.0 for full control. From this example, it is clear that the two-valued crisp set [i.e., no control (grade of membership 0.0) and full control (grade of membership 1.0)] is a particular case of the general multivalued fuzzy set A in which $\mu_A(x)$ takes its values in the interval [0,1].

Problems in engineering could be very complex and involve various concepts of uncertainty. The use of fuzzy sets in engineering has been quite extensive during this decade. The area of fuzzy control is one of the most developed applications of fuzzy set theory in engineering (Klir and Folger, 1988). Fuzzy controllers have been created for the control of robots, aircraft autopilots, and industrial processes, among others. In Japan, for example, so-called "fuzzy electric appliances," have gained great success from both technological and commercial points of view (Furuta, 1995). Efforts are underway to develop and introduce fuzzy sets as a technical basis for solving various real-world engineering problems in which the underlying information is complex and imprecise. In order to achieve this, a mathematical background in the theory of fuzzy sets is necessary. A brief summary of the fundamental mathematical aspects of the theory of fuzzy sets is presented herein.

Fundamental Notions

A fuzzy set A is represented by all its elements x_i and associated grades of membership $\mu_A(x_i)$ (Klir and Folger, 1988).

$$A = \left\{\mu_A(x_1)|x_1, \mu_A(x_2)|x_2, \ldots, \mu_A(x_n)|x_n\right\} \tag{19.15.1}$$

where x_i is an element of the fuzzy set, $\mu_A(x_i)$ is its grade of membership in A, and the vertical bar is employed to link the element with their grades of membership in A. Equation (19.15.1) shows a discrete form of a fuzzy set. For a continuous fuzzy set, the membership function $\mu_A(x)$ is a continuous function of x.

Figure 19.15.1 illustrates a discrete and a continuous fuzzy set. The larger membership grade *max* $(\mu_A(x_i))$ represents the height of a fuzzy set.

If at least one element of the fuzzy set has a membership grade of 1.0, the fuzzy set is called normalized. Figure 19.15.2 illustrates both a nonnormalized and a normalized fuzzy set.

The following properties of fuzzy sets, which are obvious extensions of the corresponding definitions for ordinary (crisp) sets, are defined herein according to Zaheh (1965) and Klir and Folger (1988).

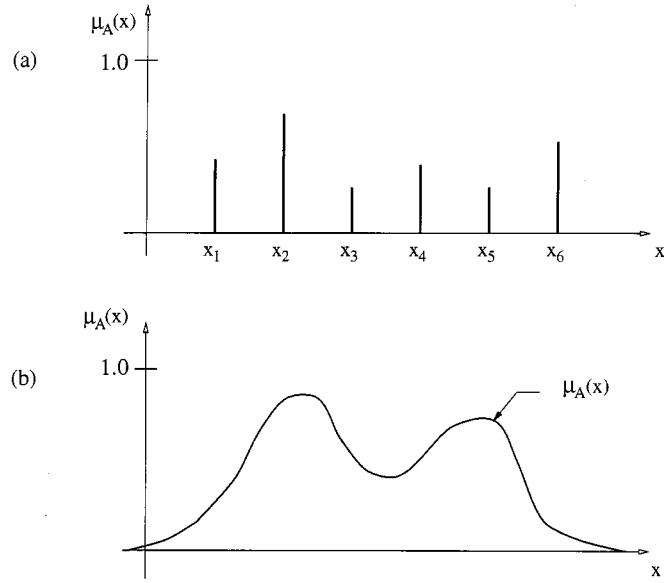


FIGURE 19.15.1 (a) Discrete and (b) continuous fuzzy set.

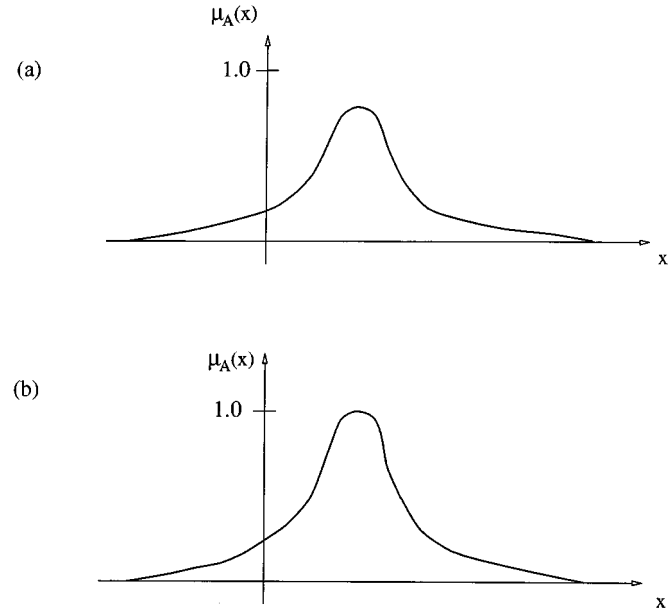


FIGURE 19.15.2 (a) Nonnormalized and (b) normalized fuzzy set.

Two fuzzy sets A and B are equal, A = B, if and only if $\mu_A(x) = \mu_B(x)$ for every element x in X (see Figure 19.15.3).

The complement of a fuzzy set A is a fuzzy set $\overline{A}$ defined as

$$\mu_{\overline{A}}(x) = 1 - \mu_A(x) \tag{19.15.2}$$

Figure 19.15.4 shows both discrete and continuous fuzzy sets and their complements.

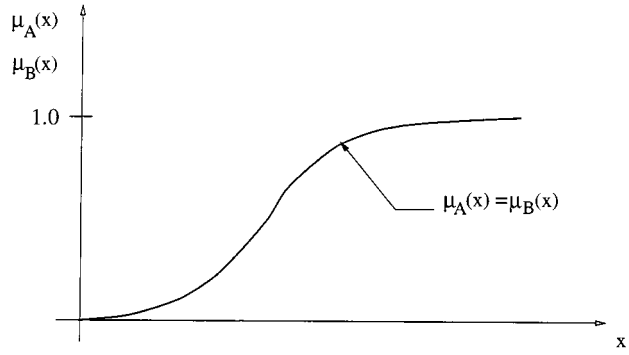


FIGURE 19.15.3 Two equal fuzzy sets, A = B.

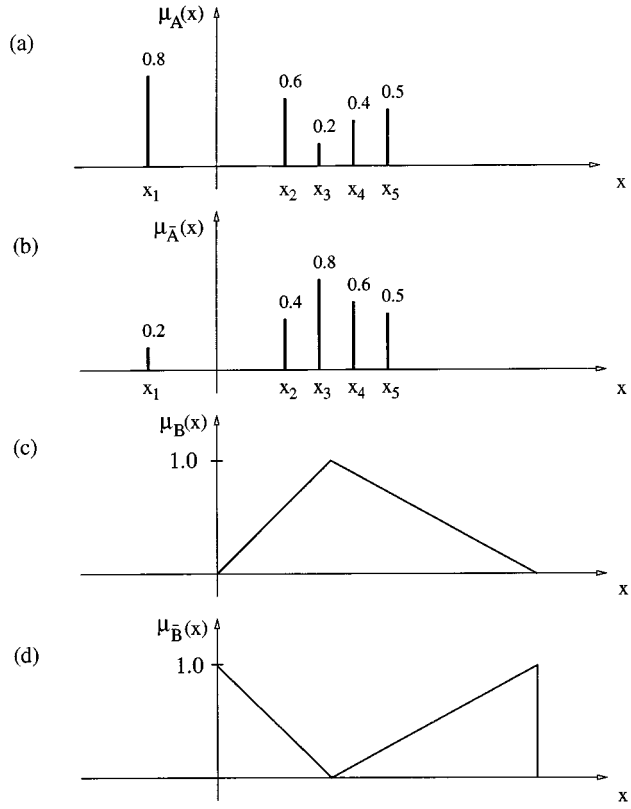


FIGURE 19.15.4 (a) Discrete fuzzy set A, (b) complement $\bar{A}$ of fuzzy set A, (c) continuous fuzzy set B, and (d) complement $\bar{B}$ of fuzzy set B.

If the membership grade of each element of the universal set X in fuzzy set B is less than or equal to its membership grade in fuzzy set A, then B is called a subset of A. This is denoted $B \subseteq A$. Figure 19.15.5 illustrates this situation.

The union of two fuzzy sets A and B with membership functions $\mu_A(x)$ and $\mu_B(x)$ is a fuzzy set $C = A \cup B$ such that

$$\mu_C(x) = \max\left[\mu_A(x), \mu_B(x)\right] \tag{19.15.3}$$

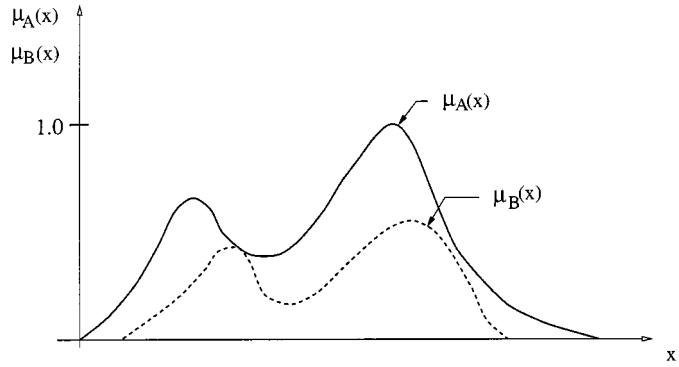


FIGURE 19.15.5 Fuzzy set A and its subset B.

for all x in X.

Conversely, the intersection of two fuzzy sets A and B with membership functions $\mu_A(x)$ and $\mu_B(x)$, respectively, is a fuzzy set $C = A \cap B$ such that

$$\mu_C(x) = \min\left[\mu_A(x), \mu_B(x)\right] \tag{19.15.4}$$

for all x in X.

Figure 19.15.6 illustrates two fuzzy sets A and B, the union set $A \cup B$ and the intersection set $A \cap B$.

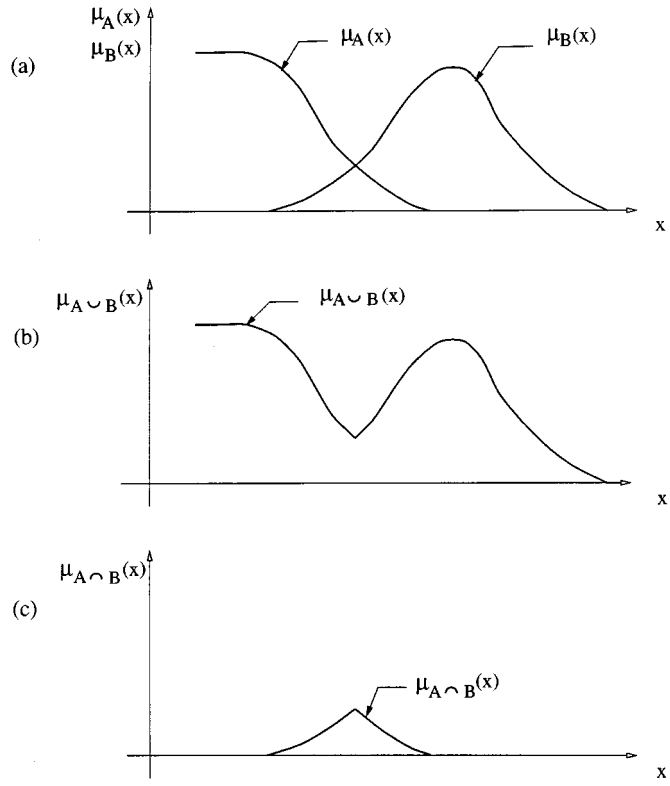


FIGURE 19.15.6 (a) Two fuzzy sets, (b) union of fuzzy sets $A \cup B$, and (c) intersection of fuzzy sets $A \cap B$.

An empty fuzzy set A is a fuzzy set with a membership function $\mu_A(x) = 0$ for all elements x in X (see Figure 19.15.7).

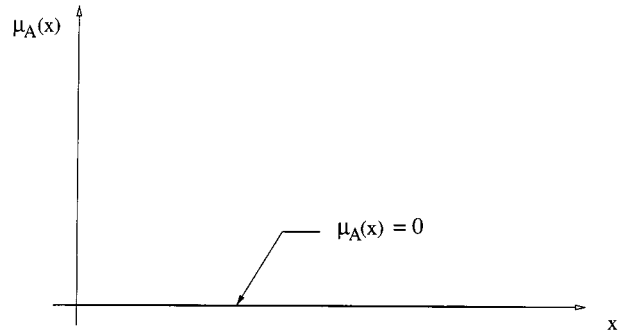


FIGURE 19.15.7 Empty fuzzy set.

Two fuzzy sets A and B with respective membership function $\mu_A(x)$ and $\mu_B(x)$ are disjoint if their intersection is empty (see Figure 19.15.8).

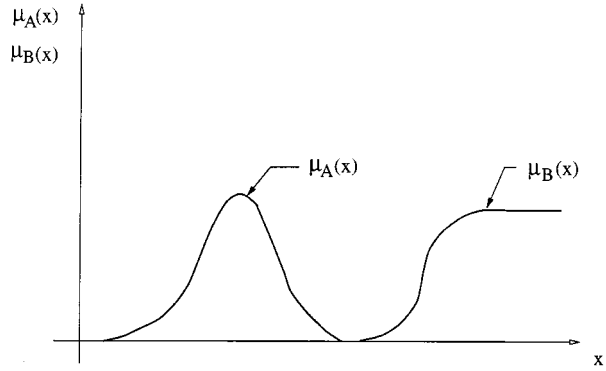


FIGURE 19.15.8 Disjoint fuzzy sets.

An α-cut of a fuzzy set A is an ordinary (crisp) set A_α containing all elements that have a membership grade in A greater or equal to α. Therefore,

$$A_\alpha = \{x | \mu_A(x) \geq \alpha\} \tag{19.15.5}$$

From Figure 19.15.9, it is clear that $\alpha = 0.5$, the α-cut of the fuzzy set A is the crisp set $A_{0.5} = \{x_5, x_6, x_7, x_8\}$ and for $\alpha = 0.8$, the α-cut of the fuzzy set A is the crisp set $A_{0.8} = \{x_7, x_8\}$.

A fuzzy set is convex if and only if all of its α-cuts are convex for all α in the interval [0,1]. Figure 19.15.10 shows both a convex and a nonconvex fuzzy set.

A fuzzy number $\tilde{N}$ is a normalized and convex fuzzy set of the real line whose membership function is piecewise continuous and for which it exists exactly one element with $\mu_{\tilde{N}}(x_0) = 1$. As an example, the real numbers close to 50 are shown by four membership functions in Figure 19.15.11.

The scalar cardinality of a fuzzy set A is the summation of membership grades of all elements of X in A. Therefore,

$$|A| = \sum_x \mu_A(x) \tag{19.15.6}$$

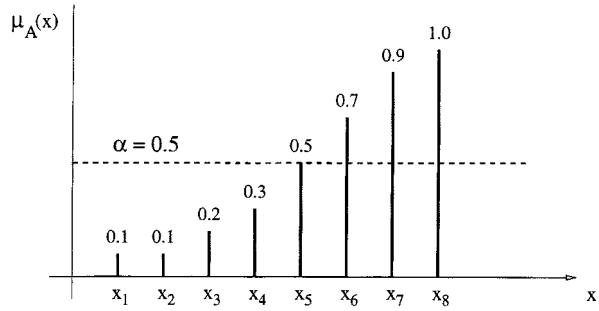


FIGURE 19.15.9 α-cut of a fuzzy set.

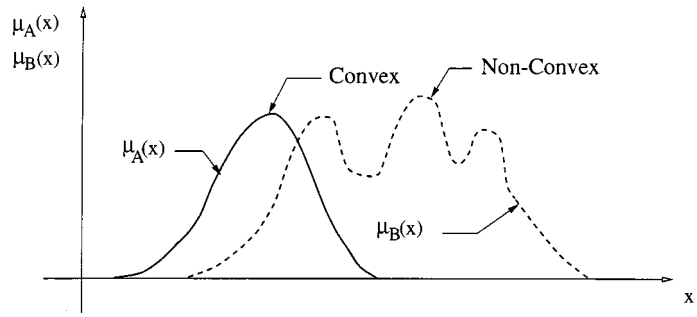


FIGURE 19.15.10 Convex and non-convex fuzzy set.

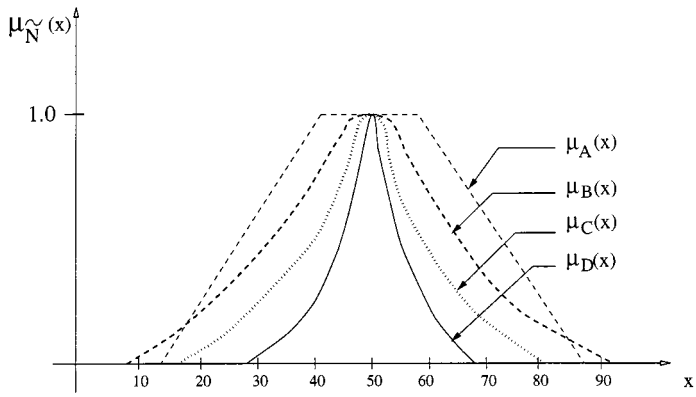


FIGURE 19.15.11 Membership functions of fuzzy sets of real numbers close to 50.

For example, the scalar cardinality of the fuzzy set A in Figure 19.15.4(a) is 2.5. Obviously, an empty fuzzy set has a scalar cardinality equal to zero. Also, the scalar cardinality of the fuzzy complement set is equal to scalar cardinality of the original set. Therefore,

$$|A| = |\overline{A}| \tag{19.15.7}$$

One of the basic concepts of fuzzy set theory is the extension principle. According to this principle (Dubois and Prade, 1980), given (a) a function f mapping points in the ordinary set X to points in the ordinary set Y, and (b) any fuzzy set A defined on X,

$$A = \left\{\mu_A(x_1)|x_1, \mu_A(x_2)|x_2, \ldots, \mu_A(x_n)|x_n\right\}$$

then the fuzzy set B = f(A) is given as

$$B = f(A) = \left\{\mu_A(x_1)|f(x_1), \mu_A(x_2)|f(x_2), \ldots, \mu_A(x_n)|f(x_n)\right\} \tag{19.15.8}$$

If more than one element of the ordinary set X is mapped by *f* to the same element y in Y, then the maximum of the membership grades in the fuzzy set A is considered as the membership grade of y in f(A).

As an example, consider the fuzzy set in Figure 19.15.4(a), where $x_1 = -2$, $x_2 = 2$, $x_3 = 3$, $x_4 = 4$, and $x_5 = 5$. Therefore, A = {0.8|–2, 0.6|2, 0.2|3, 0.4|4, 0.5|5} and $f(x) = x^4$. By using the extension principle, we obtain

$$f(A) = \left\{\max(0.8, 0.6)|2^4, 0.2|3^4, 0.4|4^4, 0.5|5^4\right\}$$

$$= \left\{0.8|16, 0.2|81, 0.4|256, 0.5|625\right\}$$

As shown by Klir and Folger (1988), degrees of association can be represented by membership grades in a fuzzy relation. Such a relation can be considered a general case for a crisp relation.

Let P be a binary fuzzy relation between the two crisp sets X = {4, 8, 11} and Y = {4, 7} that represents the relational concept “very close.” This relation can be expressed as:

$$P(X,Y) = \left\{1|(4,4), 0.7|(4,7), 0.6|(8,4), 0.9|(8,7), 0.3|(11,4), 0.6|(11,7)\right\}$$

or it can be represented by the two dimensional membership matrix

$$\begin{array}{c} \\ x_1 \\ x_2 \\ x_3 \end{array} \begin{array}{c} \begin{array}{cc} y_1 & y_2 \end{array} \\ \begin{bmatrix} 1.0 & 0.7 \\ 0.6 & 0.9 \\ 0.3 & 0.6 \end{bmatrix} \end{array}$$

Fuzzy relations, especially binary relations, are important for many engineering applications.

The concepts of domain, range, and the inverse of a binary fuzzy relation are clearly defined in Zadeh (1971), and Klir and Folger (1988).

The max-min composition operation for fuzzy relations is as follows (Zadeh, 1991; Klir and Folger, 1988):

$$\mu_{P\circ Q}(x,z) = \max_{y\in Y} \min\left[\mu_P(x,y), \mu_Q(y,z)\right] \tag{19.15.9}$$

for all x in X, y in Y, and z in Z, where the composition of the two binary relations P(X,Y) and Q(Y,Z) is defined as follows:

$$R(X,Z) = P(X,Y) \circ Q(Y,Z) \tag{19.15.10}$$

As an example, consider the two binary relations

$$P(X,Y) = \left\{1.0|(4,4), 0.7|(4,7), 0.6|(8,4), 0.9|(8,7), 0.3|(11,4), 0.6|(11,7)\right\}$$

$$Q(Y,Z) = \{0.8|(4,6),\ 0.5|(4,9),\ 0.2|(4,12),\ 0.0|(4,15),\ 0.9|(7,6),\ 0.8|(7,9),\ 0.5|(7,12),\ 0.2|(7,15)\}$$

The following matrix equations illustrate the max-min composition for these binary relations

$$\begin{bmatrix} 1.0 & 0.7 \\ 0.6 & 0.9 \\ 0.3 & 0.6 \end{bmatrix} \circ \begin{bmatrix} 0.8 & 0.5 & 0.2 & 0.0 \\ 0.9 & 0.8 & 0.5 & 0.2 \end{bmatrix} = \begin{bmatrix} 0.8 & 0.7 & 0.5 & 0.2 \\ 0.9 & 0.8 & 0.5 & 0.2 \\ 0.6 & 0.6 & 0.5 & 0.2 \end{bmatrix}$$

Zadeh (1971) and Klir and Folger (1988), define also an alternative form of operation on fuzzy relations, called max-product composition. It is denoted as $P(X,Y) \otimes Q(Y,Z)$ and is defined by

$$\mu_{P\otimes Q}(x,z) = \max_{y\in Y}\left[\mu_P(x,y), \mu_Q(y,z)\right] \tag{19.15.11}$$

for all x in X, y in Y, and z in Z. The matrix equation

$$\begin{bmatrix} 1.0 & 0.7 \\ 0.6 & 0.9 \\ 0.3 & 0.6 \end{bmatrix} \times \begin{bmatrix} 0.8 & 0.5 & 0.2 & 0.0 \\ 0.9 & 0.8 & 0.5 & 0.2 \end{bmatrix} = \begin{bmatrix} 0.8 & 0.7 & 0.5 & 0.2 \\ 0.9 & 0.8 & 0.5 & 0.2 \\ 0.6 & 0.6 & 0.5 & 0.2 \end{bmatrix}$$

illustrates the max product composition for the pair of binary relations P(X,Y) and Q(Y,Z) previously considered.

A crisp binary relation among the elements of a single set can be denoted by R(X,X). If this relation is reflexive, symmetric, and transistive, it is called an equivalence relation (Klir and Folger, 1988).

A fuzzy binary relation S that is reflexive

$$\mu_S(x,x) = 1 \tag{19.15.12}$$

symmetric

$$\mu_S(x,y) = \mu_S(y,x) \tag{19.15.13}$$

and transitive

$$\mu_S(x,z) = \max_y \min\left[\mu_S(x,y), \mu_S(y,z)\right] \tag{19.15.14}$$

is called a similarity relation (Zadeh, 1971). Equations (19.15.12), (19.15.13), and (19.15.14) are valid for all x,y,z in the domain of S. A similarity relation is a generalization of the notion of equivalence relation.

Fuzzy orderings play a very important role in decision-making in a fuzzy environment. Zadeh (1971) defines fuzzy ordering as a fuzzy relation which is transitive. Fuzzy partial ordering, fuzzy linear ordering, fuzzy preordering, and fuzzy weak ordering are also mathematically defined by Zaheh (1971) and Zimmermann (1991).

The notion of fuzzy relation equation, proposed by Sanchez (1976), is an important notion with various applications. In the context of the max-min composition of two binary relations P(X,Y) and Q(Y,Z), the fuzzy relation equation is as follows

$$\mathbf{P} \circ \mathbf{Q} = \mathbf{R} \tag{19.15.15}$$

where **P** and **Q** are matrices of membership functions $\mu_P(x,y)$ and $\mu_Q(y,z)$, respectively, and **R** is a matrix whose elements are determined from Equation (19.15.9). The solution in this case in unique. However, when **R** and one of the matrices **P**, **Q** are given, the solution is neither guaranteed to exit nor to be unique (Klir and Folger, 1988).

Another important notion is the notion of fuzzy measure. It was introduced by Sugeno (1977). A fuzzy measure is defined by a function which assigns to each crisp subset of X a number in the unit interval [0,1]. This member represents the ambiguity associated with our belief that the crisp subset of X belongs to the subset A. For instance, suppose we are trying to diagnose a mechanical system with a failed component. In other terms, we are trying to assess whether this system belongs to the set of systems with, say, safety problems with regard to failure, serviceability problems with respect to deflections, and serviceability problems with respect to vibrations. Therefore, we might assign a low value, say 0.2 to failure problems, 0.3 to deflection problems, and 0.8 to vibration problems. The collection of these values constitutes a fuzzy measure of the state of the system.

Other measures including plausibility, belief, probability, and possibility measures are also used for defining the ambiguity associated with several crisp defined alternatives. For an excellent treatment of these measures and of the relationship among classes of fuzzy measures see Klir and Folger (1988).

Measures of fuzziness are used to indicate the degree of fuzziness of a fuzzy set (Zimmermann, 1991). One of the most used measures of fuzziness is the entropy. This measure is defined (Zimmermann, 1991) as

$$d(A) = h\sum_{i=1}^{n} S\left(\mu_A(x_i)\right) \tag{19.15.16}$$

where h is a positive constant and $S(\alpha)$ is the Shannon function defined as
$S(\alpha) = -\alpha \ln\alpha - (1-\alpha)\ln(1-\alpha)$ for rational α. For the fuzzy set in Figure 19.15.4(a), defined as

$$A = \{0.8|-2,\ 0.6|2,\ 0.2|3,\ 0.4|4,\ 0.5|5\}$$

the entropy is

$$\begin{aligned} d(A) &= h(0.5004 + 0.6730 + 0.5004 + 0.6730 + 0.6931) \\ &= 3.0399\ h \end{aligned}$$

Therefore, for h = 1, the entropy of the fuzzy set A is 3.0399.

The notion of linguistic variable, introduced by Zadeh (1973), is a fundamental notion in the development of fuzzy logic and approximate reasoning. According to Zadeh (1973), linguistic variables are "variables whose values are not members but words or sentences in a natural or artificial language. The motivation for the use of words or sentences rather than numbers is that linguistic characterizations are, in general, less specific than numerical ones." The main differences between fuzzy logic and classical two-valued (e.g., true or false) or multivalued (e.g., true, false, and indeterminate) logic are that (a) fuzzy logic can deal with fuzzy quantities (e.g., most, few, quite a few, many, almost all) which are in general represented by fuzzy numbers (see Figure 19.15.11), fuzzy predicates (e.g., expensive, rare), and fuzzy modifiers (e.g., extremely, unlikely), and (b) the notions of truth and false are both allowed to be fuzzy using fuzzy true/false values (e.g., very true, mostly false). As Klir and Folger (1988) stated, the ultimate goal of fuzzy logic is to provide foundations for approximate reasoning. For a general background on fuzzy logic and approximate reasoning and their applications to expert systems, the reader is referred to Zadeh (1973, 1987), Kaufmann (1975), Negoita (1985), and Zimmermann (1991), among others.

Decision making in a fuzzy environment is an area of continuous growth in engineering and other fields such as economics and medicine. Bellman and Zadeh (1970) define this process as a "decision

process in which the goals and/or the constraints, but not necessarily the system under control, are fuzzy in nature."

According to Bellman and Zadeh (1970), a fuzzy goal G associated with a given set of alternatives X = {x} is identified with a given fuzzy set G in X. For example, the goal associated with the statement "x should be in the vicinity of 50" might be represented by a fuzzy set whose membership function is equal to one of the four membership functions shown in Figure 19.15.11. Similarly, a fuzzy constraint C in X is also a fuzzy set in X, such as "x should be substantially larger than 20."

Bellman and Zadeh (1970) define a fuzzy decision D as the confluence of goals and constraints, assuming, of course, that the goals and constraints conflict with one another. Situations in which the goals and constraints are fuzzy sets in different spaces, multistage decision processes, stochastic systems with implicitly defined termination time, and their associated optimal policies are also studied in Bellman and Zadeh (1970).

References

Bellman, R.E. and Zadeh, L.A. 1970. Decision-making in a fuzzy environment. *Management Science*, 17(4), 141–164.

Dubois, D. and Prade, H. 1980. *Fuzzy Sets and Systems: Theory and Applications.* Academic Press, New York.

Furuta, H. 1995. Fuzzy logic and its contribution to reliability analysis. In *Reliability and Optimization of Structural Systems*, R. Rackwitz, G. Augusti, and A. Borri, Eds., Chapman & Hall, London, pp. 61–76.

Kaufmann, A. 1975. *Introduction to the Theory of Fuzzy Subsets*, Vol. 1, Academic Press, New York.

Klir, G.J. and Folger, T.A. 1988. *Fuzzy Sets, Uncertainty, and Information*, Prentice Hall, Englewood Cliffs, New Jersey.

Negoita, C.V. 1985. *Expert Systems and Fuzzy Systems*, Benjamin/Cummings, Menlo Park, California.

Sanchez, E. 1976. Resolution of composite fuzzy relation equations. *Information and Control*, 30, 38–48.

Sugeno, M. 1977. Fuzzy measures and fuzzy integrals — a survey, in *Fuzzy Automata and Decision Processes*, M.M. Gupta, R.K. Ragade, and R.R. Yager, Eds., North Holland, New York, pp. 89–102.

Zadeh, L.A. 1965. Fuzzy sets. *Information and Control*, 8, 338–353.

Zadeh, L.A. 1971. Similarity relations and fuzzy orderings. *Information Sciences.* 3, 177–200.

Zadeh, L.A. 1973. The concept of a linguistic variable and its applications to approximate reasoning. Memorandum ERL-M 411, Berkeley, California.

Zadeh, L.A. 1987. *Fuzzy Sets and Applications: Selected Papers by L.A. Zadeh*, R.R. Yager, S. Ovchinnikov, R.M. Tong, and H.T. Nguyen, Eds., John Wiley & Sons, New York.

Zimmerman, H.-J. 1991. *Fuzzy Set Theory – and Its Applications*, 2nd ed., Kluwer Academic Publishers, Boston.

Further Information

The more than 5000 publications that exist in the field of fuzzy sets are widely scattered in many books, journals, and conference proceedings. For newcomers, good introductions to the theory and applications of fuzzy sets are presented in (a) *Introduction to the Theory of Fuzzy Sets*, Volume I, Academic Press, New York, 1975, by Arnold Kaufmann; (b) *Fuzzy Sets and Systems: Theory and Applications*, Academic Press, New York, 1980, by Didier Dubois and Henri Prade; (c) *Fuzzy Sets, Uncertainty and Information*, Prentice Hall, Englewood Cliffs, NJ, 1988, by George Klir and Tina Folger, and (d) *Fuzzy Set Theory and Its Applications*,, 2nd ed., Kluwer Academic Publishers, Boston, 1991, by H.-J. Zimmerman, among others.

The eighteen selected papers by Lotfi A. Zadeh grouped in *Fuzzy Sets and Applications*, John Wiley & Sons, New York, 1987, edited by R. Yager, S. Ovchinnikov, R.M. Tong, and H.T. Nguyen are particularly helpful for understanding the developments of issues in fuzzy set and possibility theory. Also, the interview with Professor Zadeh published in this book illustrates the basic philosophy of the founder of fuzzy set theory.

Kreith, F.; et. al. “Patent Law and Miscellaneous Topics”
Mechanical Engineering Handbook
Ed. Frank Kreith
Boca Raton: CRC Press LLC, 1999

Patent Law and Miscellaneous Topics

Frank Kreith
University of Colorado

Thomas H. Young
Dorsey & Whitney LLP

George A. Peters
Peters & Peters

Jeff R. Crandall
University of Virginia

Gregory W. Hall
University of Virginia

Walter D. Pilkey
University of Virginia

Michael Merker
American Society of Mechanical Engineers

Roland Winston
University of Chicago

Walter T. Welford (deceased)
Imperial College of London

Noam Lior
University of Pennsylvania

Malcolm J. Crocker
Auburn University

Barbara Atkinson
Lawrence Berkeley National Laboratory

Andrea Denver
Lawrence Berkeley National Laboratory

James E. McMahon
Lawrence Berkeley National Laboratory

Leslie Shown
Lawrence Berkeley National Laboratory

Robert Clear
Lawrence Berkeley National Laboratory

Craig B. Smith
Daniel, Mann, Johnson, & Mendenhall

20.1 Patents and Other Intellectual Property

Thomas H. Young[*]

The purpose of this chapter is to provide some very general information about intellectual property protection (especially patents) to nonlawyers.[**] It is also intended to provide some suggestions for "self-help" to engineers that will enable them to improve the possibility of obtaining and securing appropriate protection for their ideas and developments even before they consult a lawyer. One of the questions an intellectual property attorney frequently hears is: "I have a great idea. How do I protect it?" This section should provide at least a starting point for answering that question.

In the case of patents, this section is designed to assist engineers and/or inventors in understanding what is expected of them and how to assist their patent lawyers or agents in the process of protecting their inventions. Be forewarned, however; the process of obtaining patent, trade secret, and/or copyright protection on such a development is not necessarily a linear one and is often neither short nor easy.

Patents[***]

What is a patent? The authorization for a patent system stems from the United Sates Constitution, which provides:

> The Congress shall have power… to promote the progress of science and useful arts, by securing for limited times to authors and inventors the exclusive right to their resective writings and discoveries. U.S. Const. Art. I, § 8.

Although the concept of a patent system did not originate in the United Sates, the drafters of the Constitution thought patents and copyrights significant enough to provide for them in the fundamental predicates on which our government is founded. Indeed, Thomas Jefferson, as Secretary of State, was one of the first patent "examiners."

The premise of the Constitutional provision is to encourage the disclosure of inventions in the interest of promoting future innovation. Specifically, if inventors were not provided with patent protection for their innovations, they would seek to exploit them in secret. Thus, no one else would be able to benefit from and expand on their inventions, so technology would not move as rapidly as it would with full disclosure. Following enactment of the Constitution, Congress promptly adopted and has continuously maintained and updated laws implementing a patent system.

---

[*] Mr. Young has practiced intellectual property law for more than 20 years, is a partner in the Denver office of the international law firm of Dorsey & Whitney, and has been an adjunct professor of patent and trademark law at the University of Colorado law school. The author acknowledges with great appreciation the assistance of Mr. Gregory D. Leibold, an attorney focusing on intellectual property law at Dorsey & Whitney in the preparation of this article.

[**] This section is in no way intended to provide the reader with all of the information he or she may need to evaluate or obtain intellectual property protection in particular situations on their own. Additional information is available from the Patent and Trademark Office and the Copyright Office in Washington, D.C. and other published references on these subjects. For information on patents and trademarks, the reader may contact the Commissioner of Patents and Trademarks, Washington, D.C. 20231, (703)308-3457. For information regarding copyrights, contact the Register of Copyrights, Library of Congress, Washington, D.C. 20559, (202)707-3000. When applying the legal requirements to a specific issue, the reader is encouraged to consult with a knowledgeable lawyer.

[***] The patent laws are codified in 35 United States Code (i.e., "U.S.C.") § 100 *et seq*. Regulations implementing these patent laws, particularly as they relate to the operation of the Patent Office, are found in volume 37 of the Code of Federal Regulations (i.e., "C.F.R."). The internal rules of the Patent Office relating to the examination of patents are contained in the Patent Office "Manual of Patent Examining Procedure" or "M.P.E.P." For more information on patents, the following treatises may be helpful: *Patents*, Donald S. Chisum, Matthew Bender & Co., New York, 1995; *Patent Law Fundamentals*, Peter D. Rosenberg, 2nd ed., Clark Boardman Callaghan, Rochester, NY, 1995.

Strictly speaking, patents are not contracts; however, because of the premise behind the Constitutional provision, patents have sometimes been analogized to contracts between inventors and the U.S. Government. The Government agrees to allow the inventor exclusive rights to his or her invention for a period of time. In exchange, the inventor agrees to disclose his or her invention, thereby allowing other inventors the benefit of the patentee's work and enabling others to practice the invention after the patent has expired.

What rights does a patent confer? After a patent issues, the patentee has the right to exclude others from making, using, or selling the subject matter claimed in the patent. Currently, the term of a U.S. patent is 20 years from the date the patent was filed. The patentee is not required to license others to use the invention.

Nevertheless, a patent does not confer the right to do anything. In fact, it is frequently the case that an inventor will receive a patent but will be unable to use the patented invention because it infringes on another's patent. For example, if A obtained the first patent on a laser and B later obtains a patent on the use of a laser in surgery, B cannot make or use the laser in surgery, because it would infringe A's patent. By the same token, however, A cannot use this laser in surgery, because it would infringe B's patent. In addition, patent rights are territorially limited, and the use of the same invention in different countries may have different ramifications. Finally, regardless of patent rights, the use of certain technologies (e.g., pharmaceuticals) may be limited by other government regulations and practical considerations.

A significant advantage of a patent over other forms of intellectual property protection, such as copyright or trade secret, is that a patent can be enforced against anyone who utilizes the claimed technology regardless of whether that person copied or misappropriated the technology. Independent development of an infringing device is not a defense.

Who can obtain a patent? Anyone who is the first to invent something that falls into a category of subject matter that is deemed to be patentable may obtain a U.S. patent. A single patent can have several inventors if the subject matter of one or more claims in the patent was jointly conceived or reduced to practice. In that case, each inventor, in essence, obtains his or her own rights to the patent, subject to an obligation to account to the other inventor(s) for their respective share of the remuneration. The patent must be obtained by the true inventors or it may be invalid. Thus, persons who did not contribute to an invention should not be named as inventors regardless of the desirability of recognizing their moral or economic support to the process of inventing.

What subject matter is patentable? U.S. patent law provides that:

> Whoever invents or discovers any new and useful *process, machine, manufacture, or composition of matter*, or nay new and useful *improvement* thereof, may obtain a patent therefore, subject to the conditions of patentability and requirements of this title. [emphasis added.] 35 U.S.C. § 101.

Although the categories of patentable items seem archaic, Congress acknowledged that it intended this language to "include anything under the sun that is made by man."

The development of significant new technologies, such as computers and biotechnology, have challenged the limits of proper patentable subject matter. Nevertheless, with some diversions along the way, courts have continued to expand those boundaries so that they now include computer software and engineered life forms when embodied in properly drafted patent claims. In general, however, laws of nature, scientific truths, and mathematical algorithms are not patentable subject matter in and of themselves. New and useful applications of those concepts, however, may be patented.

What are the standards for patentability? In order to merit a patent, an invention must be new, useful, and nonobvious. The "new" and "useful" standards are defined precisely as one might expect. One may not obtain a patent on an invention that has been invented before and, therefore, is not novel or on one that has no practical utility.

The "nonobvious" standard is a more difficult one. Even if an invention has not been created before, in order to be patentable, it must not have been obvious to one of ordinary skill in the technical field of the invention at the time the invention was made. In other words, a valid patent cannot be obtained on

an invention that merely embodies a routine design or the application of principles within the ordinary skill of the art. Whether an invention meets the standard for nonobviousness involves a factual inquiry into the state of the art, the level of ordinary skill in the art, and the differences between what was known in the art and the invention at the time it was made. In addition, both the Patent Office and courts will look at objective evidence of nonobviousness, including the context in which the invention was made (e.g., whether or not there was a long-standing problem that had not been solved by others), recognition of the invention, and its success in the marketplace. Nevertheless, the determination of nonobviousness is not precise. Indeed, the "beauty" of many of the most significant inventions is embodied in their simplicity. As such, they are temptingly "obvious" after the fact, even though courts are theoretically constrained from utilizing hindsight in determining nonobviousness.

How is a U.S. patent obtained? The process of obtaining a patent starts with the invention. Invention, under U.S. patent law, has two parts. First, there is the "conception" of the invention, which literally refers to the date when the inventor first thought of the novel aspect of his or her invention. Second, there is the "reduction to practice" of the invention, which can refer to a variety of activities. In the case of a mechanical invention, for example, reduction to practice occurs when a working version of the machine is built embodying the invention. However, there is also a legal concept called "constructive" reduction to practice, which occurs when the inventor files a patent application with the Patent Office.

The date of invention is very important in the United States. Unlike every other country in the world, the United States awards patents to the first person to invent it, rather than the first person to file a patent application. Importantly, the date of conception serves as the date of invention in the United States so long as the inventor was "diligent" in reducing the invention to practice. If the inventor is not continuously "diligent" (and there is some question as to what that word means, exactly), then the date of invention is considered the date on which the inventor was continuously diligent until the invention was reduced to practice.

The date of invention can be critical to obtaining a patent in the United States. In foreign countries, however, the only date that matters is the date on which the inventor first filed the patent application. Filing a patent application in the United States will normally preserve the inventor 's rights abroad if the appropriate foreign patent applications are filed within 1 year of the U.S. filing date and the other requirements of the Patent Cooperation Treaty are met. Ideally, an inventor should file a patent application directly after the conception of the invention in order to achieve the earliest effective date in both the United States and abroad.

Before filing a patent application, an inventor should have a "prior art search" performed. *Prior art* is a term used to refer, in part, to any printed materials published before the inventor files the application that are relevant to the issues of novelty and nonobviousness. Having a search done for prior art is a good way to ensure that the invention for which the application is written is patentable and may also provide some insight into whether or not practice of the invention would infringe the rights of others. The prior art search enables a preliminary determination of the patentability of the invention and, if it appears patentable, the identification of the patentable features to be focused upon in drafting the application. Nevertheless, there is always some prior art that will not be accessible using economically viable means at the time of application, so an inventor can never be completely sure about the novelty of his/her invention at the time of filing.

An inventor can apply for a patent from the Patent Office either *pro se* (i.e., on his own) or through a registered patent agent or attorney. The process of obtaining a patent is a complex one, and the inventor's chances of obtaining a valid patent of the broadest possible scope are greatly increased by the use of a qualified agent or attorney. Lists of registered patent attorneys and agents are available from the Patent Office. Patent attorneys may also be located through the *Martindale-Hubbell Law Directory*, state and local bar associations, and other publications, directories, and professional organizations.

This is not to say that inventors, themselves, have not successfully obtained patents from the Patent Office. Nevertheless, busy patent examiners are easily frustrated by *pro se* applicants' lack of familiarity with patent application requirements and Patent Office rules, and those frustrations are evidenced, consciously or unconsciously, in Patent Office "rejections." Familiarity with the stated requirements of

the Patent Office and knowledge of its informal workings greatly increase the chances of successfully obtaining a valid patent with the broadest possible scope.

A patent application contains a number of parts, all of which generally fall into two main categories. The first group, known as the "specification," contains a detailed written description of the invention including relevant drawings or graphs and any examples. The purpose of the specification is to "enable" one of ordinary skill in the art to make and use the invention. In addition, the applicant must disclose the "best mode" of practicing the invention. The "enablement" and "best mode" requirements are intended to fulfill the constitutional purpose of full disclosure of the invention. Thus, a patent that was otherwise properly granted by the Patent Office may nevertheless be invalidated if it is later determined that the inventor failed to teach others how to utilize the invention or tried to maintain the best mode of practicing the invention as a secret.

The second major part of the patent application is the "claims," which are the separately numbered paragraphs appearing at the end of a patent. The claims define the scope of protection that the patent will confer, and they must be clear, definite, and unambiguous. The goal of the applicant in prosecuting a patent is to obtain claims that cover the invention as broadly as possible without including subject matter that is obvious or not novel. For example, if A invented the two-wheeled bicycle, he might claim a vehicle utilizing two round rotating objects for conveyance. Such a claim might prove to be invalid, for example, if carts were previously known. In such a situation, it would be better to claim the bicycle more narrowly as a vehicle with two wheels in tandem.

Once the application is drafted and submitted to the Patent Office, the Patent Office classifies the invention and sends it to an appropriate "art unit." The Patent Office is divided into separate art units that employ examiners who are knowledgeable about, or at least familiar with, particular fields of technology. Eventually, an examiner will read the application, focusing mainly on the claims since they define the legal parameters of the invention. After performing his own patentability search and reviewing the application for other technical defects, the examiner may either "allow" the application to become a patent or reject the application on the grounds that it lacks substantive merit and/or fails to comply with the formalitites of a proper patent application. If the examiner rejects some of the claims, he will explain why the claims are not patentable in light of the specific prior art references. This will start a series of communications (in writing, by telephone, or in person) between the examiner and the applicant (or the registered attorney of record). During that process, the claims of the patent will be amended as becomes necessary in an attempt to define patentable subject matter.

If the application is eventually allowed, then the patent will be valid from the date of issuance until 20 years after the date the application was originally filed. If the application is not allowed by the examiner and is finally rejected, the inventor can (1) abandon the application and the effort to obtain a patent, (2) file a "continuation" application and start the process over, or (3) appeal the rejection to the Board of Patent Appeals and Interferences and eventually to other courts. It is not uncommon for the process from filing of the application until allowance to take several years.

After the application is submitted, it is appropriate to advise the public that a patent is "pending." This remains true until the application has been abandoned or has issued as a patent. Generally, neither the contents of the application nor even the fact that an application has been filed are publicly available from the Patent Office absent the consent of the applicant. After the patent issues, the patent number should be marked on products embodying the patent. This notice serves to start the period running for the collection of damages for patent infringement. Absent such notice, damages do not begin to run until the infringer receives actual notice of the patent.

Several aspects of this process bear particular note. First, once an application has been filed, the Patent Office does not permit the introduction of "new matter" into the application. Although typographical and other simple errors in the application may be corrected, an applicant may not insert additional written material or drawings (e.g., an additional embodiment or further improvement). The reason for this is simple — because the date of invention and the date of filing are both important for U.S. and foreign priority purposes, the Patent Office could not operate efficiently if the subject matter of the application were constantly amended. There would be numerous continuing disputes as to the effective date to be

afforded the subject matter of an amended application. Instead, if it is necessary to amend the application substantively, a new "continuation-in-part" application must be filed. If it is filed while the original or "parent" application is still pending, the subject matter common to both applications will retain the original filing date, while the new material will only be afforded the filing date of the later continuation-in-part application. The prohibition on adding "new matter," therefore, places a premium on filing an original application that completely and accurately describes the invention.

Second, the patent application process frequently is not linear. Seldom does an invention constitute a static concept or a single finished embodiment. The concept may change as its full ramifications are slowly revealed through practical experience, and its embodiment in physical articles, compositions, or processes may change after the patent application is filed. Thus, as the invention evolves, it may be necessary to file further continuation-in-part applications to cover the developments which appear during this evolutionary process. To achieve the ultimate objective of obtaining patent protection on commercially viable aspects of the invention, it may be necessary for the inventor and legal counsel to reevaluate the merits of the original application and to take appropriate corrective action in the patent prosecution process.

How can the inventor help? There are a variety of ways in which the inventor can and, in fact, should aid in the process of obtaining a patent.

First, and perhaps foremost, an inventor should keep detailed notes on his research and development throughout the inventive process. All research and/or discoveries should be documented, dated, and witnessed by a noninventor because they may become important later for purposes of determining who was the first to invent. The date of invention may also be useful in convincing the Patent Office or a court that certain prior art dated before the filing of the patent application should, nevertheless, not be applied. In preparing this documentation, it is important to realize that the testimony of the inventor himself or documents authenticated only by the inventor are generally not sufficient to prove priority. There are many reasons for this. Suffice it to say that there is a premium on having the records witnessed contemporaneously by at least one noninventor. Getting a noninventor to witness pages of a lab notebook is one way to accomplish this. Alternatively, the Patent Office will, for a nominal fee, accept and provide validation of "disclosure" documents evidencing the date of conception. This is probably the most unquestionable evidence of a date of conception.

An inventor can also assist by performing an informal prior art search of his own before contacting a patent attorney. Filing a patent application is expensive. An inventor should be as confident as possible that he has created something novel before spending the money to draft an application. Typically, a patent attorney will urge that a formal search be performed to ensure that the prior art that the Patent Office examiner is likely to access has been considered. Due to the proliferation of electronic databases for scientific and industry literature, however, there may be additional prior art which is more readily accessible to the inventor. A search of this material can be very helpful to the patent attorney in isolating the patentable features of the invention and improving the chances of obtaining a valid patent.

All prior art located by or known to the inventor should be disclosed to the patent attorney so that it can be evaluated and disclosed to the Patent Office, if necessary. In that regard, both an applicant and his attorney are under an affirmative duty to disclose relevant prior art to the Patent Office, and failure to do so can lead to invalidation of the patent.

Once a prior art search has been performed, an inventor should file an application as early as practically possible. As noted previously, this will protect the inventor's rights in foreign countries as well as provide constructive reduction to practice of the invention if it has not already been embodied in a working prototype. Nevertheless, some delay in filing may be desirable in order to permit further testing and corroboration. This is permissible, and in many instances desirable, bearing in mind that under U.S. patent law an application *must* be filed within 1 year of the first public use, disclosure, sale, or offer of sale of the invention. This deadline cannot be extended and, if violated, will automatically result in the invalidation of the patent. Many other countries do not have a 1 year "grace" period; a patent application is immediately barred once there has been a public disclosure in that country or elsewhere. Although there is an "experimental use" exception in the United States, it is limited to technical experiments (as

opposed to test marketing) that are appropriately documented. To be safe, the inventor should contact a patent attorney before publicly disclosing or using an invention or offering it for sale. If any of those events have already occurred, they should be immediately called to the patent attorney's attention.

The inventor is also instrumental in drafting the patent application. The inventor must make sure that the specification is accurate, enables one skilled in the art to practice the invention, and discloses the best way to make and use the invention. Although the art of claim drafting is somewhat of an acquired skill, the inventor should also be very involved in that process to make sure that the claims cover the invention and are not easily designed around. Sometimes a patent attorney will draft claims that do not expressly state a physical element or process step, but utilize "means-plus-function" language. In that case, it is highly desirable to describe as many means as possible for fulfilling that element in the application itself, and the inventor is the best source of that type of information. In other words, the inventor should continuously be asking questions of his attorney about whether or not certain variations of his invention will be adequately protected by the claims as drafted. Finally, the inventor can and should be instrumental in helping the attorney identify the technical/practical differences between the invention and prior art. In short, the prosecution of a patent application should be a team effort between inventor and attorney at each step in the process.

Trade Secrets*

In some instances it may be desirable to protect new technology as a trade secret rather than by patent. For example, an invention may not be patentable; it may be one of numerous small items or "know how" that improve one's business. Such items are typically protected as trade secrets. While there is no formal method of acquiring trade secret protection as there is for patents and copyrights, attention must still be paid to their identification and protection. Also, unlike patents and copyrights, trade secrets are protected under state, rather than federal laws. Therefore, the nuances of trade secret protection may vary depending on which state's laws apply. There are, however, some general principles of trade secrets that are common to most jurisdictions.

What is a trade secret? A "trade secret" may include anything that gives a person an advantage over his competitors and is not generally known to them. It includes compilations of information such as customer lists, compositions, process techniques and parameters, and software. In determining whether or not something is a protectable trade secret, courts look at the following factors: (1) the extent to which the information is known outside of the trade secret owner's business; (2) the extent to which it is known by employees or others involved in the trade secret owner's business; (3) the extent of measures taken by the trade secret owner to guard the secrecyof the information; (4) the value of the information to the trade secret owner and to his competitors; (5) the amount of effort or money expended by the trade secret owner in developing the information; and (6) the ease or difficulty with which the information could be properly acquired or duplicated by others.

What protection does a trade secret provide? A trade secret protects the owner from improper appropriation of the secret through use or disclosure by a person having an obligation not to do so. The major advantage of a trade secret is that it may last indefinitely; the major defect is that if it is lost, it generally cannot be reclaimed. Unlike patent infringement, a person accused of stealing a trade secret can successfully defend such an accusation by showing that he independently developed the subject matter of the secret. Trade secret protection, therefore, is only as good as the inability of competitors to "reverse engineer" or independently develop it.

How can one obtain a trade secret? Whether or not one has a trade secret is ultimately determined judicially in an action for enforcement. The court will look at all of the factors previously noted. At that time it is usually too late to take the actions necessary to establish trade secret protection. Thus, there is a premium for the periodic review and implementation of a program for trade secret protection. Among

* See, generally, the Uniform Trade Secrets Act § 1(4), 14 U.L.A. 537 (1980) and *Milgrim on Trade Secrets*, Roger M. Milgrim, Matthew Bender & Co., New York, 1995.

the steps that can be taken to protect trade secrets are (1) identifying the types of materials that are deemed to be trade secrets by an organization and notifying employees of the organization's policy to treat this information as trade secrets; (2) restricting access to the trade secrets to only those individuals who have a need to know and who are obligated by contract or otherwise not to use or disclose them; (3) taking physical precautions to limit access to trade secrets such as using locked file cabinets, vaults, etc. These are the most fundamental steps that can be taken.

Although legal counsel is frequently consulted in establishing a plan for protecting trade secrets, including the drafting of appropriate confidentiality agreements, the physical steps required should be put in place by the trade secret owner prior to legal consultation.

Copyrights*

Copyright protection is usually associated with writings, songs, paintings, sculpture, and other artistic endeavors. However, it also extends to certain technology, particularly software, databases, and certain architectural plans.

What protection does a copyright provide? Copyright protection is a matter of federal law. Under the federal Copyright Act, protection is available for "original" works of appropriate subject matter, such as those mentioned in the previous paragraph. A copyright provides the exclusive right to reproduce, publicly distribute, publicly perform, publicly display, and make derivative works from an original work of authorship. However, it protects only against "copying" of the protected work and does not provide protection where a work has been independently developed by another. Copying can be presumed where the third party had access to the copyrighted work and there is "substantial similarity" between the third party's work and the original. Further, in certain limited instances, a small portion of a work may be reproduced without liability as a "fair use." The parameters of the "fair use" doctrine are complicated, however, and they are beyond the subject of this section.

Only a modicum of creativity is needed to satisfy the requirement of "originality." On the other hand, protection is limited. It has frequently been stated that a copyright only protects the manner in which something is expressed, rather than the idea or content, which can be an extremely difficult distinction to make. For example, there is significant debate as to the appropriate scope of copyright protection, for both software and databases. Nevertheless, an advantage of a copyright is that it may last for a relatively long period of time. At a minimum, copyright in works created at this time last for the life of the author plus 50 years.

How is a copyright obtained? Theoretically, in order to obtain a copyright an author need only show that a work is original to him (i.e., that he is in fact the author). Federal protection for copyright attaches as soon as a writing is fixed in a tangible medium (i.e., as soon as it is put on paper, or in memory on a computer, or anywhere else that it can be "perceived"). Nevertheless, there are two important steps that should be taken to realize the full scope of that protection.

First, it is highly advisable to affix a copyright notice to the work. "Notice" is provided by affixing to the work: (1) the word "Copyright" or the "©" symbol, (2) the date of first publication, and (3) the name of the copyright owner. Notice should be affixed to the work in an obvious position (such as one of the first pages of a book, an entry screen of a computer program, etc.). A copyright notice should be placed on all drafts of material that may be subject to protection under copyright law. If an author fails to put the notice on his work, copyright protection is not lost; however, an infringer may then use the defense that he innocently thought that the work was within the public domain. Adequate notice precludes the "innocent" infringer defense, and, because it does not cost anything, notice is always a good idea.

In addition, registration of the copyright with the Register of Copyrights at the Library of Congress is also highly recommended. This procedure is relatively simple and inexpensive. It consists of filling

* The federal laws on copyright are codified in the Lanham Act, § 1 *et seg*. For further information on copyrights, the following treatise may be helpful: *Nimmer on Copyright*, Melville B. & David Nimmer, Matthew Bender & Co., New York, 1995.

out a government form and submitting it with the appropriate filing fee and a sample of the work sought to be copyrighted. If the application is acceptable, the Copyright Office returns a copy of the application stamped with the copyright registration number.

In the case of software and databases, there has been concern that an unscrupulous competitor might attempt to purloin valuable information from the sample submitted with the application. In recognition of this problem, the Copyright Office has promulgated rules allowing the submission of only a limited portion of the software or database adequate to identify the copyrighted work.

Although federal registration is not necessary to perfect a copyright, it does have some significant advantages. First, it is necessary to obtain a registration in order to bring suit against an infringer. While it is possible to register the work on an expedited basis immediately prior to commencing the lawsuit, it is not advisable. Early registration (i.e., within 3 months of the first publication of the work) allows an author to elect to sue for "statutory" damages, which can be considerably easier to prove and, possibly, more munificent than "actual" damages. In addition, an author may be able to recover attorneys' fees from a defendant in some cases, but only if the work in question was registered before or soon after publication. In sum, if an author believes that a work is important and might be copied, he should promptly register the work.

Trademarks*

Although trademark protection does not encompass technology per se, any summary of intellectual property protection would be incomplete without a few comments on the scope and availability of trademark rights.

What is a trademark? A trademark is anything that serves to identify the source of a product or service. Trademark protection can be obtained on a wide variety of items such as words, symbols, logos, slogans, shapes (if they are not functional), tones (such as NBC's three-note tone), and even colors. A trademark owner may assert his rights against a third party that is engaged in conduct that is likely to cause consumers to believe that his goods or services emanate from, are associated or affiliated with, or are sponsored by the trademark owner.

The purpose of trademark protection is twofold: (1) to ensure the public that goods or services offered under a trademark have the quality associated with the trademark owner and (2) to preserve the valuable goodwill that the trademark owner has established by promoting the mark. Trademark rights may last indefinitely.

How are trademark rights established? Trademarks are protected under both federal and state law. There are two ways to obtain rights in a trademark: use or federal registration. Rights in a mark can be acquired simply by using the mark in connection with the goods and services. If the mark is a distinctive one (i.e., coined or "made-up" like "POLAROID") rights are established immediately. On the other hand, if the mark is descriptive, the use must be so prominent and lengthy that the mark has acquired a "secondary meaning," (i.e., the public has come to recognize the term as identifying the source of the goods and services rather than as a reference to the product or service itself). In either case, to establish rights through use, it is desirable to use the designation "™" in connection with the mark, thereby indicating that the owner of the mark asserts that it is a trademark under common law.

In contrast, the process of obtaining a federal trademark registration involves the filing of an application with the U.S. Patent and Trademark Office. The application identifies the mark, the goods and services with which it is used, and, if it is already in use, the date of first use and the date of first use in interstate commerce. A filing fee is also required but is relatively minimal. A trademark examiner will check the application substantively, including the review of previously issued trademarks, to see if there is any conflict. If the examiner allows the application, the mark is published in the *Official Gazette*, which is

* The federal trademark laws are codified at 17 U.S.C. § 1 *et seg*. For further information on trademarks, the following treatise may be helpful: *McCarthy on Trademarks and Unfair Competition*, J. Thomas McCarthy, Clark Boardman Callaghan, Rochester, NY, 3rd ed., 1995.

a weekly publication of the Patent Office. If no one opposes the registration within 30 days after publication, then a registration will be issued. If there is an objection, then an opposition proceeding is initiated, and the applicant and the opposer essentially litigate to determine who should obtain the trademark registration. Provided that certain documents are subsequently filed with the Patent Office to affirm and renew the registration, the registration may last indefinitely. The existence of a federally registered trademark is designated by the symbol "®", which is frequently seen on the shoulder of the mark.

A federal trademark registration is a powerful tool in protecting a trademark. By law it serves as "constructive" (i.e., assumed) notice of the registrant's claim of rights throughout the United States. In essence, the registration acts like a recorded deed to real estate, and anyone who subsequently purchases that property without checking the recorded information does so subject to the interests of record. In contrast, the rights of one who attempts to acquire a trademark only through use are generally limited to the geographic area in which the mark is actually used. Others who subsequently use the mark in other areas of the country may also establish rights. Thus, the first step in selecting a trademark is to search the records of the Patent Office to determine whether or not a confusingly similar mark has previously been registered. Typically, a trademark search performed by a professional search organization will also reveal state registrations and other common law uses that might cause conflicts with the mark under consideration.

One other advantage to federal registration is that the registrant may begin to acquire protection on a mark prior to actual use. This is accomplished by filing an "intent-to-use" application on a mark that the applicant has made a *bona fide* decision to use. The application is examined as described previously. Although the mark must actually be placed in use before the registration is issued, the registration will be effective from the date of its filing. Thus, the public is on notice of the applicant's potential rights as soon as the application is filed. The examination process will also give the applicant significant comfort that the mark is available for use before investing a great deal of money in its promotion.

Final Observations

Selecting an appropriate method of acquiring intellectual property protection for a new development may involve several forms of protection. Software, for example, may be protected by patent, trade secret, and copyright protection and may be sold using a registered trademark. In addition, other legal means may be used to protect the software, including appropriate contractual provisions limiting and restricting the rights of a user acquired by license. Those contractual commitments may survive, even if the intellectual property protection is lost.

It is hoped that this section has provided at least a general overview by which the nonlawyer can begin to understand how to protect intellectual property. There are almost never any easy answers to the question of how to protect products, ideas, and services, and it is always advisable to consult a qualified attorney with specific questions. A knowledgeable client, however, can make a significant difference in achieving the strongest protection available.

20.2 Product Liability and Safety

George A. Peters

Introduction

Almost all engineers, at some time in their career, can expect some direct contact or indirect involvement with the legal system. The contact may be in the form of having to answer written interrogatories on technical issues for a company defendant or plaintiff, being personally deposed and having to respond to oral questions under oath, appearing for a company during trial, assisting lawyers in a lawsuit, or personally being a defendant in a lawsuit. Most important is having the ability to translate legal requirements into engineering specifications to assure compliance with the law.

The old maxim "ignorance of the law is no excuse" should be supplemented by an understanding that ignorance of the law may result in unnecessary mistakes (illegal acts) and personal fear when first confronted by an unknown aspect of the legal process. It is essential that the engineer have sufficient understanding to avoid gross errors and omissions, to react appropriately to legal proceedings, and to want to build a basic foundation of knowledge that can be quickly enhanced when needed. This section is only a brief overview that might be illustrative and helpful in both understanding potential legal liability and how to avoid, prevent, or proactively minimize any such legal exposure. If product liability is legal fault for an unsafe design, the question then becomes how to achieve an appropriate level of safety.

Legal Concepts

Fundamental to determining who might be legally "at fault" is the concept of *negligence* which is utilized worldwide in the apportionment of damages (legal redress). Negligence is the failure to exercise ordinary or reasonable care, which persons of ordinary prudence would use to avoid injury to themselves or others. The exact definition is given to jurors, usually in the form of approved jury instructions (the actual operative law), who then apply the law given to them to the facts and circumstances of the case before them. The defenses to allegations of negligence (absence of due care) are, usually, *contributory negligence* (fault) on the part of the plaintiff or *assumption of the risk* on the part of the plaintiff (that risk that is specific and voluntarily assumed, not general or coerced). If there is a violation of a statute or regulation there may be a rebuttable presumption of negligence. Compliance with a technical standard may be some evidence of the exercise of due care. Concepts of *strict liability* involve the presence of a defect that legally caused personal injury or property damage. There are many definitions of a *defect*, such as a failure to perform as safely as an ordinary consumer would expect when used in an intended or reasonably foreseeable manner or "excessive preventable risks."

Foreseeability means that the personal injury, property damage, or environmental harm must have been predictable or knowable at the time of design, manufacture, or sale of the product. Generally, the law requires only *reasonable efforts* to prevent defects, deficiencies, or unsafe conditions. In other words, there should be efforts to predict possible harm and reasonably practical risk reduction efforts to minimize the harm.

The term *engineering malpractice* includes conduct that has personal legal consequences (professional liability) for the individual engineer, conduct that has adverse legal consequences for his or her employer (such as product liability or toxic torts), and conduct having moral or ethical consequences even though it may be legally protected (Peters, 1996a).

There are many other supplemental legal concepts and each state or jurisdiction has summaries of the law (Witkin, 1987–1990), approved jury instructions (Breckinridge, 1994), statutes enacted by the legislature, compendiums of case law decisions by the courts, and regulations issued by executive agencies (all of which are constantly being revised and expanded). Thus, the engineer should always consult with an attorney-at-law before making an interpretation or taking any action that might be within the province of a licensed professional.

There are thousands of technical standards issued by professional associations, trade groups, standards formulation organizations, and government agencies. Compliance with such standards is the first line of liability prevention, but such standards should be exceeded by a confortable margin to accommodate design, material, fabrication, and in-use process variance (Council, 1989). However, there are other important liability prevention measures that should be undertaken during the design process and described in a product liability prevention plan.

Risk Assessment

A central theme in liability prevention is "risk assessment". The first step in such an assessment is to identify all probable *hazards* (those faults, failures, or conditions that could cause harm), then determine the quantitative *risk* for each (the frequency and severity of harm), and, finally, render a subjective judgment as to *danger* (the presence of excessive preventable danger). It is important to determine what kind of risk assessment is being made and to identify its objective as follows:

1. *Compliance*. The exact method of conducting a risk assessment may be specified by procurement specifications, industry standards, or government regulations. The design objective may be only compliance with the assessment requirement. However, in the process of performing a written risk assessment and classifying the risk into some severity level, it may result in a beneficial safety audit of the product. Where there is "residual risk," the design of specific warnings and instructions may be required.
2. *Major Compliance Tasks*. Where safe performance is very important, a detailed engineering analysis may be required by contract or regulation. This might involve listing all probable hazards for each component, part, and subsystem, then attempting to quantify the risk estimate for each hazard at the 10^{-6} level of attempted precision. Since this requires a major effort, the primary design and management objective should be product improvements and informed product assurance.
3. *Comparative Analysis*. Some risk assessments involve only an overall risk estimate which is then compared with other similar products or a wide range of products. The objective or use may be for marketing purposes or liability defense. Since the results are gross or macroscopic, such risk assessments generally do not result in product improvement.
4. *Risk Ratings*. Some trade associations may provide generic risk ratings for materials or products. The objective may be for an "informed choice" of procedures and field equipment in order to satisfy a legal "duty of care" or to help determine the "best practical means" of job performance.
5. *System Effectiveness*. Some risk assessments are performed early in the design process, perhaps as part of a reliability analysis, for the purpose of predicting final system effectiveness. The objective may be to check design efforts, reallocate available funds, or refocus management audits to achieve a desired level of system effectiveness. The process may be used to assure that desired system-of-systems performance is achievable.

From the societal viewpoint, some level of risk can be tolerable, acceptable, required, and specified. What is desired at a given locality or time period can be made known by the local common law, government regulations, trade standards, practices, customs, and expectations. From the engineering viewpoint, *risk levels are controllable, adjustable, manageable, and a consequence of the application of appropriate engineering techniques, skills, and information resources*.

Engineering Analysis

Rather than rely only upon an engineer's subjective and often biased judgment as to what constitutes a safe design, it is advisable to perform specific objective engineering analyses for design safety. This usually includes some systematic approach to identifying failure modes and field hazards, their risk consequences, and alternative design options that might improve the safety performance. This should

be supplemented by formal design review sessions that consider information from or about customers and learned intermediaries in the fabrication, packaging, shipping, distribution, and marketing system. Written hazard analyses should include consideration of legal duties (Peters, 1991a,b), liability prevention techniques (Peters, 1996b), and cultural attributes and technical information resources in the worldwide marketplace (Murakami, 1987, 1992). There should be a systems perspective, a cradle-to-ultimate-disposal philosophy, a true understanding of customer needs and characteristics, and appropriate application of specific design safety techniques.

Testing is required to *verify* the engineering analyses, to *prove* the inherent safety of the product or process system, and to *discover* all potential problems before they become manifest postsale. As part of a *product liability mitigation plan*, procedures and costs should be determined for possible accident investigations, product recalls, retrofits, and injury reparations. A continuing (postsale) *product surveillance plan* should cover all foreseeable users, servicing, maintenance, repair, modification, transport, disposal, and recycling of the product and its components. This permits early discovery and resolution of safety problems. It should also include a means for updating a knowledge bank of scientific, engineering, legal, insurance, patent, and foreign standards information useful for future design efforts as well as postsale obligations.

Human Error

Human error is a major source of undesirable variance in human–machine interfaces. Unfortunately, many engineers neglect or virtually ignore the human factors aspects of product and system design. There should be some effort to control human performance and prevent human failure by designing for a wide range of human dimensions, characteristics, and predictable responses (Peters, 1996a). If possible, the physical attributes, kinetics, and creative perceptual abilities of human operators, maintenance and repair personnel, and bystanders should be utilized in design. Human factors should be part of any early engineering analysis, with appropriate testing and safeguarding for human error that cannot otherwise be eliminated. This includes mechanical guards, tamper-resistant features, safe-stop and limited movement switches, proximity sensors with directional control, built-in time for protective behavioral reactions, and the prevention of inadvertent activation and operation. One of the most important sources of helpful information on human error comes from incident and accident reconstruction, but the scope and bias of the inquiry may severely limit the design usefulness of the data obtained; for example,

1. The *fatalistic approach* where human error is considered as being inevitable in an imperfect world. If there is "no fault" other than by the person committing an error or omission, there is little incentive to investigate in detail to determine other factors for purposes of corrective action. This fosters the continuance of tolerable human error, persistent undesirable human error, and a lack of true recognition of causation and preventive actions.
2. The *behavioral approach* which has a focus on individual behavior in an attempt to develop "safer people," safer attitudes, and to develop motivation to "act responsibly." This may result in closer supervision and additional training, with some short-term benefits, but it does not permanently alter the error-inducing situation.
3. The *situational approach* to human error is to blame the situation, the work environment, group interactions, sociotechnical factors, and the overall circumstances of the situation. There is some benefit from a broader perspective to human error since it provides a better understanding of causation.
4. The *product design approach* has an emphasis on the interaction between the user and the engineered product to provide information useful to the design engineer.
5. The *multifactorial approach* is based on the assumption that there is multiple causation for each injury, damage, loss, harm, or error. If special attention is given to each substantial factor or cause, valuable design-oriented information can result. This multifaceted perspective of accident reconstruction has the greatest benefit and is compatible with concepts of pure comparative negligence and the allocation of damages in proportion to the degree of fault.

During any accident investigation or accident reconstruction, care should be exercised to prevent any "spoliation" or distruction of evidence. This requires trained specialists, since even slight changes to the product may obliterate information that becomes critical in later product evaluations.

Warnings and Instructions

As a last resort, for residual risk, appropriate use of warnings and instructions is essential for product liability prevention and the safe use of products (Peters, 1993). Such communications require a specific design engineering effort, plus relevant testing, if they are to be effective. They include warning devices, warnings on labels and packaging, material safety data sheets, instructions for training, insertions in owner's or operator's manuals, and postsale advertisements and letters to all owners. Some regulations require that they be in two languages and in nonlanguage pictorials. Such hazard communications and procedural directions should not be the result of a cursory afterthought, but an ongoing integral part of the design process. There may be neuropsychological considerations in the presentation of information by visual or auditory displays, machine condition and status indicators, and computer-generated information that requires specialized knowledge and testing. The basic premise is to design a referent about a specific hazard so the target individual is adequately informed as to risk and has a reasonable choice as to avoidance behavior. Warnings that fail to communicate their intended message are functionally useless. There is considerable scientific and engineering information regarding the design of warnings and instructions, and the failure to heed such information may result in legal allegations about a failure to warn, instruct, test, or appropriately market a product. The issue becomes what proof exists about whether or not warnings, instructions, or representations would have significantly influenced user behavior, purchaser choices (as, for example, in available options), and use of protective equipment. Warnings are an important liability prevention and design safety objective.

References

Breckenridge, P.G., Ed. 1994. *California Jury Instructions.* 2 vols., West Publishing, St. Paul, MN.

Council Directive of 14 June 1989 on the approximation of the laws of the Member States relating to machinery (89/392/EEC), as amended 20 June 1991 (91/368/EEC) and 14 June 1993 (93/44/EEC). *Note*: the Council is the Council of the European Communities. Conformity is indicated by the CE mark on machinery and safety components which must be accompanied by an EC declaration of conformity (93/465/EEC).

Murakami, Y., Ed. 1987, 1992. *Stress Intensity Factors Handbook*, Vols. 1 and 2 (1987), Vol. 3 (1992), The Society of Material Science (Japan), and Pergamon Press, Elmsford, NY.

Peters, G.A. 1991a. The globalization of product liability law, *Prod. Liability Law J.*, Butterworth Legal Publishers, 2(3), 133–145.

Peters, G.A. 1991b. Legal duty and presumptions that are compatible with current technology and future world trade, *Prod. Liability Law J.*, Butterworth Legal Publishers, 2(4), 217–222.

Peters, G.A. 1993. Warnings and alarms, Chap. 4 in Vol. 5 of *Automotive Engineering and Litigation*, Peters, G.A. and Peters B.J., Eds., John Wiley & Sons, New York, 93–120.

Peters, G.A. 1996a. Engineering malpractice and remedies: advanced techniques in engineering liability, *Technol. Law Insurance*, 1, 3–9.

Peters, G.A. 1996a. Human error prevention, Chap. 8 in *Asbestos Health Risks*, Vol. 12 of the *Sourcebook on Asbestos Diseases*, G.A. Peters and B.J. Peters, Eds., Michie, Charlottesville, VA, 207–234.

Peters, G.A. 1996b. Liability prevention techniques, in *Proceedings of the JSME International Symposium on Product Liability and Failure Prevention*, Fukuoka, Japan, 167–184.

Witkin, B.E. 1987–1990. *Summary of California Law*, 9th ed., 13 vols., Bancroft-Whitney, San Francisco.

Further Information

For more detailed information, on product liability and safety, read

Peters, G.A. and Peters B.J., Eds. *Sourcebook on Asbestos Diseases: Medical, Legal, and Engineering Aspects,* 14 vols. Michie, Charlottesville, VA, 1980–1997.

Peters, G.A. and Peters B.J., Eds. *Automotive Engineering and Litigation,* 6 vols. John Wiley & Sons, New York, 1984–1993.

20.3 Bioengineering

Jeff R. Crandall, Gregory W. Hall, and Walter D. Pilkey

The interdisciplinary field of **bioengineering** combines personnel and principles from the physical and life sciences. Although a relatively young field, it is developing rapidly by incorporating experimental and computational techniques from traditional disciplines of engineering and applying them to biological problems. In particular, finite-element techniques and computer modeling are revolutionizing the manner in which bioengineering is conducted and interpreted.

Although based on the principles of traditional disciplines, bioengineering differs from classical engineering because the human body is a living system capable of responding to a changing environment. This results in several challenges for the bioengineer:

- Material properties can vary due to intrinsic and extrinsic factors, e.g., physical properties change with specimen age, gender, and health;
- The human body is capable of self-repair and adaptation, e.g., muscle size increases in response to physical training;
- Many structures within the body cannot be isolated for laboratory testing, e.g., the heart cannot be removed to study circulatory system mechanics.

The discipline of bioengineering includes, but is not limited to, the topics of

- Biomechanics
- Biomaterials
- Biomedical instrumentation and sensors
- Biotechnology
- Genetic engineering
- Medical imaging
- Clinical engineering.

Due to the vast breadth of the field, the authors of this section have not attempted to provide an exhaustive overview of the field of bioengineering but rather have tried to select those areas believed to be of most interest to mechanical engineers. Therefore, this summary of bioengineering is intended to acquaint the reader with the concepts and terminology within the fields of **biomechanics** and **biomaterials** and to provide basic material properties for both tissues and implanted devices.

Biomechanics

The field of biomechanics applies the theories and methods of classical mechanics to biological systems. More precisely, it is concerned with the forces that act on and within a biological structure and with the effects that these forces produce. An overview of topics within the field of biomechanics is provided in Figure 20.3.1. It is evident that the study of stress and strain distributions in biological materials for the development of constitutive equations is a major research emphasis. A review of the constitutive equations and the associated material properties are provided in this section for the most commonly tested tissues.

Hard Tissue Mechanics

The term *bone* refers to two types of hard tissue structure: cortical and cancellous bone. Cortical bone is the dense, structured compact bone that composes the shaft of long bones such as the femur and tibia. Cancellous or trabecular bone is found within the ends of tubular bones and the body of irregularly shaped bones. The mechanical and structural properties of the two types of bone can vary substantially. In addition, bone is capable of **remodeling** in response to its environment. Table 20.3.1 summarizes the general physical and material properties for the two bone types.

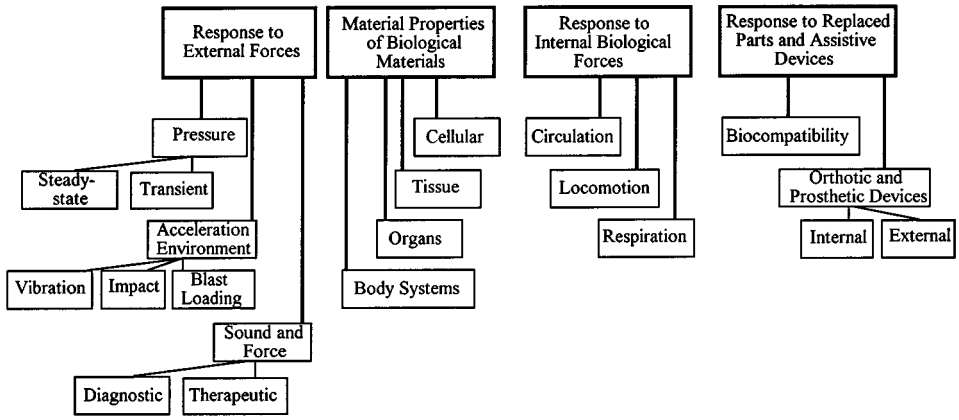


FIGURE 20.3.1 Topics of biomechanics.

TABLE 20.3.1 Physical and Material Properties of Cancellous and Cortical Bone

Bone	Density (kg/m^3)	Poisson's Ratio	Elastic Modulus (GPa)	Tensile Strength (MPa)	Compressive Strength (MPa)	Ref.[a]
Cortical	1700–2000	0.28–0.45	5–28	80–150	106–224	1
Cancellous	100–1000	—	0.1–10	—	1–100	2

[a] 1, Nigg and Herzog (1994); 2, Mow and Hayes (1991).

Like most biological materials, bone behaves as an anisotropic, nonhomogeneous, viscoelastic material. Therefore, the values in Table 20.3.1 exhibit a wide range of scatter and variability due to the simplified model of bone as a linearly elastic isotropic material. It is generally adequate, however, to model bone as a linearly elastic anisotropic material at the strain rates found in most experiments. To address the anisotropy of bone, bone is generally considered to exhibit either transverse isotropic or orthotropic behavior. The constitutive equation for a linearly elastic material can be written using a single-index notation for stress and strain as

$$\sigma_i = c_{ij}\varepsilon_j \tag{20.3.1}$$

where the standard summation convention is used with the indexes possessing a range of 6. The stress-strain relationship can be similarly expressed in terms of the compliance matrix $\mathbf{s_{ij}}$ such that

$$\varepsilon_i = \mathbf{S}_{ij}\sigma_j \tag{20.3.2}$$

Equation (20.3.3) represents the compliance matrix of an orthotropic material in terms of the Young's moduli (E_i), the Poisson's ratio (ν_{ij}), and the shear moduli (G_{ij}).

$$\mathbf{S}_{ij} = \begin{bmatrix} 1/E_1 & -\nu_{21}/E_2 & -\nu_{31}/E_3 & 0 & 0 & 0 \\ -\nu_{12}/E_1 & 1/E_2 & -\nu_{32}/E_3 & 0 & 0 & 0 \\ -\nu_{13}/E_1 & -\nu_{23}/E_2 & 1/E_3 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1/G_{23} & 0 & 0 \\ 0 & 0 & 0 & 0 & 1/G_{31} & 0 \\ 0 & 0 & 0 & 0 & 0 & 1/G_{12} \end{bmatrix} \tag{20.3.3}$$

For an orthotropic material the compliance matrix can be expressed in terms of 12 components, 9 of which are independent. The additional symmetry of the transverse isotropic model results in a further simplification with

$$E_1 = E_2, \quad \upsilon_{12} = \upsilon_{21}, \quad \upsilon_{13} = \upsilon_{31}, \quad G_{23} = G_{31}$$
$$G_{12} = \frac{E_1}{2(1 + \upsilon_{12})} \tag{20.3.4}$$

Table 20.3.2 provides a summary of the material constants for bone. Both mechanical and ultrasonic testing have been used to determine the independent elastic coefficients for bone. The anisotropy of bone requires that mechanical tests be applied in several different directions in order to determine all of the independent elastic coefficients. Ultrasound has the advantage that all elastic coefficients can be measured on a single specimen.

TABLE 20.3.2 Material Constants for Cortical Bone from the Human Femur

Model	E_1 (GPa)	E_2 (GPa)	E_3 (GPa)	G_{12} (GPa)	G_{13} (GPa)	G_{23} (GPa)	ν_{12}	ν_{13}	ν_{23}	ν_{21}	ν_{31}	ν_{32}
TI	11.5	11.5	17.0	3.6	3.3	3.3	0.58	0.31	0.31	0.58	0.46	0.46
Orth.	12.0	13.4	20.0	4.53	5.61	6.23	0.376	0.222	0.235	0.422	0.371	0.350

Note: TI = transverse isotropic; Orth. = orthotropic.

Mechanics of Soft Tissue

The biomechanical properties of soft tissues depend on both the chemical composition and the structure of the tissue. Most soft tissue structures within the body demonstrate aspects of nonhomogeneous, anisotropic, nonlinear viscoelastic behavior. Given the complexity of the constitutive equations, the material properties are difficult to measure and many tests are required. To simplify the test procedures, homogeneity and linearity are frequently assumed.

The simplest representation of viscoelastic material behavior uses combinations of three discrete models comprising linear springs and dashpots: the Maxwell solid, the Voigt model, and the Kelvin model. While these models are generally linear approximations of the nonlinear behavior of biological materials, they can often describe material behavior with reasonable accuracy and can help to visualize tissue behavior.

For improved characterization of the soft tissue response, Fung (1993) developed an approximate theory that was based on the theory of linear viscoelasticity but incorporated nonlinear stress-strain characteristics. More-complex nonlinear viscoelastic models can provide additional improvements in describing tissue response but require extensive experimental testing to determine the model coefficients.

Cartilage

In most joints of the body, the ends of the articulating bones are covered with a dense connective tissue known as hyaline articular cartilage. The cartilage is composed of a composite organic solid matrix that is swollen by water (75% by volume). The cartilage serves to distribute load in the joints and to allow relative movement of the joint surfaces with minimal friction and wear. The coefficient of friction for articular cartilage ranges from 0.001 to 0.1 (Duck, 1990; Mow and Hayes, 1991).

The biomechanical properties of articular cartilage are summarized in Table 20.3.3. The experiment of preference for the testing of articular cartilage has historically been the indentation test. Analysis of the test results has used elastic contact theory, the correspondence principle, and the assumption of material incompressibility ($\nu = 0.5$). This analysis ignores the nonhomogeneous and directional properties of articular cartilage and does not take into account considerable finite deformational effects or the flow

TABLE 20.3.3 Biomechanical Properties of Articular Cartilage

Tissue	Ultimate Compressive Strength (MPa)	Ultimate Compressive Strain	Ultimate Tensile Strength (MPa)	Ultimate Tensile Strain	Elastic Modulus (MPa)	Poisson's Ratio	Ref.[a]
Cartilage	5.0–8.0	0.136	2.8–40	0.182	1.63–220	0.42–0.47	1, 2, 3

[a] 1, Duck (1990); 2, Skalak and Chien (1987); 3, McElhaney (1976).

of the interstitial fluid relative to its porous permeable solid matrix. More-recent models of cartilage have used a biphasic (i.e., an elastic solid and a fluid phase) model to describe more accurately the mechanical response of cartilage.

Muscle

Three types of muscle make up the muscular system: cardiac muscle, smooth or involuntary muscle, and skeletal or voluntary muscle. The focus of this section will be on the skeletal muscle used to maintain the body's posture and to provide movement of the body's segments. The response of skeletal muscle is determined by a combination of active and passive components. The force exerted by a muscle is dependent on the length at which it is stimulated, on the velocity of contraction, on the duration of contraction, and on factors such as fatigue, temperature, and prestretching. A general estimate of the strength of the muscle assumes that its strength is proportional to the physiological cross-sectional area, defined as the muscle volume divided by its true fiber length. The average unit force per cross-sectional area that is exerted by a muscle ranges from 20 to 80 N/cm^2. For more-precise calculations, the relationship between the maximum force of muscle and instantaneous rate of change of its length must be considered. Hill's equation is an empirical relationship expressing the rate of muscle shortening as a function of the isotonic force

$$V = \frac{b\left(F_0 - F\right)}{F + a} \quad \text{or} \quad F = \frac{F_0 b - av}{b + v} \tag{20.3.5}$$

where v is the velocity of shortening, F_o is the force at zero velocity (isometric condition), and F is the instantaneous force. The constants a and b have units of force and velocity, respectively, determined empirically using the relationship

$$K = \frac{a}{F_0} = \frac{b}{V_{\max}} \tag{20.3.6}$$

where $v_{max} = bF_o/a$, the shortening velocity against no load. For most muscles, the range of the muscle constant is $0.15 < k < 0.25$ (Skalak and Chien, 1987).

Tendon/Ligaments

Tendons and ligaments are connective tissues composed primarily of collagen fibers and are normally loaded in tension. Tendons transmit forces from muscles to bone in order to

- Execute joint motion and
- Store energy.

Ligaments attach articulating bones across a joint in order to

- Guide joint movement;
- Limit the joint range of motion;

- Maintain joint congruency; and
- Assist in providing joint stability.

The biomechanical properties of ligaments and tendons are provided in Table 20.3.4.

TABLE 20.3.4 Biomechanical Properties of Tendon and Ligament

Tissue	Ultimate Tensile Strength (MPa)	Ultimate Tensile Strain	Elastic Stiffness (MPa)	Ref.[a]
Ligament	2.4–38	0.20–1.60	2–111	1, 2, 3
Tendon, calcaneal	30–60	0.09–0.10	600	4, 5

[a] 1, Yamada (1970); 2, Skalak and Chien (1987); 3, Nahum and Melvin (1993); 4, Bronzino (1995); 5, Duck (1990).

The general stress-strain behavior of connective soft tissue can be represented by four distinct loading regions (Figure 20.3.2). Region I is commonly referred to as the "toe region" and is associated with the straightening or alignment of the randomly ordered structural fibers. Most physiological loading occurs in this region where there is a small increase in stress for a given increase in strain. Region II characterizes the linear region where the straightened fibers are more uniformly loaded. The tangent to the stress-strain response curve in this linear region is frequently referred to as the elastic stiffness or tangent modulus rather than the elastic modulus in order to emphasize that the soft tissue is not truly behaving as a perfectly elastic material. At the initiation of Region III, small fluctuations in the stress can be observed resulting from microfailures in the soft tissue. The final loading profile shown in Region IV exhibits catastrophic failure of the soft tissue.

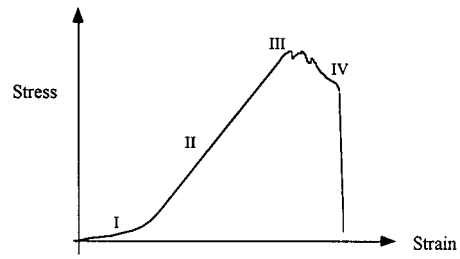


FIGURE 20.3.2 General stress-strain behavior of connective soft tissue.

Factors Affecting Biomechanical Properties

Due to the nature of biological testing, a great deal of variability in the material properties of tissue is evident in the published literature. Unlike traditional engineering materials, there are no standard protocols or procedures for the testing of biological materials. Although some variability can be attributed to differences in the experimental approaches undertaken by the researchers, mechanical behavior of biological materials can also be affected by

- Geometric characteristics;
- Loading mode (i.e., tension, compression, torsion);
- Rate and frequency of loading;
- Specimen preparation (i.e., temperature, preservation method, hydration);
- Age or health of specimen donor;
- Gender;
- Temperature;
- Anatomic location of load and specimen;
- Species.

Rate of Loading. The rate of load can affect both the material properties and the patterns of failure of biological materials. Due to their viscoelastic nature, many biological materials are stronger and stiffer at high rates of loading.

Loading Mode. The stress-strain behavior of biological materials is highly dependent on the orientation of the tissue structure with respect to the direction of loading (Table 20.3.5). Many tissues such as cortical bone, tendon, and muscles have organized structures which result in anisotropy of the physical properties. In some cases, the anisotropy exists but is not incorporated in the constitutive model or experiments and results in increased variability of the data.

TABLE 20.3.5 Material Properties Variation in Bone Due to Specimen Orientation and Loading Mode

Cortical Bone Orientation	Mode	Ultimate Strength (MPa)	Elastic Modulus (MPa)	Ref.
Longitudinal	Tension	133	—	Currey (1984)
	Compression	193	17	
	Shear	68	3.3	

Anatomic Location. Although homogeneity is often assumed, biological tissues typically exhibit significant differences in material properties depending on the anatomic location. Table 20.3.6 illustrates the variability of material properties with anatomic location using articular cartilage from the knee.

TABLE 20.3.6 Dependence Tensile Modulus (*E*) of Human Articular Cartilage on Location

Region	Medial Anterior	Medial Central	Medial Posterior	Lateral Anterior	Lateral Central	Lateral Posterior	Ref.
E (MPa)	159.6	93.2	110.2	159.1	228.8	294.1	Mow and Hayes (1991)

Age. The material properties of biological materials can vary depending on whether the specimens are obtained from immature, mature, or aging tissue. In general, the ultimate strength of tissues decreases with increasing age. Bone, muscle, tendon, cartilage, and ligaments all show decreases in the ultimate strength and elastic moduli after maturity with increasing age (Table 20.3.7). Several tissues, such as skin and teeth, are exceptions to this general rule and can exhibit increases in some material properties with age.

TABLE 20.3.7 Ratio of Age Changes for Ultimate Tensile Strength

Tissue	10–19years	20–29years	30–39years	40–49years	50–59years	60–69years	70–79years
Cortical bone	0.93	1.00	0.98	0.91	0.76	0.70	0.70
Cartilage	1.02	1.00	0.93	0.80	0.56	0.33	0.29
Muscle	1.27	1.00	0.87	0.73	0.67	0.60	0.60
Tendon	1.00	1.00	1.00	1.00	1.00	0.95	0.78
Skin	—	1.00	1.26	1.26	1.08	0.96	0.77

Storage/Preservation. The increasing complexity of biomechanical testing requires a considerable period of time for each individual test and necessitates tissue storage and preservation. It is essential to ensure that the storage and preservation of biological materials is controlled so that the material property measurements in the laboratory accurately reflect the properties of the tissue in the living state. Biological materials are preserved using one of three methods depending on the required storage interval:

- Refrigeration for short-term storage;
- Freezing for long-term storage;
- Embalming for long-term storage.

Refrigeration and freezing have no effect on the properties of hard tissue. Although the effects of refrigeration on soft tissue properties are tissue dependent, stable properties can generally be achieved within 3 days of storage. Freezing of soft tissue can result in ice cavities that disrupt the tissue architecture and result in property changes. The effects of embalming on soft and hard tissue structures exhibit conflicting results.

Because many biological materials are composed primarily of water, humidity can strongly affect the stress-strain relationships (Table 20.3.8). Therefore, care must be taken to keep biological specimens moist prior to and during testing. The control of the moisture level of specimens is most important when analyzing the elastic properties of surface tissues such as skin and hair. Young's modulus of skin at 25 to 30% relative humidity may be 1000-fold greater than that at 100% humidity (Duck, 1990).

TABLE 20.3.8 Comparison of Material Properties between Wet and Dry Cortical Bone

Test Condition	Ultimate Tensile Strength (kg/mm²)	Ultimate Percentage Elongation (%)	Elastic Modulus (kg/mm²)	Ref.
Wet tibia	14.3 ± 0.12	1.50	1840	Yamada (1970)
Dry tibia	17.4 ± 0.12	1.38	2100	Yamada (1970)

Species. Animal tissues have been used extensively in biomechanical testing because of their availability. When extrapolating results to humans, care must be taken since significant differences can exist in the structure and material properties between humans and other animals due to physiological and anatomic differences. Table 20.3.9 shows a comparison of elastic moduli of femoral bone specimens obtained from different animals.

TABLE 20.3.9 Comparison of Bone Elastic Modulus between Different Species

Tissue Source	Tissue	Elastic Modulus (MPa)	Ref.
Human	Femur	18	Currey (1984)
Cow	Femur	23	
Sheep	Femur	22	
Tortoise	Femur	10	

Impact Biomechanics

The prevention of injury through the development of countermeasures is most effectively achieved through biomechanical testing and analysis. The prevalence of injuries resulting from motor vehicle crashes, sporting activities, and industrial accidents has led to the development of a branch of biomechanics referred to as impact biomechanics. In order to achieve the principal aims of prevention of injury through environmental modification, this branch of biomechanics must develop an improved understanding of the mechanisms of injury, descriptions of the mechanical response of the biological materials involved, and information on the human tolerance to impact. Injury criteria for the human body as interpreted by anthropometric dummies are provided in Table 20.3.10. The injury criteria are provided in terms of engineering parameters of resultant acceleration $a(t)$, force $F(t)$, velocity $V(t)$, displacement $s(t)$, and anthropometric scaling factors such as the chest depth D.

Computational Biomechanics

Computational mechanics provides a versatile means of analyzing biomechanical systems. Modeling uses either rigid-body models composed of rigid masses, springs, joints, and contact surfaces or flexible-

TABLE 20.3.10 Adult Midsize Male Injury Criteria for an Impact Environment

Body Region	Injury Criteria	Formulation	Threshold
Head	Head injury criteria (HIC)	$\text{HIC} = (t_2 - t_1)\left[\frac{1}{(t_2 - t_1)}\int_{t_1}^{t_2} a(t)\,dt\right]^{2.5}$	1000
Chest	Compression criteria	$s(t)$	7.5 cm
	Acceleration criteria	$a(t)$	65 g
	Viscous criteria ($V * C$)	$V(t) * s(t)/D$	1 m/sec
Femur	Force criteria	$F(t)$	10 kN

TABLE 20.3.11 Comparison of Rigid Body and Finite Element Modeling Methods

Model Characteristics	Multibody Model	Finite-Element Method
Complexity	Relatively simple	Relatively complex
Fidelity	Requires engineering intuition	Can achieve high fidelity
Efficiency	Very efficient	Computationally expensive
Model Elements	Springs, point masses, rigid bodies, ellipsoids, joints, and contact planes	Flexible elements: bricks, beams, plates, shells

body models with finite or boundary elements (Table 20.3.11). Software is now available that incorporates the two modeling techniques and uses multibody modeling to capture overall kinematics of a biomechanical system and flexible body modeling to provide an in-depth study of those regions of particular interest.

The complexity of biological systems and limited constitutive model data often require that simplifying assumptions be made during the development of models. Therefore, it is necessary to verify the model before conducting parametric and sensitivity studies.

Biomaterials

The application of engineering materials to replace, support, or augment structures in the human body has become a major thrust of medical research in the last half of the 20th century. These efforts have helped to improve quality of life and longevity in individuals suffering from a variety of diseases and injuries. Since the materials involved with this repair are implanted within the biochemically and mechanically active environment of the body, the response of both the host and material has been investigated.

A biomaterial is defined as a nonviable material used in a medical device that is intended to interact with biological systems. The interaction of a biomaterial with its host can be classified into four categories (Black, 1992):

- *Inert* — implantable materials which elicit a minimal host response.
- *Interactive* — implantable materials which are designed to elicit specific, beneficial responses, such as ingrowth and adhesion.
- *Viable* — implantable materials, possibly incorporating live cells at implantation, which are treated by the host as normal tissue matrices and are actively resorbed and/or remodeled.
- *Replant* — implantable materials consisting of native tissue, cultured ***in vitro*** from cells obtained previously from the specific implant patient.

The concept behind *inert* implants is to introduce an "innocuous" implant to the host that will perform its mechanical role without altering the local environment. Decades of research on host and implant response have informed the biomedical community that no implant is inert and that every material will elicit a response from the host. For many biomaterial applications, the objective is to elicit a *minimal*

host response. Those implants that are *interactive* with the biological environment do so in a desired and planned fashion. *Viable* implants are at their infancy, but already show promise in assisting and possibly directing tissue repair. The concept of an entire organ *replant* is still unrealized, but current strides in deciphering the human genetic code will make replants of **autologous** materials a clinical method of the future.

Material and Host Response

The Physiological Environment. The first reaction of the body to an implant is to "wall off" or build a fibrous tissue capsule around the implant. The biomaterial will be enclosed in a fibrous capsule of varying tissue thickness that will depend on the material and the amount of motion between the tissues and the implant. When the biomaterial elicits a minimal encapsulating response, it is possible for the tissue to integrate with the implant mechanically. This has been demonstrated in total-joint replacement when new bone is formed to interdigitate with a porous surface. Another approach to **biocompatibility** has been to use the tissue response to a material to benefit the function of the implant. This approach is used when bone chemically bonds to the hydroxyapetite coatings on hip implants.

The interaction between implant material and the host occurs at both the chemical and mechanical level. There are several parameters that affect the response of the biomaterial to the host environment and the overall performance:

- *The pH and salt content of the implant site* — A Pourbaix diagram for saline environments should be consulted to see if the metal will be passivated.
- *Lipid content of the implant site* — Lipophilic materials may swell or lose strength.
- *Mechanical environment* — Implant materials may wear, creep, and/or fatigue.
- *Corrosive potential of the material* — Metallic implants are susceptible to galvanic, crevice, and other types of corrosion.
- *Method of manufacture and handling* — Scratches and other defects will increase the likelihood of crack development and/or crevice corrosion.
- *Method of sterilization* — Some methods of sterilization will cross-link polymers or affect the degradation of biodegradable materials.

The host response to an implanted material is also affected by both chemical and mechanical factors. A few parameters that affect host response are

- *Leached materials from the implant* — Leached material, consisting of metallic ions, polymeric molecules, or solvents from manufacturing, may denature proteins. The pathway and effects of leached material should be examined.
- *Surface tension and electric charge of the material*.
- *Stiffness of the biomaterial* — Stress shielding of the bone by a stiff implant has been shown to affect bone remodeling.
- *Geometry of the implant* — Sharp edges have been shown to increase the fibrous tissue capsule thickness.
- *Method of cleaning and sterilization* — Solvents, filings, or residues left from the manufacturing, handling, or sterilization processes will likely cause unwanted host response and increase the probability of infection.

The expected life span of an implant varies with application. Some implants are used for temporary structural support, such as intermedullary nails for fracture fixation. In this application, corrosion is not as critical an issue as cost and mechanical fatigue. Another approach to temporary applications has been the use of biodegradable materials that eventually dissolve into the body. Biodegradable materials have been used for fracture fixation and drug delivery with success.

Current Biomaterials

Metallic Biomaterials. Metals are typically selected for biomedical applications because of their high tensile strength (Table 20.3.12). With the high tensile strength of metals comes a high elastic modulus that can lead to stress shielding of the bone. Bone, which remodels in response to its loading environment, resorbs in regions of low stress which can lead to implant loosening.

TABLE 20.3.12 Common Metallic Biomaterials

Metal	Form	Density (g/cm^3)	*E* Modulus (GPa)	Yield Strength (MPa)	Fatigue Strength @ 10^7 Cycles (MPa)	Characteristics	Medical Applications	Ref.[a]
Ti-6Al-4V	AN	4.4	127	830–896	620	Chemically inert, poor wear properties, good fatigue properties, high strength-to-weight ratio, closest modulus to that of bone	Total hip and knee stems	1, 5
CP-Ti	—	120	470	—				
Co-Cr-Mo, ASTM F-75	C/AN	7.8	200	450–492	207–310	Excellent wear properties, castable, Co and Cr ions have been found mutagenic *in vitro*	Articular surfaces in hips and knees, dental implants, hip stems	2, 3
W/AN	9.15	230	390	—				
AISI-316LVM Stainless	AN	7.9	210	211–280	190–230	Inexpensive	Temporary applications, bone screws, bone plates, suture	5–7
30% CW	—	230	750–1160	530–700				
Tantalum	CW	16.6	190	345	—	Very inert, very dense	Transdermal implant testing, suture	4

Note: AN: annealed, CW: cold worked, C: cast, W: wrought, CP: chemically pure.

[a] 1, ASTM F136-79; 2, ASTM F75-82; 3, ASTM F90-82; 4, ASTM F560-86, p. 143, 1992; 5, Green and Nokes (1988); 6, ASTM F55-82; 7, Smith and Hughes (1977).

Another issue when using metals in the body is corrosion since very few metals are in a passivated state while ***in vivo***. Metallic implant materials are subject to several types of corrosion: galvanic, crevice, pitting, intergranular, and stress/fatigue. For metallic biomaterial selection, it is best to refer to the corrosion potential of a material in seawater, rather than tap water since seawater is a reasonable approximation for the ***in vivo*** environment. Implant designs that involve dissimilar metals or multiple pieces have increased susceptibility to galvanic corrosion (Cook et al., 1985). A reasonable prediction of the reactivity between dissimilar metals can be made upon consideration of the galvanic series.

Polymers. The number of polymers available for biomedical applications has been increasing rapidly in the last few decades. Polymers are viscoelastic materials whose mechanical properties, host response, and material response depend on molecular weight, degree of cross-linking, temperature, and loading rate, among other factors. Therefore, it should be recognized that tabular data representing the mechanical properties of polymers only indicates the general range of properties for a class of polymers (Tables 20.3.13 and 20.3.14). Common applications of polymers in biomedical engineering include articular wear components, drug-delivery devices, cosmetic augmentation, blood vessels, and structural applications.

Biodegradable Polymers. There are a plethora of degradable polymers available to the modern biomedical engineer. Each is unique in mechanical properties, degradation time, degradation products, and sensitivity of strength to degradation. Recent efforts in polymer development have focused on biodegradable materials that serve as either temporary support or drug-delivery devices during the healing

TABLE 20.3.13 Bulk Material Properties of Some Thermoset Polymers

Thermoset	Density (g/cm^3)	Modulus (GPa)	Yield Strength (MPa)	Ultimate Strength (MPa)	Elongation (%)	Applications	Comments	Ref.[a]
Polyethylene terephthalate (dacron, polyester)	1.40	2.41	62.06	150–250	70–130	Suture, mesh, vascular grafts, heart valves, fluid transfer implants, artificial tendons	Subject to negligible creep, heat resistance up to 250°F, steam sterilizable, poor wear resistance, susceptible to hydrolysis and loss of strength	1, 2
PMMA thermoset	1.15–1.20	2.4–3.1	15.8	9.7–32	2.4–5.4	Acrylic bone cement	Very biologically inert, poor tensile strength, monomer lowers blood pressure, may cause thermal necrosis during curing at 95°C, radiolucent without addition of barium sulfate	2, 3
Polyurethane	1.10	5.9	—	45	750	Implant coatings	Good resistance to oil and chemicals	2
Silicone Rubber						Cosmetic augmentations, nose, ear, chin, and breast implants	Polymer backbone is silicon, rubbery mechanical properties, lipophylic, material elicits a minimal host response and has very flexible, cartilage-like stiffness, may cause local sclerosis and inflammatory phenomena. Recent *in vitro* studies on silicone have demonstrated an absence of cytotoxic response. There is no current epidemiological data to support termination of its use	2, 4, 5
Heat vulcanized	1.12–1.23	<1.4	—	5.9–8.3	350–600			
High performance	0.98–1.15	2.4	—	8.3–10.3	700			

[a] 1, Park (1979), 2, Lee et al. (1995); 3, Lautenschlager et al. (1984); 4, Frisch (1984); 5, Polyzois et al. (1994).

TABLE 20.3.14 Thermoplastic Biomedical Polymers

Thermoplastic	Density (g/cm^3)	Modulus (GPa)	Yield Strength (MPa)	Ultimate Strength (MPa)	T_{glass} (°C)	Elongation (%)	Applications	Comments	Ref.[a]
PMMA thermoplastic	1.19	2.4–3.1	—	50–75	105	2–10	Contact lenses, blood pump	Very biologically inert, excellent optical properties, radiolucent without barium sulfate added, losses strength when heat sterilized	1–3
Polypropylene	0.85–0.98	1.5	—	30–40	–12	50–500	Disposable syrine, suture, artificial vascular grafts	High flex life, excellent environment stress cracking resistance	2, 1
Polyvinylchloride (PVC) rigid	1.35–1.45	3.0	—	40–55	70–105	400	Blood and solution bags, catheters, dialysis devices	High melt viscosity, thus difficult to process	2, 1
Polysulfone	1.23–1.25	2.3–2.48	65–96	106	—	20–75	Hemodialysis, artificial kidney circulatory assist, composite matrix	High thermal stability and chemical stability, unstable in polar organic solvents such as ketones or chlorinated hydrocarbons	2, 3
Polytetra-flouroethylene (Teflon)	2.15–2.20	0.5–1.17	—	17–28	—	320–350	Catheter, artificial vascular grafts	Heat sterilizable, low coefficient of friction, not recommended as a bearing surface, inert in solid form	1, 5
UHMWPE molded and machined extruded	0.93–0.97 0.93–0.94	— 1.24	21 21–28	34 34–47	— —	300 200–250	Articular bearing surfaces	UHMW has MW > 2E6 g/mol, very inert, induces minimal swelling *in vivo*, high concentration of wear particles leads to bone resorbtion, creeps under load, has heat resistance up to 180°F which is too low for steam sterilization	4

[a] 1, Park (1979); 2, Lee et al. (1995); 3, Dunkle (1988); 4, ASTM 648-83; 5, ASTM F754.

TABLE 20.3.15 Properties of Degradable Polymer Fiber

Polymer	Crystallinity	T_{melt}/T_{glass} (°C)	Modulus/Strength (GPa/MPa)	Ultimate Elongation (%)	Common Material Applications
Polyglycolic acid PGA	High	230/36	8.4/890	30	Suture, nerve guidance channels, chondrocyte scaffolds
Poly-L-lactic acid PLLA	High	170/56	8.5/900	25	Suture, stents, bone plates, screws
Polyglactine910	High	200/40	8.6/850	24	Skin regeneration
Polydioxanone	High	106/<20	8.6/850	35	Monofilament suture

Source: Kimura, Y., in *Biomedical Applications of Polymeric Biomaterials,* Tsuruta, T. et al., Eds., CRC Press, Boca Raton, FL, 1993. With permission.

process. One of the primary advantages to the use of biodegradable implants and suture is that they do not require surgical removal. This aspect decreases both the risk of infection and cost. The main mechanism of *in vivo* degradation is hydrolytic degradation, although some enzymes may also serve a role (Piskin, 1995). For use of these biomaterials, an in-depth knowledge of the candidate polymer is recommended. A basic description of several degradable polymers is provided in Table 20.3.15.

Ceramics. Ceramic materials are some of the strongest and hardest material structures used in engineering. Ceramics are composed of primarily inorganic compounds and are generally characterized by unstable crack growth that leads to poor tensile strength (Table 20.3.16). There are two categories of ceramic biomaterials: relatively inert and **bioactive**. Relatively inert ceramics have high compressive strength and hardness and are commonly used for wear applications. Bioactive ceramics are designed to bond with the host tissue and do not possess great strength characteristics (Table 20.3.17).

Optimum biocompatibility of ceramic degradable biomaterials occurs only with special proportions of the material constituents. Unfortunately, the confines of these proportions limit the mechanical properties of the material. These materials are not strong enough for structural applications, but have been successfully used as coating material to enhance tissue bonding, as dental restorative material, and as filler in bone cement (Bajpai and Billotte, 1995).

Composite Biomaterials. Composite biomaterials combine the properties of two or more different materials in order to obtain a biomaterial with tailored properties. With composite biomaterials, the opportunity exists to fabricate an implant that is stronger and lighter than conventional implants while exhibiting a stiffness that is very similar to surrounding tissues. The material properties of composite materials depend on the matrix material, reinforcement material, volume fraction of reinforcement material, interfacial bond strength, orientation of reinforcement, number of inclusions, and other factors. Use of composite biomaterials is in its infancy but has already been used to improve the strength, stiffness, and toughness in current biocompatible materials.

In biodegradable composite materials, the method used to acquire final implant geometry will affect the service life of the implant. Machining has been shown to expose fiber ends on the surface and to promote the wicking of surrounding fluid into the device, which can increase the rate of degradation.

Materials of Natural Origin. Materials from natural origin may be xenogenous (i.e., obtained from other species) or autogenous (i.e., obtained from the patient). Xenogenous biological materials from nonhuman animals are commonly used for soft tissue replacement. These materials exhibit mechanical properties that are very similar to the surrounding living human tissue but may cause an immune response due to their non-self proteins. Examples of a xenogenous biomaterial are the porcine heart valve, catgut suture, and collagen-based implants. Recent research on heart and kidney transplants from primates have demonstrated the feasibility of using living xenogenous biomaterials.

TABLE 20.3.16 Material Properties of Relatively Inert Biomedical Ceramics

Material	Density (g/cm^3)	Grain Size (nm)	*E* Tensile Modulus (GPa)	Hardness (HV) (N/mm^2)	Fatigue Strength (MPa)	Comments	Applications	Ref.[a]
Alumina (AL_2O_3)	3.9–4	3000–4000	380	23,000	550	Very inert, excellent wear and friction properties, may increase tissue aluminum causing bone demineralization	Articular bearing surfaces, dental implants, total joint prostheses	1, 2, 6–8
LTI carbon	1.7–2.2	3–4	18–28	150–250	280–560	Excellent blood compatibility, low density	Coating or structural material for heart valves and blood vessels	3
Vitreous carbon	1.4–1.6	1–4	24–31	150–200	70–210	Known as glassy carbon due to lack of crystal structure	Artifical heart valves	3
Zirconia (ZrO_2)	6.1	<0.5	200	16,000	1200	Nonreactive in rhesus monkey bone *in vivo*, excellent wear and friction properties	Articular bearing surfaces	4, 5, 8

[a] Boutin et al. (1988); 2, ASTM F560-78; 3, Intermedics Orthopedics (1983); 4, Christel et al. (1989); 5, Ducheyne and Hastings (1984); 6, Graves et al. (1972); 7, Toni et al. (1994), 8, Hentrich et al. (1971).

TABLE 20.3.17 Bioactive Ceramic Materials

Material	Chemical Content	Comments	Ref.[a]
Hydroxyapatite	$Ca_{10}(PO_4)_6(OH)_2$	Actual mineral phase of bone, used as a coating or solid, synthetic versions available, good mechanical properties, excellent biocompatibility, subject to osteoclastic resorbtion, chemically bonds to bone	1, 2
Bioglass 46S5.2	46.1% SiO_2, 26.9% CaO, 24.4% Na_2O, 2.6% P_2O_5 mol%	Fine-grained, glassy ceramics, 46S5.2 exhibits best tissue bonding, other constitutent ratios are available	3, 4
Ceravital	Bioglass materials with Al_2O_3, TiO_2, and Ta_2O_5 added	Very similar composition to bioglass with additional metal oxides to control dissolution rate	4

[a] 1, Gessink et al. (1987); 2, Kay (1988); 3, Hench and Ethridge (1982); 4, Bajpai and Billotte (1995).

Autogenous materials are used when skin and bone are relocated on a patient. As human ability to manipulate the human genetic code continues, it may be possible to replace failed human organs with identical organs that have been grown either *in vitro* or *in vivo*.

Biomaterial Testing

The human body is a biologically and mechanically active environment that reacts to all materials. Biomaterials should be initially selected for their engineering mechanical properties to serve a specific function. Careful consideration must be given to the fatigue and creep properties of the material because of the rigorous loading incurred by many implants. The anterior cruciate ligament of a moderately active person, for example, will be subjected to 4.2 million loading cycles annually (Black, 1992). This frequency of cyclic loading is consistent with the cyclic loading of other orthopedic structures.

Once a candidate material has been selected, the material must be tested to determine both the material response to its proposed biological environment and the host response to that material. Valuable host and material response information may be obtained from the use of *in vitro* tests that simulate a specific physiological environment. The final step before implantation into a human is to test the performance of a candidate material with an animal model.

In Vitro Studies. Experimental *in vitro* studies should be conducted prior to testing on living animals. *In vitro* studies provide essential information on the response of the candidate biomaterials to a simulated host environment and on the chemical interactions that ensue. It is important to observe the effect of the biomaterial on proteins to determine if any coating, denaturing, or other processes result. These studies are inexpensive and can be tailored to simulate specific implant sites. Initial tests may simply expose the candidate material to the expected inorganic chemical and thermal conditions. More complex *in vitro* tests may include appropriate, viable, active cells with their associated cell products. The cell survival count, reproduction rate, metabolic activity, effective motion activity, and cell damage should be observed during testing (Black, 1992). Some sample *in vitro* materials tests are described in Table 20.3.18.

Biomaterials intended for vascular applications must be tested for their interaction with blood. Examination of blood-biomaterial interactions should include protein absorption, platelet interactions, intrinsic coagulation, fibrinolytic activity, erythrocytes, leukocytes, and complement activation. Since the circulatory system has both an oxygen- and nutrient-rich arterial side and a carbon dioxide-rich venous side, a candidate material must be tested in the environment for which it is intended.

In Vivo Studies. Animal studies are extremely valuable for examining the host and material response of a specific test material. These studies must be in complete accordance with the Animal Welfare Act (7 U.S.C. 2131, December 23, 1985). Initial animal tests are nonfunctional and involve placing a test specimen in an anatomic location similar to the expected implant location. These locations may be subcutaneous, intramuscular, intraperitoneal, transcortical, or intramedullar. Depending on the study, the duration of implantation can vary from weeks to years. Factors that affect the outcome of nonfunctional

TABLE 20.3.18 Sample *In Vitro* Tests

In Vitro Test	Brief Description
ASTM F895-4 cytotoxicity	Place sterile agar over mouse fibroblast cells. Place the test material on top and incubate for 24 hr; examine the culture for the extent of cell lysis
Ames test for mutagenicity	Expose autotrophic *Salmonella typhinurium* bacteria cells to the test material and histidine for 48 hr at 37°C; if cells mutate to non-autotrophic state, histidine levels will be lower after testing
Lee–White coagulation rate	Place a small volume of fresh-drawn blood on a test surface and similar amount on control surface; compare coagulation times
Ex vivo blood compatibility	Tap into the bloodstream of a canine; let blood flow at typical rates and pressure through a test specimen for a short time; examine specimen for platelet adhesion and blood for hemolysis

tests include the degree of relative motion between implant and host and the geometry of the test specimen.

Functional testing of biomedical materials is performed in order to evaluate a test material or design as it performs its intended function. Species are selected for an animal model based on similarities with human models and then scaled to size. Some factors to consider are the life span, activity level, metabolic rate, and size of the test animal. Results are determined from macroscopic and histological evaluations.

Defining Terms

Autologous: Related to self; materials obtained from the same organism.

Bioactive: The ability of a material to chemically interact with the host environment in a predetermined manner.

Biocompatibility: The ability of a biomaterial to perform with an appropriate host response in a specific manner.

Bioengineering: The application of principles from engineering, applied mathematics, and physics to the study of biological problems.

Biomaterial: A nonviable material used in a medical device that is intended to interact with biological systems.

Biomechanics: The study of forces that act on and within a biological structure and the effects that these forces produce.

In vitro: Within a glass; observable in an artificial environment such as a test tube.

In vivo: Within the living body.

Remodeling: Changes in internal architecture and external conformation of biological tissues in accordance with applied strain.

Sterilization: The complete elimination or destruction of all living microorganisms. The most common methods are steam, radiation, and chemical sterilization.

References

Bajpai, P.K. and Billotte, W.G. 1995. Ceramic biomaterials; in *The Biomedical Engineering Handbook*, J.D. Bronzino, Ed., CRC Press, Boca Raton, FL, chap. 41, 552–580.

Black, J. 1992. *Biological Performance of Materials; Fundamentals of Biocompatibility*, Marcel Dekker, New York.

Boutin, P., Christel, P., Dorlot, J.M., Meunir, A., de Roquancourt, A., Blanquaert, D., Herman, S., Sedel, L., and Witvoet, J. 1988. *J. Biomed. Mater. Res.*, 22,1203.

Bronzino, J.D., Ed. 1995. *The Biomedical Engineering Handbook*, CRC Press, Boca Raton, FL.

Christel, P., Meunier, A., Heller, M., Torre, J.P., and Peille, C.N. 1989. *J. Biomed. Mater. Res.*, 23, 45.

Cook, S.D., Renz, E.A., Barrack, R.L., Thomas, K.A., Harding, A.F., Haddad, R.J., and Milicic, M. 1985. *Clin. Orthop.*, 194(236).

Cowin, S.C. 1989. *Bone Mechanics,* CRC Press, Inc., Boca Raton, FL.

Currey, J.D. 1984. *Mechanical Adaptations of Bone*, Princeton University Press, Princeton, NJ.

Ducheyne, P. and Hastings, G.W., Eds. 1984. *Metal and Ceramic Biomaterials*, CRC Press, Boca Raton, FL.

Duck, F. 1990. *Physical Properties of Tissue*, Academic Press, San Diego, CA.

Dunkle, S.R. 1988. Engineering plastics, in *Engineered Materials Handbook*, Vol. 2, C.A. Dostal, Ed., ASM Int., Metals Park, OH, 200ff.

Frisch, E. 1984. Polymeric materials and artificial organs, in *ACS Symposium* 256, C.G. Gebelein, Ed., American Chemical Society, Washington, D.C., 63ff.

Fung, Y.C. 1993. *Biomechanics: Mechanical Properties of Living Tissues*, 2nd ed. Springler-Verlag, New York.

Gessink, R.G., deGroot, K., and Klein, C. 1987. Chemical implant fixation using hydroxyapatite coatings, *Clin. Orthop. Relat. Res.*, 226(147).

Green, M. and Nokes, L.D.M., Eds. 1988. *Engineering Theory in Orthopaedics: An Introduction*, Ellis Horwood, Chichester, England.

Hench, L.L. and Ethridge, E.C. 1982. *Biomaterials: An Interfacial Approach*, Academic Press, New York.

Kay, J.F. 1988. Bioactive surface coatings: cause for encouragement and caution, *J. Oral Implantol.*, 16(43).

Kimura, Y. 1993. Biomedical polymers, in *Biomedical Applications of Polymeric Biomaterials*, T. Tsuruta, T. Hayashi, K. Kataoka et al., Eds., CRC Press, Boca Raton, FL.

Lautenschlager, E.P., Stupp, S., and Keller, J.C. 1984. In *Functional Behavior of Orthopedic Biomaterials*, Vol. II., P. Ducheyne and G.W. Hastings, Eds., CRC Press, Boca Raton, FL, 87ff.

Lee, H.B., Kim, S.S., and Khang, G. 1995. Polymeric biomaterials, in *The Biomedical Engineering Handbook*, J.D. Bronzino, Ed., CRC Press, Boca Raton, FL, 581–597.

McElhaney, J.H., Roberts, V.L., and Hilyard, J.F., 1976. Handbook of Human Tolerance, Japanese Automobile Research Institute, Tokyo, Japan.

Mow, V.C. and Hayes, V.C. 1991. *Basic Orthopedic Biomechanics*, Raven Press, New York.

Nahum, A.M. and Melvin, J.W., Eds. 1993. *Accidental Injury: Biomechanics and Prevention*, Springler-Verlag, New York.

Nigg, B.M. and Herzog, W., Eds. 1994. *Biomechanics of the Musculo-Skeletal System*, John Wiley and Sons, Chichester, England.

Nordin, M. and Frankel, V.H. 1989. *Basic Biomechanics of the Musculoskeletal System*, Lea and Febiger, Malvern, PA.

Park, J.B. 1979. *Biomaterials: An Introduction*, Plenum Press, New York.

Piskin, E. 1995. Biodegradable polymers as biomaterials, *J. Biomat. Sci. Polym. Ed.* 6(9), 775–795.

Polyzois, G.L., Hensten-Pettersen, A., and Kullmann, A. 1994. An assessment of the physical properties and biocompatibility of three silicone elastomers, *J. Prosth. Dent.* 71(5), 500-504.

Skalak, R. and Chien, S. 1987. *Handbook of Bioengineering*, McGraw-Hill, New York.

Smith, D.J.E. and Hughes, A.N. 1977. The influence of frequency and cold work on fatigue strength of 316L stainless steel in air and 0.17M saline, *AWRE Rep.,* 44/83/189, Atomic Weapons Research Establishment, Aldermaston, Berkshire, U.K.

Yamada, H. 1970. In *Strength of Biological Materials*, F.G. Evans, Ed., Williams and Wilkins, Baltimore, MD.

Further Information

A comprehensive summary of the biomechanical properties of tissues can be found in the *Strength of Biological Materials* (Yamada, 1970). A good introduction to material response, material testing, and host response is presented in *Biological Performance of Materials; Fundamentals of Biocompatibility* (Black, 1992).

A reference that goes more in-depth on biomaterial chemical contents, applications, and manufacturing is *The Biomedical Engineering Handbook*, edited by Joseph Bronzino (1995).

The *Journal of Biomedical Materials Research* is a monthly publication that reports advancements in the field of biomaterial development and testing. For subscription information contact: Journal of Biomedical Materials Research, Subscription Fulfillment and Distribution, John Wiley and Sons, Inc., 605 3rd Avenue, New York, NY 10158.

The *Journal of Biomechanics* covers a wide range of topics in biomechanics, including cardiovascular, respiratory, dental, injury, orthopedic, rehabilitation, sports, and cellular. For subscription information, contact: Elsevier Science, Inc., 660 White Plains Road, Tarrytown, NY 10591-5153.

20.4 Mechanical Engineering Codes and Standards

Michael Merker

What Are Codes and Standards?

A **standard** can be defined as a set of technical definitions, requirements, and guidelines for the uniform manufacture of items; safety, and/or interchangeability. A **code** is a standard which is, or is intended to be, adopted by governmental bodies as one means of satisfying legislation or regulation. Simply put, they can range from a general set of minimum requirements to very specific "how to" instructions for designers and manufacturers.

Voluntary standards, and they can run from a few paragraphs to hundreds of pages, are written by experts who sit on the many committees administered by standards-developing organizations (SDOs), such as ASME International.

They are not considered voluntary because they are created by volunteers; rather they are voluntary because they serve as guidelines, but do not of themselves have the force of law. ASME International publishes its standards; accredits users of standards to ensure that they have the capability of manufacturing products that meet those standards; and provides a stamp that accredited manufacturers may place on their product, indicating that it was manufactured according to the standard. ASME International cannot, however, force any manufacturer, inspector, or installer to follow ASME International standards. Their use is voluntary.

Why are voluntary standards effective? Perhaps the American Society for Testing and Materials (ASTM) said it best in its 1991 annual report. "Standards are a vehicle of communication for producers and users. They serve as a common language, defining quality and establishing safety criteria. Costs are lower if procedures are standardized; training is also simplified. And consumers accept products more readily when they can be judged on intrinsic merit."

A dramatic example of the value and impact codes and standards have had on our society is provided by ASME International's Boiler and Pressure Vessel Code. Toward the end of the 19th century, boilers of every description, on land and at sea, were exploding with terrifying regularity for want of reliably tested materials, secure fittings, and proper valves. They would continue to do so into the 20th century. Engineers could take pride in the growing superiority of American technology, but they could not ignore the price of 50,000 dead and 2 million injured by accidents annually.

The mechanical engineers who tackled the problems in 1884 began by seeking reliable methods for testing steam boilers. The need for the establishment of universally accepted construction standards would take many more years and resulted in the first edition of the Boiler and Pressure Vessel Code being published by ASME International in 1915.

Codes and Standards–Related Accreditation, Certification, and Registration Programs

Accreditation

Shortly after the Boiler Code was first published, the need emerged for a recognizable symbol to be affixed to a product constructed in accordance with the standards. ASME International commissioned appropriate seals that are now the internationally acknowledged symbols of the Society. The symbol is stamped onto the product.

But how does a manufacturer obtain permission to use one of the symbols? Through the ASME International **accreditation** process, the manufacturer's quality control process is reviewed by an ASME International team. If the quality control system meets the requirements of the applicable ASME International code or standard and the manufacturer successfully demonstrates implementation of the program, the manufacturer is accredited by ASME International. This means that the manufacturer may certify the product as meeting ASME International standards and may apply the stamp to the product.

The stamp consists of a modified cloverleaf (from the shape of the ASME International logo), with letter(s) in the center. The letter(s) indicate the code or section of the code met by the product upon which it is placed. Boiler and pressure vessel stamps issued are

A Field Assembly of Power Boilers,
E Electric Boilers,
H Heating Boilers, Steel Plate, or Cast-Iron Sectional,
HV Heating Boiler Safety Valves,
HLW Lined Potable Water Heaters,
M Miniature Boilers,
N Nuclear Components,
NPT Nuclear Component Partial,
NA Nuclear Installation/Assembly,
NV Nuclear Safety Valves,
PP Pressure Piping,
RP Reinforced Plastic Pressure Vessels,
S Power Boilers,
U, U2 Pressure Vessels,
UM Miniature Pressure Vessels,
UV Pressure Vessel Safety Valves,
V Boiler Safety Valves.

ASME International also has accreditation programs for nuclear materials, offshore safety and pollution prevention equipment, fasteners, authorized inspection agencies, organizations that certify elevator inspectors, and reinforced thermoset plastic corrosion-resistant vessels.

Certification

ASME International has also expanded its scope of activity to cover **certification** of individuals. The first program became available in 1992 and covered the qualification and certification of resource recovery facilities operators. They have since added programs to cover operators of hazardous waste incinerators and medical waste incinerators. Programs to certify an individual's knowledge and ability in the area of geometric dimensioning and tolerancing and operators of fossil fuel–fired plants are under development.

Registration

Registration is similar to accreditation; however, it is the term used more frequently in the international arena, particularly when dealing with the International Organization for Standardization (ISO) 9000 program on quality assurance systems. ASME International's ISO 9000 Registration Program has been accredited by the American National Accreditation Program for Registrars of Quality Systems (ANSI-RAB) and the Dutch Council for Certification (RvC) in the following industrial sectors:

Primary Metals Industries
Fabricated Metal Products
Industrial and Commercial Machinery and Equipment
Reinforced Thermoset Plastic Tanks and Vessels
Engineering Services

How Do I Get Codes and Standards?

Role of ASME International in Mechanical Engineering Standards

ASME International is a nonprofit educational and technical organization with more than 125,000 members, most of whom are practicing engineers. About 20,000 are students. ASME International has a wide variety of programs: publishing, technical conferences and exhibits, engineering education,

government relations, and public education, as well as the development of codes and standards, all aimed at serving the engineering profession, the public, industry, and government.

The ASME International Board of Governors has delegated the codes and standards activity to a 22-member Council on Codes and Standards, which directs all aspects of the program. Under the Council are ten boards, also made up of ASME International members and other interested persons; supervisory boards in turn oversee committees, each responsible for a specific area of standard development.

Committees in one form or another have dealt with standards since the first test code in 1884. Currently, there are more than 100 main committees dealing with over 500 standards that are under regular review and revision. Once a standard is accepted, it is printed and made available to manufacturers, regulatory agencies, designers — anyone with an interest in that particular subject. Close to 4000 individuals serve on these committees and their subcommittees, subgroups, and working groups.

After a standard has been considered and reconsidered at meetings and through many drafts, it is sent to the main committee, representing all interests, which votes on the standard. But this is not the final step. Before the draft becomes a standard and is published and ready for distribution, it is made available for public comment and must be approved by the appropriate ASME International supervisory board. This process is illustrated graphically in Figure 20.4.1.

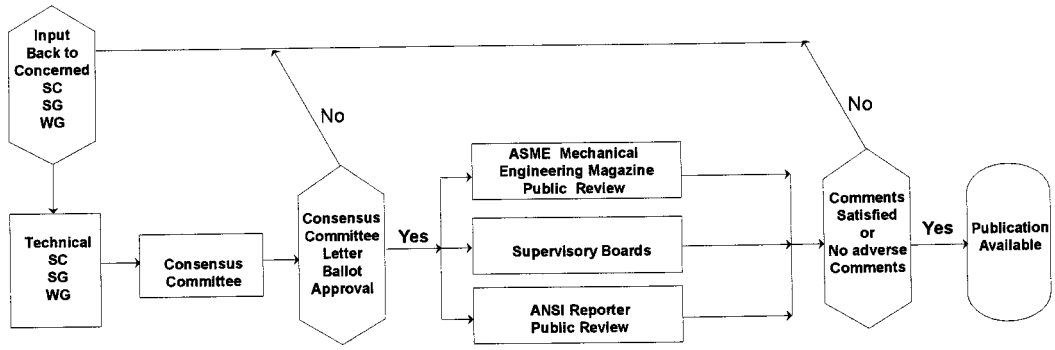


FIGURE 20.4.1 A typical path for standards approval.

ASME International has been a consistent supporter of the policy of prior announcement of meetings, open meeting rooms, **balanced committees**, public announcements and reviews, appeal mechanisms, and overall procedural **due process**.

Role of ANSI in These and Other Related Standards

In 1911, ASME International was one of a number of organizations which recommended the establishment of an organization to help eliminate conflict and duplication in the development of voluntary standards in the U.S. Such an organization was formed in 1918 and is currently known as the American National Standards Institute (ANSI). ANSI is the U.S. member of the ISO and the administrator of the U.S. National Committee of the International Electrotechnical Committee (IEC). ANSI also serves as a bookseller of domestic and international standards.

The intent of obtaining ANSI approval of a standard is to verify that in establishing the standard, the originating organization has followed principles of openness and due process and has achieved a **consensus** of those directly affected by the standard.

What Standards Are Available?

Listing of Topics Covered by ASME International Standards

Abbreviations
Accreditation
Air Cooled Heat Exchangers
Air Cylinders & Adapters
Air Heaters
Atmospheric Water Cooling Equipment
Automatically Fired Boilers
Automotive Lifting Devices
Backwater Valves
Boilers
Bolts
Building Services Piping
Cableways
Cargo Containers
Carriers
Castings and Forgings
Centrifugal Pumps
Chemical Plant and Petroleum Refinery Equipment
Chucks and Chuck Jaws
Cleanouts
Coal Pulverizers
Compressors
Consumable Tools
Conveyors
Coordinate Measuring Machines
Cranes
Deaerators
Density Determination
Derricks
Dial Indicators
Diaphragm Seals
Dies
Diesel and Burner Fuels
Digital Systems
Dimensional Metrology
Dimensioning and Tolerancing
Drafting
Drains
Dumbwaiters
Ejectors
Elevators
Escalators
Exhausters
Fans
Fasteners
Feedwater Heaters
Fittings
Flanges
Floor Drains
Flue Gas Desulfurization
Fluid Flow in Pipes
Fuel Gas Piping
Gage Blanks
Gage Blocks
Gas Flow Measurement
Gas Transmission and Distribution Piping Systems
Gas Turbine Power Plants
Gas Turbines
Gaseous Fuels
Gaskets
Gauges
Graphic Symbols
Hand Tools
High Lift Trucks
Hoists
Hooks
Hydroelectric Equipment
Incinerators
Indicated Power
Industrial Sound
Industrial Trucks and Vehicles
Internal Combustion Engine Generator Units
Ion Exchange Equipment
Jacks
Keys
Keyseats
Knurling
Letter Symbols
Lifts
Limits and Fits
Line Conventions and Lettering
Linear Measurements
Liquid Transportation Systems
Low Lift Trucks
Machine Guarding
Machine Tools
Manlifts
Material Lifts
Measurement
Mechanical Power Transmission Apparatus
Mechanical Springs
Metric System

Milling Machines
Model Testing
Monorails
Moving Walks
Nuclear Facilities and Technology
Nuts
Offshore Oil and Gas Operations
Oil Systems
Optical Parts
Pallets
Particulate Matter
Performance Test Codes
Pins
Pipe Dimensions
Pipe Threads
Piping
Pliers
Plumbing
Pressure Transducers
Pressure Vessels
Pumps
Quality Assurance
Reamers
Refrigeration Piping
Resource Recovery Facility Operators
Retaining Rings
Rivets
Safety and Relief Valves
Screw Threads
Screws
Slings
Slip Sheets
Spray Cooling Systems
Stainless Steel Pipe
Stands
Steam-Generating Units
Steel Stacks
Storage/Retrieval Machines
Storage Tanks
Surface Texture
Temperature Measurement
Thermometers
Tools
Transmission Apparatus
Transmission Chains
Turbines
Valves
Washers
Waste Facility Operators
Water Hammer Arresters
Weighing Scales
Welded Aluminum-Alloy Storage Tanks
Wheel Dollies
Wheelchair Lifts
Whirlpool Bathtub Appliances
Wind Turbines
Window Cleaning
Wrenches

Where Do I Go If the Subject I Want Is Not on This List?

With the constant creation and revision of standards, not to mention the large number of different SDOs, it is impractical to search for standards-related information in anything other than an electronic format. The latest information on ASME International's codes and standards can be found on the World Wide Web (WWW). ASME International's home page is located at http://www.asme.org. It contains a searchable catalog of the various codes and standards available as well as general information about the other areas ASME International is involved with. Additional information on drafts out for public review and committee meeting schedules will be added as the home page continues to evolve.

Another useful Web site is provided by the National Standards System Network (NSSN). This is a project which is still under development, but has the goal of presenting a comprehensive listing of all bibliographic information on standards and standards-related material. Eventually, this site may also provide direct electronic access to the standards themselves. The NSSN page is located at http://nssn.org. This site is also useful in that it provides links to many of the SDO's Web sites.

Defining Terms

Accreditation: The process by which the ability of an organization to produce a product according to a specific code or standard is evaluated. It is not an actual certification of a specific product, but does provide a third-party evaluation of the manufacturer's competence to certify that individual products are in compliance with the applicable standards.

Balanced committee: A committee in which the consensus body responsible for the development of a standard comprises representatives of all categories of interest that relate to the subject (e.g., manufacturer, user, regulatory, insurance/inspection, employee/union interest). A balanced committee ensures that no one interest group can dominate the actions of the consensus body.
Certification: The process by which an individual's training or abilities to perform a task according to a specific code or standard is evaluated.
Code: A standard which is, or is intended to be, adopted by governmental bodies as one means of satisfying legislation or regulation.
Consensus: This means that substantial agreement has been reached by directly and materially affected interest groups. It signifies the concurrence of more than a simple majority, but not necessarily unanimity. Consensus requires that all views and objections be considered and that an effort be made toward their resolution.
Due process: A procedure by which any individual or organization who believes that an action or inaction of a third party causes unreasonable hardship or potential harm is provided the opportunity to have a fair hearing of their concerns.
Registration: Similar to accreditation, it is the term used more frequently in the international arena, particularly when dealing with the ISO 9000 program on quality assurance systems.
Standard: A set of technical definitions, requirements, and guidelines for the uniform manufacture of items; safety, and/or interchangeability.

Further Information

The Engineering Standard, A Most Useful Tool by Albert L. Batik is a comprehensive work covering the impact of standards in marketing and international trade as well as their application to traditional areas such as design, manufacturing, and construction.

The Code: An Authorized History of the ASME Boiler and Pressure Vessel Code by Wilbur Cross provides an in-depth look at events which led to need for codes and the pioneers who created the Boiler and Pressure Vessel Code.

20.5 Optics*

Roland Winston and Walter T. Welford (deceased)

Geometrical Optics

Geometrical optics is arguably the most classical and traditional of the branches of physical science. By the time Newton wrote his *Principia*, geometrical optics was already a highly developed discipline. Optical design of instruments and devices has been worked out and improved over the centuries. From the telescopes of Galileo to the contemporary camera lens, progress while impressive has been largely evolutionary with modern design benefiting enormously from the availability of fast, relatively inexpensive digital computers. It is noteworthy that even in the last 20 years progress has been made by extending the classical model to problems where image formation is not required, or desired. We shall touch on these developments of "nonimaging optics" later in this section. But first we treat classical geometrical optics. In this model, a "point source" emits rays which are straight lines in a vacuum or in a homogeneous isotropic dielectric medium. Light travels at different speeds in different dielectrics. Its speed is given by c/n, where c is the speed in vacuum (299,792,458 m sec^{-1}) and n, the *refractive index*, depends on the medium and on the frequency of the light.

A ray is refracted at an interface between two media. If $\mathbf{r}$ and $\mathbf{r}'$ are unit vectors along the incident and refracted directions, n and n' are the respective refractive indexes, and $\mathbf{n}$ is the unit normal to the interface, then the ray directions are related by

$$n\mathbf{n} \times \mathbf{r} = n'\mathbf{n} \times \mathbf{r}' \tag{20.5.1}$$

which is the law of refraction, *Snell's law*, in vector form. More conventionally, Snell's law can be written

$$n \sin I = n' \sin I' \tag{20.5.2}$$

where I and I' are the two angles formed where the normal meets the interface, the angles of incidence and refraction. The two rays and the normal must be coplanar. Figure 20.5.1 illustrates these relationships and shows a reflected ray vector $\mathbf{r}''$. Equation (20.5.1) can include this by means of the convention that after a reflection we set n' equal to $-n$ so that, for reflection,

$$\mathbf{n} \times \mathbf{r} = -\mathbf{n} \times \mathbf{r}'' \tag{20.5.3}$$

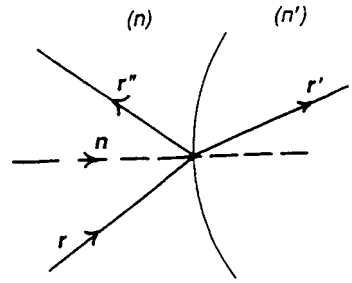


FIGURE 20.5.1

With a bundle or pencil of rays originating in a point source and traversing several different media, e.g., a system of lenses, we can measure along each ray the distance light would have traveled in a given time t; these points delineate a surface called a *geometrical wavefront*, or simply a *wavefront*. Wavefronts are surfaces orthogonal to rays (the Malus–Dupin theorem). (It must be stressed that wavefronts are a

*From Welford, W.T., *Useful Optics,* reproduced with permission of University of Chicago Press.

concept of geometrical optics and that they are *not* surfaces of constant phase, phase fronts, of the light waves in the scaler or electromagnetic wave approximations. However, in many situations the geometrical Earth are a very good approximation of phase fronts.) Thus, if successive segments of a ray are of length d_1, d_2, …, a wave front is a locus of constant Σnd or, passing to the limit, $\int n\, dl$. This quantity is called an *optical path length*. (See Figure 20.5.2.)

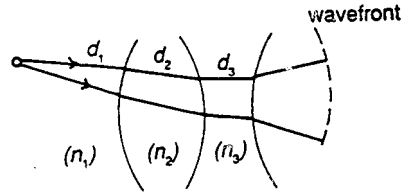


FIGURE 20.5.2

Optical path lengths enter into an alternative to Snell's law as a basis for geometrical optics. Consider any path through a succession of media from, say, P to P'. We can calculate the optical path length W from P to P', and it will depend on the shape of this path, as shown in Figure 20.5.3. Then *Fermat's principle* states that, if we have chosen a physically possible ray path, the optical path length along it will be stationary (in the sense of the calculus of variations) with respect to small changes of the path. (The principle as originally formulated by Fermat proposed a *minimum* time of travel of the light. Stationarity is strictly correct, and it means roughly that, for any small transverse displacement δx of a point on the path, the change in optical path length is of order δx^2.) For our purposes, Fermat's principle and Snell's law are almost equivalent, but in the case of media of continuously varying refractive index it is sometimes necessary to invoke Fermat's principle to establish the ray path. Apart from such cases, either one can be derived from the other.

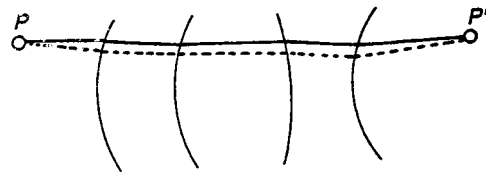


FIGURE 20.5.3

Either Fermat or Snell can be used to develop the whole edifice of geometrical optics in designing optical systems to form images.

Symmetrical Optical Systems

The axially symmetric optical system, consisting of lenses and/or mirrors with revolution symmetry arranged on a common axis of symmetry, is used to form images. Its global properties are described in terms of *paraxial* or *Gaussian* optics. In this approximation only rays making small angles with the axis of symmetry and at small distances from the axis are considered. In Gaussian optics, we know from symmetry that rays from any point on the axis on one side of the system emerge on the other side and meet at another point on the axis, the *image point* . This leads to the well-known formalism of principal planes and focal planes shown in Figure 20.5.4. A ray entering parallel to the axis passes through F', the second, or image-side, principal focus on emerging from the system, and a ray entering through F, the first principal focus, emerges parallel to the axis. A ray incident on the first, or object-side, principal plane P at any height h emerges from the image-side principal plane P' at the same height h so that the principal planes are *conjugated* planes of unit magnification. Excluding for the moment the special case

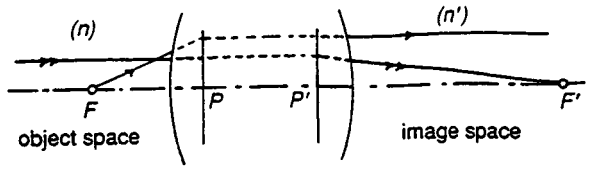


FIGURE 20.5.4

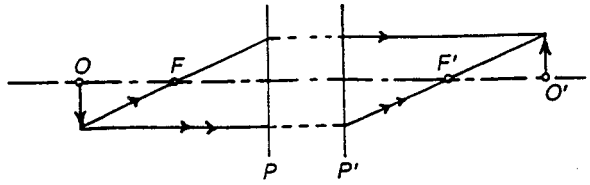


FIGURE 20.5.5

in which a ray entering parallel to the axis also emerges parallel to the axis, these four points yield a useful graphical construction for objects and images, as depicted in Figure 20.5.5.

The two focal lengths f and f' are defined as

$$f = PF, \qquad f' = P'F' \tag{20.5.4}$$

Their signs are taken according to the usual conventions of coordinate geometry, so that in Figure 20.5.4 f is negative and f' is positive. The two focal lengths are related by

$$n'/f' = -n/f \tag{20.5.5}$$

where n and n' are the refractive indexes of the object and image spaces, respectively.

Conjugated distances measured from the principal planes are denoted by l and l', and the conjugate distance equation relating object and image positions is

$$n'/l' - n/l = n'/f' = -n/f \tag{20.5.6}$$

The quantity on the right — that is, the quantity on either side of Equation (20.5.5) — is called the *power* of the system, and is denoted by K.

Another form of the conjugate distance equation relates distances from the respective principal foci, z and z'.

$$zz' = ff' \tag{20.5.7}$$

This equation yields expressions for the transverse magnification:

$$\eta'/\eta = -f/z = -z'/f' \tag{20.5.8}$$

This is useful to indicate paraxial rays from an axial object point O to the corresponding image point O' as in Figure 20.5.6 with convergence angles u and u' positive and negative, respectively, as drawn in the figure. (Paraxial angles are small but diagrams like Figure 20.5.6 can be drawn with an enlarged transverse scale. That is, convergence angles and intersection heights such as h can all be scaled up by the same factor without affecting the validity of paraxial calculations.) Then, if η and η' are corresponding object and image sizes at these conjugates, the following relation exists between them:

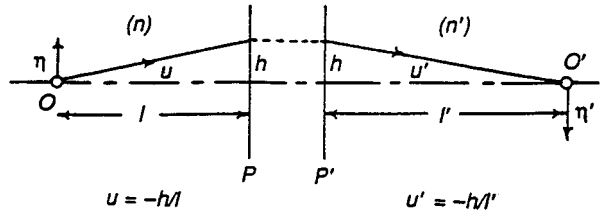


FIGURE 20.5.6

$$nu\eta = n'u'\eta' \tag{20.5.9}$$

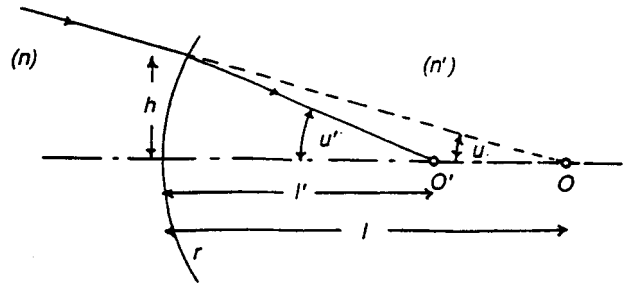
In fact, for a given paraxial ray starting from O, this quantity is the same at any intermediate space in the optical system. That is, it is an invariant, called the Lagrange invariant. It has the important property that its square is a measure of the light flux collected by the system from an object of size η in a cone of convergence angle u.

The above discussion covers all general Gaussian optic properties of symmetrical optical systems. We next look at particular systems in detail. To do this, we abandon the skeleton representation of the system by its principal planes and foci and consider it as made up of individual refracting or reflecting surfaces.

Figures 20.5.7 and 20.5.8 show the basic properties of a single spherical refracting surface of radius of curvature r and of a spherical mirror. In each case r as drawn is positive. These diagrams suggest that the properties of more complex systems consisting of more than one surface can be found by tracing paraxial rays rather than by finding the principal planes and foci, and this is what is done in practice. Figure 20.5.9 shows this with an iterative scheme outlined in terms of the convergence angles. The results can then be used to calculate the positions of the principal planes and foci and as the basis of aberration calculations. For details see Welford (1986). The actual convergence angles which can be admitted, as distinguished from notional paraxial angles, are determined either by the rims of individual components or by stops deliberately inserted at places along the axis chosen on the basis of aberration theory. Figure 20.5.10 shows an *aperture stop* in an intermediate space of a system. The components of the system to the left of the stop from an image (generally virtual) which is "seen" from the object position (this image is usually virtual, i.e., it is not physically accessible to be caught on a ground-glass screen like the image in an ordinary looking glass); this image is called the entrance pupil, and it limits the angle of beams that can be taken in from the object. Similarly, on the image side there is an exit pupil, the image of the stop by the components to the right, again usually virtual. This pupil may also determine the angles of beams from off-axis object point O and O'; the central ray of the beam from O passes through the center of the entrance pupil (and therefore through the center of the aperture stop and the center of the exit pupil) and it is usually called the *principal*, *chief*, or *reference ray* from this object point. The rest of the beam or pencil from O may be bounded by the rim of the entrance pupil, or it may happen that part of it is *vignetted* by the rim of one of the components.

Although the aperture stop is usually thought of as being inside an optical system, as in a photographic objective, it is sometimes placed outside, and one example is the *telecentric stop* shown in Figure 20.5.11. The stop is at the object-side principal focus, with the result that in the image space all the principal rays emerge parallel to the optical axis. A telecentric stop can be at either the object-side or the image-side principal focus, and the image conjugates can be anywhere along the axis. The effect is that the pupil on the opposite side of the telecentric stop is at infinity, a useful arrangement for many purposes. It may happen that the telecentric stop is between some of the components of the system.

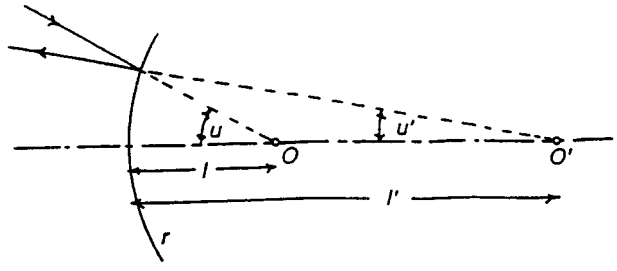
The above information is all that is needed to determine how a given symmetrical optical system behaves in Gaussian approximation for any chosen object plane. Suitable groups of rays can be used to set out the system for mechanical mounting, clearances, etc.; however, it is often easier and adequate in terms of performance to work with the *thin-lens model* of Gaussian optics. This model uses complete



$$\frac{n'}{l'} - \frac{n}{l} = \frac{n' - n}{r} = K$$

$$n'u' - nu = -hK$$

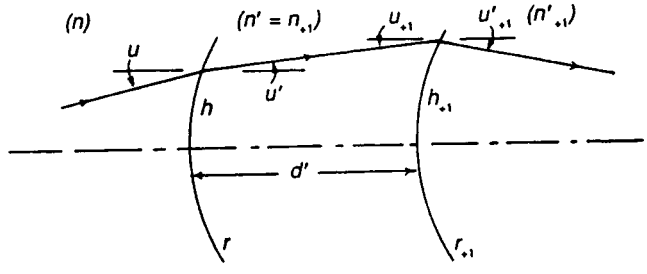
FIGURE 20.5.7



$$\frac{1}{l'} + \frac{1}{l} = \frac{2}{r} = -K$$

$$u' + u = hK$$

FIGURE 20.5.8



$$n'u' - nu = -hK$$

$$u_{+1} = u'$$

$$h_{+1} = h + d'u'$$

$$n'_{+1}u'_{+1} - n_{+1}u_{+1} = -h_{+1}K_{+1}$$

FIGURE 20.5.9

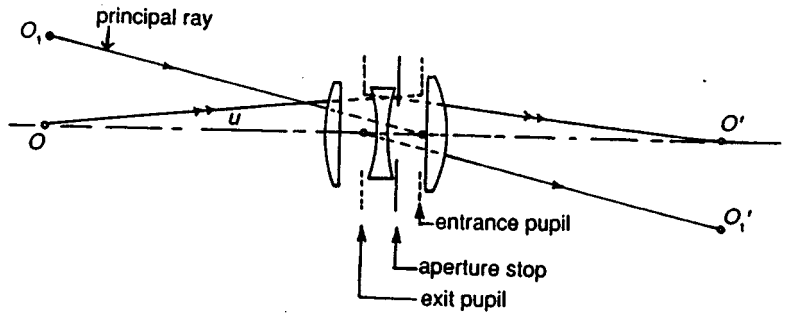


FIGURE 20.5.10

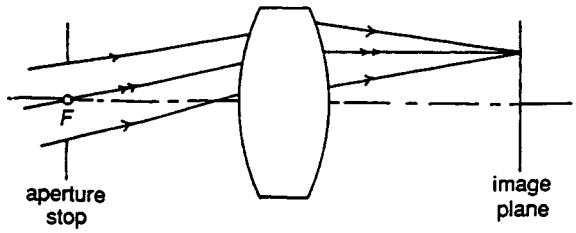


FIGURE 20.5.11

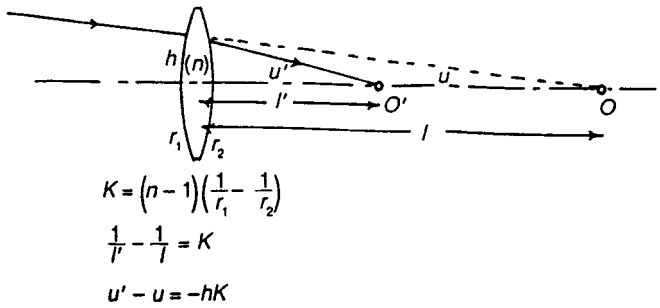


FIGURE 20.5.12

lenses of negligible thickness instead of individual surfaces. Figure 20.5.12 shows the properties of a thin lens. A system of thin lenses can be ray traced to find its properties, locate foci and ray clearances, etc., and very often the results will be good enough to use without further refinement. This is particularly true of systems involving unexpanded laser beams, where the beam diameters are quite small.

We omitted from our discussion of Figure 20.5.4 the special case in which a ray incident parallel to the optical axis emerges parallel to the axis, as in Figure 20.5.13. This is an *afocal* or *telescopic* system; it forms an image at infinity of an object at infinity, and the angular magnification is given by the ratio of the ray incidence heights. An afocal system also forms images of objects at finite distances, as indicated by the rays drawn in the figure. The transverse magnification is then constant for all pairs of conjugates.

Plane Mirrors and Prisms

A single-plane mirror used to deflect or rotate an optical axis needs no explanation, but some useful points can be made about combinations of mirrors. Two mirrors at an angle θ turn the beam through 2θ about the line of intersection of the mirror planes, whatever the angle of incidence on the first mirror, as in Figure 20.5.14. The diagram is drawn for a ray in the plane perpendicular to the line of intersection

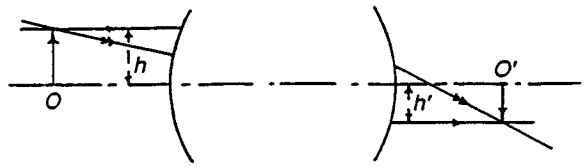


FIGURE 20.5.13

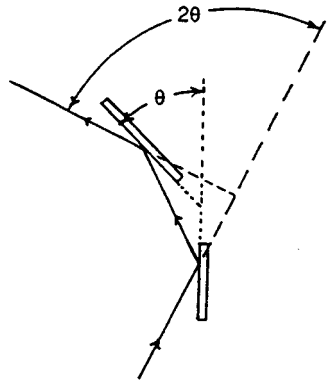


FIGURE 20.5.14

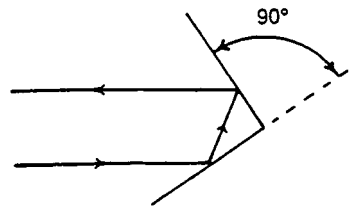


FIGURE 20.5.15

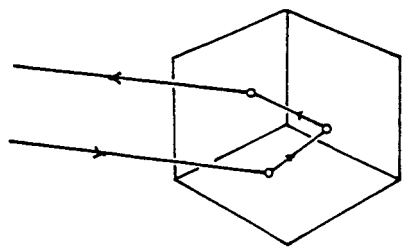


FIGURE 20.5.16

of the mirror planes, but it is equally valid if the ray is not in this plane, i.e., the diagram is a true projection. In particular, if the mirrors are at right angles, as in Figure 20.5.15, the direction of the ray is reversed in the plane of the diagram. Three plane mirrors at right angles to each other, forming a corner of a cube as in Figure 20.5.16, reverse the direction of a ray incident in *any* direction if the ray meets all three mirrors in any order.

These properties are more often used in prisms in the corresponding geometry. Total internal reflection, as in, for example, the right-angle prism (Figure 20.5.17), is a great advantage in using prisms for turning beams. The condition for total internal reflection is

$$\sin I > 1/n \tag{20.5.10}$$

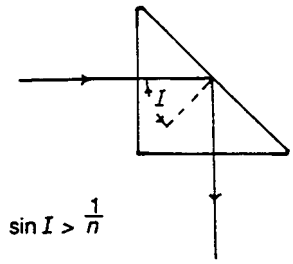


FIGURE 20.5.17

The *critical angle* given by $\sin I = 1/n$ is less than 45° for all optical glasses, and probably for all transparent solids in the visible spectrum. Total internal reflection is 100% efficient provided the reflecting surface is clean and free from defects, whereas it is difficult to get a metallized mirror surface that is better than about 92% efficient. Thus, with good anti-reflection coating on the input and output surfaces a prism, such as that shown in Figure 20.5.17, transmits more light than a mirror.

Roof prisms and cube-corner prisms, the analogs of Figure 20.5.15 and 20.5.16, have many uses. The angle tolerances for the right angles can be very tight. For example, roof edges form part of the reversing prism system in some modern binoculars, and an error ε in the right angle causes an image doubling in angle of $4n\varepsilon$. The two images are those formed by the portions of the beam incident at either of the two surfaces, which should have been at exactly 90°.

In addition to turning the axis of a system, mirror and prism assemblies sometimes rotate the image in unexpected ways. The effect can be anticipated by tracing, say, three rays from a notional object such as the letter F (i.e., an object with no symmetry). A more direct and graphic method is to use a strip of card and mark arrows on each end as in Figure 20.5.18a. The card is then folded without distorting it as in Figure 20.5.18b to represent, say, reflection at the hypotenuse of the right-angle prism, and the arrows show the image rotation. The process is repeated in the other section, as in Figure 20.5.18c. Provided the folding is done carefully, without distortion, this procedure gives all image rotations accurately for any number of successive reflections.

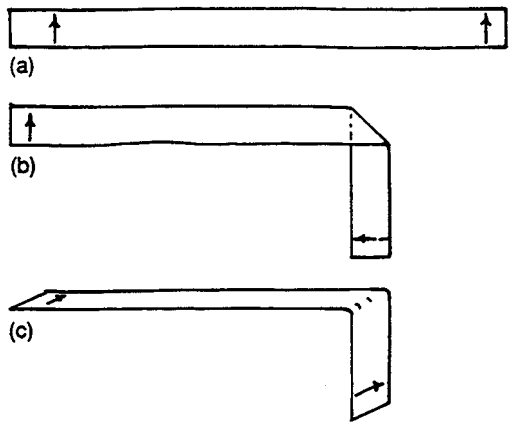


FIGURE 20.5.18

The Dove prism (Figure 20.5.19) is an example of an image-rotating prism. When the prism is turned through an angle ϕ about the direction of the incident light, the image turns in the same direction through 2ϕ. A more elaborate prism with the same function is shown in Figure 20.5.20. The air gap indicated between the hypotenuses of the two component parts needs to be only about 10 μm or so thick to ensure total internal reflection. Any prism or mirror assembly like this with an odd number of reflections will serve as an image rotator. Figure 20.5.20 illustrates an elegant advantage of prisms over mirrors; the system can be made compact by using the same optical surface both for reflection and for transmission.

FIGURE 20.5.19

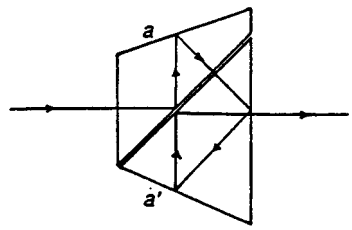


FIGURE 20.5.20

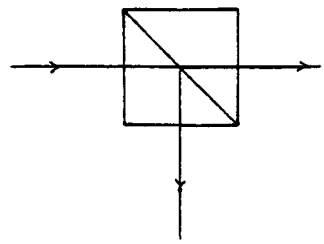


FIGURE 20.5.21

Figure 20.5.21 shows a typical beam-splitting (or combining) prism, a component of many diverse optical systems. The beam-splitting surface may be approximately neutral, in which case it would be a thin metal layer, or it may be dichroic (reflecting part of the spectrum and transmitting the rest), or it may be polarizing (transmitting the p-polarization and reflecting the s-polarization of a certain wavelength range). In the last two cases the reflecting-transmitting surface is a dielectric multilayer and its performance is fairly sensitive to the angle of incidence.

Prisms as devices for producing a spectrum have been largely replaced by diffraction grating (See Figure 20.5.22). The latter have several advantages for direct spectroscopy, but here are a few specialized areas where prisms are better. Losses in grating through diffraction to unwanted orders are a nuisance in certain astronomical applications where every photon counts. Another example of an area where prisms are preferable is wavelength selection in multiwavelength lasers: a prism inside the laser resonator with adequate angular dispersion ensures that only one wavelength will be produced, and one scans through the available wavelengths by rotating the prisms at and away from the position of minimum deviation. The significance of the minimum deviation position is that the effects of vibrations and placement errors are least. Also, if the shape of the prism is isosceles, the resolving power will be a maximum at minimum deviation. The main formulas relating to dispersing prisms are as follows.

Spectroscopic resolving power:

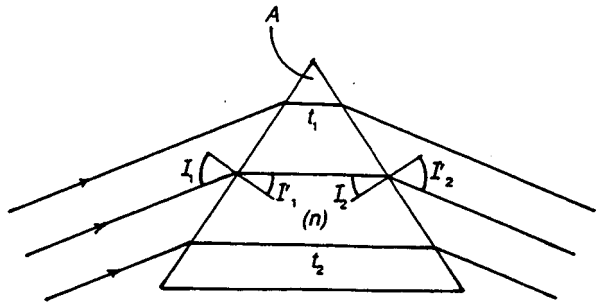


FIGURE 20.5.22

$$\lambda/\Delta\lambda = \left(t_1 - t_2\right) dn/d\lambda \tag{20.5.11}$$

where $t_1 - t_2$ is the difference between the path lengths in glass from one side of the beam to the other.

Angular dispersion:

$$dI_2'/d\lambda = \sin A\left(dn/d\lambda\right)/\left(\cos I_1' \cos I_2'\right) \tag{20.5.12}$$

$$= 2\sin\left(A/2\right)\left(dn/d\lambda\right)/\cos I_2 \tag{20.5.13}$$

at minimum deviation.

Spectrum line curvature:

$$1/\text{radius} = \left[\left(n^2 - 1\right)/nf\right]\sin A/\left(\cos I_1' \cos I_2'\right) \tag{20.5.14}$$

$$= \left[\left(n^2 - 1\right)/n^2 f\right]2\tan I_1 \tag{20.5.15}$$

at minimum deviation, where f is the focal length of the lens which brings the spectrum to a focus.

The spectrum line curvature refers to the observation that the image of the entrance slit of the spectroscope produced by placing a lens after the prism is actually parabolic. The parabola is convex toward the longer wavelengths. The reason the image is curved is that rays out of the principal plane of the prism are deviated more than rays in the plane, a straightforward consequence of the application of Snell's law. For rays with angle out of the plane the extra deviation can be parametrized by an additional contribution to the index of refraction given by

$$dn \approx \varepsilon^2\left(n^2 - 1\right)\left(2n\right) \tag{20.5.16}$$

If the length of the slit image is L, then $\varepsilon \approx L/(2f)$, where f is the focal length of the lens. Moreover, from Equation (20.5.12) we have

$$dI_2'/dn = \sin A/\left(\cos I_1' \cos I_2'\right) \tag{20.5.17}$$

$$= \left(2/n\right)\tan I_1 \tag{20.5.18}$$

at minimum deviation. The curvature of the slit image readily follows from these relations.

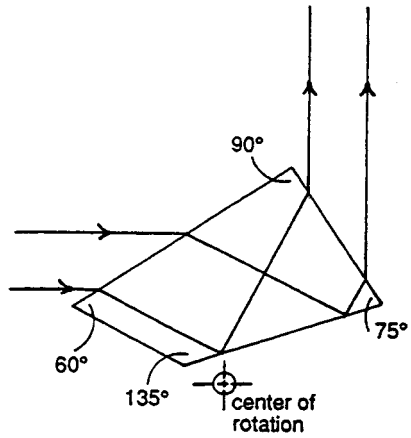


FIGURE 20.5.23

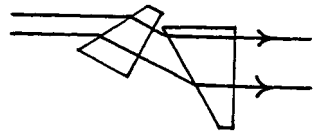


FIGURE 20.5.24

The typical dispersing prism of constant deviation shown in Figure 20.5.23 has the property that, if it is placed in a collimated beam, the wavelength which emerges at right angles to the incident beam is always at minimum deviation so that the spectrum is scanned by rotating the prism about a suitable axis such as the one indicated.

A prism used a long way from minimum deviation will expand or contract a collimated beam in one dimension. Figure 20.5.24 shows a pair of prisms used in this way to turn a laser beam of elliptical profile (from a diode laser) into a beam of circular profile by expanding it in the plane of the diagram only.

Nonimaging Optics

In one important respect conventional *image-forming* optical design is quite inefficient, that is, in merely concentrating and collecting light. This is well illustrated by an example taken from solar energy concentration (Figure 20.5.25). The flux at the surface of the sun (≈63 W/mm^2) falls off inversely with the square of the distance to a value ≈1.37 mW/mm^2 above the Earth's atmosphere or typically 0.8 to 1 mW/mm^2 on the ground. The second law of thermodynamics permits an optical device (*in principle*) to concentrate the dilute solar flux at Earth so as to attain temperatures up to but not exceeding that of the sun's surface. This places an upper limit on the solar flux density achievable on Earth and correspondingly on the concentration ratio of any optical device. From simple geometry, this limiting concentration ratio is related to the sun's angular size (2θ) by $C_{max} = 1/\sin^2\theta \approx 1/\theta^2$ (small angle approximation). We will call this thermodynamic limit the *sine law of concentration*. Therefore, since $\theta = 0.27°$ or 4.67 mrad, $C_{max} \approx 46{,}000$. When the target is immersed in a medium of refractive index n, this limit is increased by a factor n^2, $C_{max} = n^2/\sin^2\theta$. This means that a concentration of about 100,000 will be the upper limit for ordinary ($n \approx 1.5$) refractive materials. In experiments at the University of Chicago we actually achieved a solar concentration of 84,000 by using a nonimaging design with a refractive medium (sapphire). We would not even have come close using conventional designs, not for any fundamental reason but because imaging optical design is quite inefficient for delivering maximum concentration. For example, consider the paraboloidal mirror of a telescope used to concentrate sunlight at its focus (Figure 20.5.26). We can relate the concentration ratio to the angle 2ϕ subtended by the paraboloid at

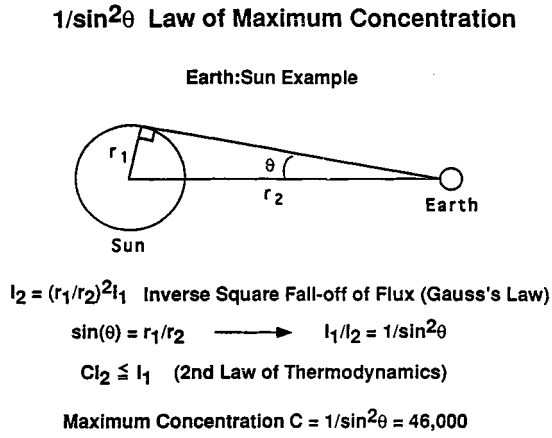


FIGURE 20.5.25

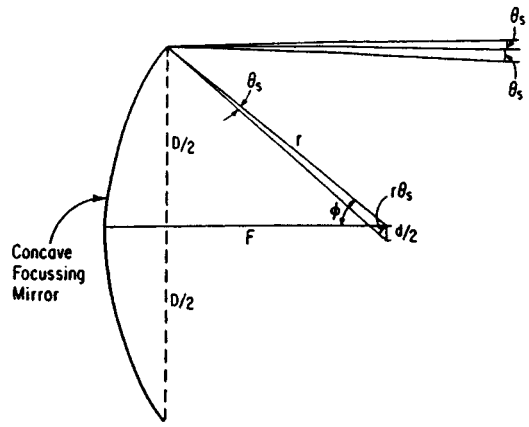
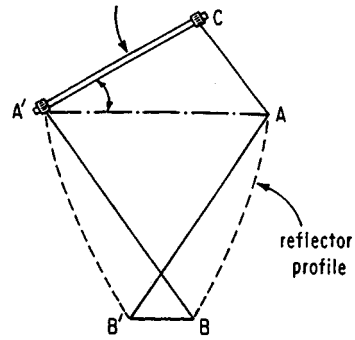


FIGURE 20.5.26

its focus and the sun's angular size ($2\theta_s$), $C = (\sin\phi\cos\phi/\theta_s)^2 = (^1/_4)\sin^2 2\phi/\theta_s^2$, where we have used the small angle approximation for θ_s. Notice that C is maximized at $\phi = \pi/4$, or $C = 1/(4\theta_s^2) = (^1/_4)\,C_{max}$. In fact, this result does not depend on the detailed shape of the paraboloid and would hold for any focusing mirror. One fares no better (and probably worse) with a lens (a refracting telescope), since the optimum paraboloid in the above example is equivalent in concentrating performance to a lens with focal ratio $f = 1$ which has been corrected for spherical aberration and coma. Such high-aperture lenses are typically complex structures with many components. The thermodynamic limit would require a lens with focal ratio $f = 0.5$ which, as every optical designer knows, is unattainable. The reason for this large shortfall is not hard to find. The paraboloid images perfectly on-axis, but has severe off-axis aberration (coma) which produces substantial image blurring and broadening. Nonimaging optics began in the mid 1960s with the discovery that optical systems could be designed and built that approached the theoretical limit of light collection (the sine law of concentration). The essential point is that requiring an image is unnecessarily restrictive when only concentration is desired. Recognition of this restriction and relaxation of the associated constraints led to the development of nonimaging optics. A nonimaging concentrator is essentially a "funnel" for light. Nonimaging optics departs from the methods of traditional optical design to develop instead techniques for maximizing the collecting power of concentrating elements and systems. Nonimaging designs exceed the concentration attainable with focusing techniques by factors of four or more and approach the theoretical limit (ideal concentrators). The key is simply to dispense with image-forming requirements in applications where no image is required.

Since its inception, the field has undergone three periods of rapid conceptual development. In the 1970s the "string" (see Figures 20.5.27 and 20.5.28) or "edge-ray" method (see Welford and Winston, 1989) was formulated and elaborated for a large variety of geometries. This development was driven by the desire to design wide-angle solar concentrators. It may be succinctly characterized as $\int n\, dl$ = constant along a string. (Notice that replacing "string" by "ray," Fermat's principle, gives all of imaging optics.) In the early 1980s, a second class of algorithms was found, driven by the desire to obtain ideally perfect solutions in three dimensions (3D) (The "string" solutions are ideal only in 2D, and as figures of revolution in 3D are only approximately ideal, though still very useful.) This places reflectors along the lines of flow of a radiation field set up by a radiating lambertian source. In cases of high symmetry such as a sphere or disk, one obtains ideal solutions in *both* 2D and 3D. The third period of rapid development has taken place only in the past several years; its implications and consequences are still in the process of being worked out. This was driven by the desire to address a wider class of problems in illumination that could not be solved by the old methods, for example, uniformly illuminating a plane (e.g., a table or a wall) by a lambertian light source (e.g., a fluorescent light). It is well known that the far-field illuminance from a lambertian source falls off with a power of the cosine of the radiating angle θ. For example, cylindrical radiators (such as a fluorescent lamp) produce a $\cos^2\theta$ illuminance on a distant plane, strip radiators produce a $\cos^3\theta$ illuminance, while circular disk radiators produce a $\cos^4\theta$ illuminance. But suppose one desires a predetermined far-field illuminance pattern, e.g., uniform illuminance? The old designs will not suffice; they simply transform a lambertian source radiating over 2π into a lambertian source radiating over a restricted set of angles. The limitation of the old designs is that they are too static and depend on a few parameters, such as the area of the beam A_1 and the divergence angle θ. One needs to introduce additional degrees of freedom into the nonimaging designs to solve a wider class of problems.



$$\int_W^{B'} \text{ndl} = \text{Constant. } AC + AB' = A'B + BB'$$
$$AC = AA' \sin\theta$$
$$AB' = A'B$$
$$\Rightarrow AA' \sin\theta = BB'$$

FIGURE 20.5.27

Edge-Ray Optics

One way to design nonimaging concentrators is to reflect the extreme input rays into the extreme output rays. We call this the "edge-ray method." An intuitive realization of this method is to wrap a string about both the light source and the light receiver, then allow the string to unwrap from the source and wrap around the receiver. In introducing the string picture, we follow an insight of Hoyt Hottel (Massachusetts Institute of Technology), who discovered many years ago that the use of strings tremendously simplified the calculation of radiative energy transfer between surfaces in a furnace. Our string is actually a "smart"

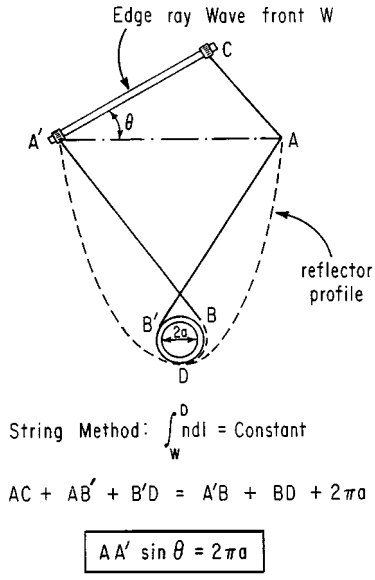


FIGURE 20.5.28

string; it measures the optical path length (ordinary length times the index of refraction) and refracts in accordance with Snell's law at the interface between different refracting materials. $\int n\,dl$ = constant. The locus traced out turns out to be the correct reflecting surface! Let's see how this works in the simplest kind of problem nonimaging optics addresses: collecting light over an entrance aperture AA' with angular divergence $\pm\theta$ and concentrating the light onto an exit aperture BB' (Figure 20.5.27). We attach one end of the string to the edge of the exit aperture B and the loop at the other end over a line WW' inclined at angle θ to the entrance aperture (this is the same as attaching to a "point at infinity"). We now unwrap the string and trace out the locus of the reflector taking care that string is taut and perpendicular to WW'. Then we trace the locus of the other side of the reflector. We can see with a little algebra that when we are done the condition for maximum concentration has been met. When we start unwrapping the string, the length is $AB' + BB'$; when we finish, the same length is $WA' + A'B$. But $WA' = AA'\sin\theta$, while $AB' = A'B$. So we have achieved $AA'/BB' = 1/\sin\theta$ which is maximum concentration! To see why this works, we notice that the reflector directs all rays at $\pm\theta$ to the edges BB' so that rays at angles $> \pm\theta$ are reflected out of the system and rejected. Now there is a conservation theorem for light rays called conservation of phase space or "etendue" which implies that if the rays at angles $> \pm\theta$ are rejected, then the rays that have angles $<\theta$ are all collected. Next we can try a more-challenging problem, where the "exit aperture" is a cylinder of radius a (Figure 20.5.28). Now we attach the string to a point on the cylinder and wrap it around the cylinder. When the string is unwrapped, we find that $AA'/2\pi a = 1/\sin\theta$ which is maximum concentration on the surface of the cylinder! Such designs are useful for solar-thermal concentrators since the typical receiver is a tube for carrying fluid. A solar plant for powering air conditioners that uses this design is shown in Figure 20.5.29. As already mentioned, there is an alternative method for designing "ideal" optical systems which bears little resemblance to the "string method" already described. We picture the aggregate of light rays traversing an optical system as a fluid flow in an abstract space called phase space.This is the "space" of ray positions and ray directions multiplied by the index of refraction, so it has twice the number of dimensions of ordinary space. By placing reflectors along the lines of flow of this vector field, nonimaging designs are generated. Flow-line designs are perfect in three dimensions, while the string designs rotated about an axis are not. On the other hand, the number of flow-line designs are much more restricted. For details see Welford and Winston (1989).

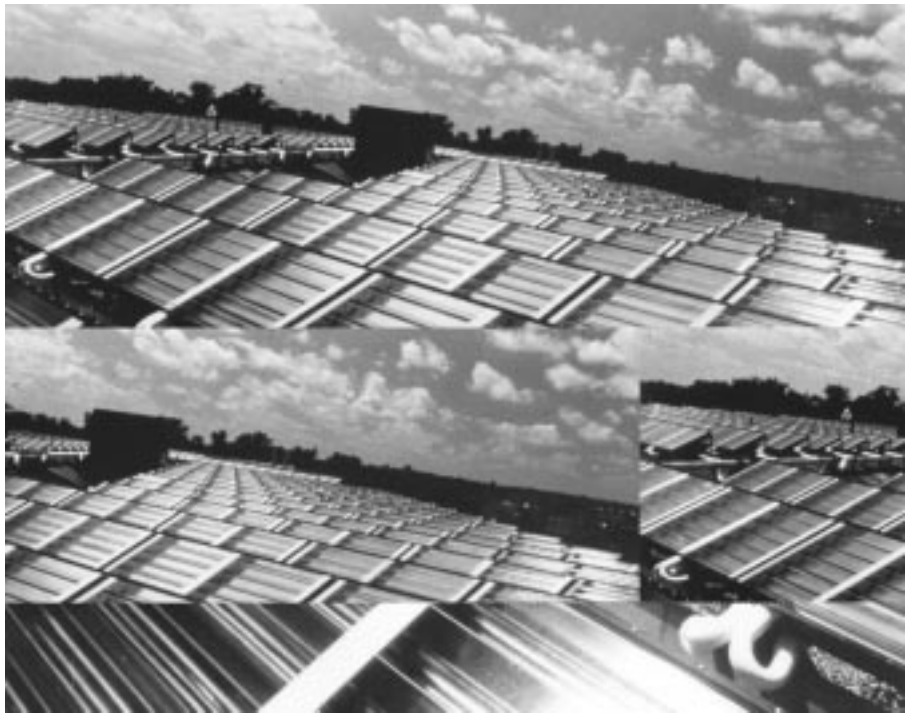


FIGURE 20.5.29

Lasers

Lasers are by now a ubiquitous light source with certain properties in addition to coherence and monochromaticity which have to be taken account of in some applications. Aside from many research applications of lasers, the HeNe laser at 632.8 nm wavelength has been available for several decades for alignment, surveying, and the like. But the explosive uses of lasers have come only recently with the advent of the solid-state diode laser. To appreciate the convenience of diode lasers one can draw upon the analogy between transistors and vacuum tubes. Inexpensive diode lasers are in widespread use in compact disk players, optical disk readers, and fiber-optics communications. The list of consumer applications is growing rapidly (e.g., laser pointers). In addition, diode lasers are used in arrays to optically drive (pump) more powerful lasers. The radiation that lasers emit can be highly coherent, and, except for the wavelength, of the same general character as the radiation from a radio frequency oscillator. We can identify four elements common to nearly all lasers (we follow the discussion in Mandel and Wolf, 1995, which should be consulted for details):

1. An optical resonator, generally formed by two or more mirrors;
2. A gain medium in which an inverted atomic population between the laser energy levels is established;
3. An optical pump or energy source to excite the gain medium;
4. A loss mechanism by which energy is dissipated or dispersed.

Figure 20.5.30 shows a typical form of laser, in which the resonator is a Fabry–Perot interferometer, and the amplifier is a gas plasma tube wherein a discharge is maintained. To reduce reflection losses from the plasma tube, its end windows are generally arranged at the Brewster angle for linearly polarized light at the laser frequency. The end mirrors are usually provided with multilayer dielectric coatings to make them highly reflecting. Of course, the output mirror needs to have its reflectivity, R, less than

100%, so that $(1 - R)$ is commonly the main source of the energy loss that has to be compensated by the gain medium. The cavity mirrors play the important role of feeding photons belonging to the laser modes back into the laser cavity. Most of the spontaneously emitted photons traveling in various other directions are lost. However, photons associated with a cavity resonance mode interact repeatedly with the atoms of the gain medium, and their number grows through stimulated emission, as illustrated in Figure 20.5.31. Once one mode is sufficiently populated, the probability for stimulated emission into that mode exceeds the spontaneous emission probability. In general, when the rate at which photons are fed into the optical cavity mode exceeds the rate at which they are lost from the cavity by the loss mechanism, the amplitude of the laser field starts to grow until a steady state is reached. At that point the rate of radiation by the laser equals the net rate at which energy is supplied. It is easy to see that this is achievable by an inverted population between the two atomic laser levels, with more atoms in the upper laser state than in the lower. If N_2, N_1 are the upper state and lower state populations, the rate of absorption of laser photons by the atomic system is proportional to N_1, and the rate of stimulated emission of laser photons by the system is proportional to N_2, with the same constant of proportionality for both. If N_2 exceeds N_1 sufficiently, all the radiation losses can be made good by the atomic system. It can be shown that the condition for laser action (in a single mode) is

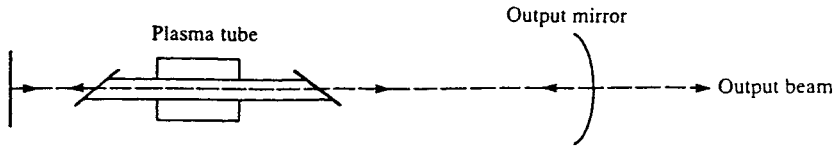


FIGURE 20.5.30 A simple form of laser.

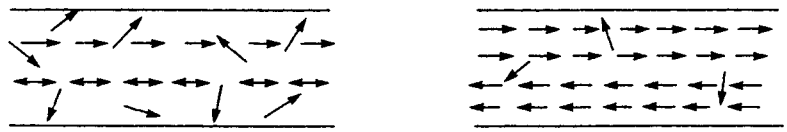


FIGURE 20.5.31 Illustration of the growth of the stimulated emission probability with the intensity of the mode.

$$N_2 - N_1 > 2A(1 - R)/\lambda^2 \tag{20.5.19}$$

where A is the cross-sectional area of the laser and λ is the wavelength.

We next summarize how the special properties of laser beams are to be taken into account in optical design. The well-known Gaussian intensity profile of laser beams persists if it is taken through a sequence of lenses along the axis, and at certain points that can be more-or-less predicted by paraxial optics a "focus" is formed. But when, as often happens, the convergence angle in the space in which this occurs is small, say, 1 mrad or less, some departures from the predictions of the paraxial optics occur. We shall examine these effects, since they are of importance in many of the systems already mentioned.

Gaussian Beams

In paraxial approximation the simplest form of a single-mode beam is the TEM_{00} Gaussian beam shown in Figure 20.5.32. Starting from the narrowest part, known as the waist, the beam diverges with spherical phase fronts. The complex amplitude at the waist has the form

$$A = A_0 \exp\left(-r^2/\omega_0^2\right) \tag{20.5.20}$$

where ω_0 is called the beam width and r is a radial coordinate.

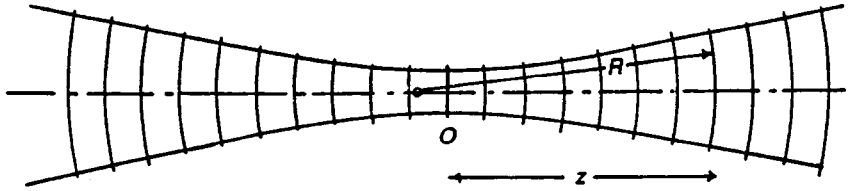


FIGURE 20.5.32

At a distance z along the beam in either direction, the complex amplitude is, apart from a phase factor,

$$A = \left(\omega_0/\omega\right) A_0 \exp\left(-r^2/\omega_0^2\right) \tag{20.5.21}$$

where ω is given by

$$\omega(z) = \omega_0\left[1 + \left(\lambda z/\pi\omega_0^2\right)\right]^{1/2} \tag{20.5.22}$$

At a distance z from the waist, the phase fronts have a radius of curvature R given by

$$R(z) = z\left[1 + \left(\pi\omega_0^2/\lambda z\right)^2\right] \tag{20.5.23}$$

The beam contour of constant intensity A_0^2/e^2 is a hyperboloidal surface of (small) asymptotic angle given by

$$\theta = \lambda/\pi\omega_0 \tag{20.5.24}$$

It can be seen that the centers of curvature of the phase fronts are not at the beam waist; in fact the phase front is plane at that point. The geometrical wavefronts are not exactly the same as true phase fronts, and if in this case we postulate that geometrical wavefronts should have their centers of curvature at the beam waist, we have an example of this. However, the difference is small unless the convergence angle is very small or, more precisely, when the Fresnel number of the beam is not much larger than unity:

$$\text{Fresnel number } = \omega^2/\lambda R \tag{20.5.25}$$

There is nothing special about Gaussian beams to cause this discrepancy between phase fronts and geometrical wavefronts; a similar phenomenon occurs with beams which are sharply truncated by the pupil (“hard-edged” beams). But it happens that it is less usual to be concerned with the region near the focus of a hard-edged beam of small Fresnel number, whereas Gaussian beams are frequently used in this way. Thus, Born and Wolf, *Principles of Optics* (1959) show that the phase front at the focus of a hard-edged beam is also plane, but with rapid changes of intensity and phase jumps across the zeros of intensity.

Tracing Gaussian Beams

If the beam is in a space of large convergence angle, say, greater than 10 mrad, it can be traced by ordinary paraxial optics, i.e., using the assumption that for all practical purposes the phase fronts are the same as geometrical wavefronts. In a space of small convergence angle it is necessary to propagate the beam between refracting surfaces by means of the proper Gaussian beam formulas and then use paraxial optics to refract (or reflect) the phase front through each surface in turn. To do this, we need

two more formulas to give the position, z, and size of the beam waist starting from the beam size and phase front curvature at an arbitrary position on the axis, i.e., given ω and R. These are

$$z = R\left[1 + \left(\lambda z/\pi\omega^2\right)\right]^{-1} \tag{20.5.26}$$

and

$$\omega_0 = \omega\left[1 + \left(\pi\omega^2/\lambda R\right)^2\right]^{-1/2} \tag{20.5.27}$$

Equations (20.5.22) to (20.5.27) enable a Gaussian beam to be traced through a sequence of refracting surfaces as an iterative process. Thus, starting from a beam waist of given size ω_0 (and angle given by Equation (20.5.24)), we move a distance z to the first refracting surface. At this surface the beam size ω is given by Equation (20.5.22) and the radius of curvature R of the phase front is given by Equation (20.5.23). The radius of curvature R' of the refracted phase front is obtained by paraxial optics using the equation of Figure 20.5.7 and taking R and R' as the conjugate distances l and l'. Then the position and size of the new beam waist are found from Equations (20.5.26) and (20.5.27). These procedures can be carried through all the refracting surfaces of the optical system.

It can be seen from Equation (20.5.26) that z and R are substantially equal when $\lambda R/\pi\omega^2$ is very small. When this is so, there is no need to use these special equations for transferring between surfaces; the iterative equations in Figure 20.5.9 can be used, with the understanding that the paraxial convergence angle u is the equivalent of the asymptotic angle θ in Equation (20.5.24).

There are no simple equations for hard-edged beams corresponding to Equations (20.5.23) to (20.5.27) for use with very small convergence angles. Numerical calculations of the beam patterns near focus have been published for some special cases, and these show, as might be expected, very complex structures near the "focal" region; however, that is defined.

Truncation of Gaussian Beams

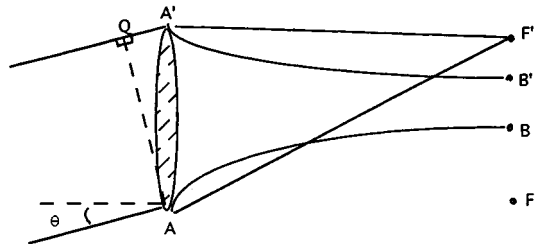
The theoretical origin of the Gaussian beam is as a paraxial solution of the Helmholtz equation, i.e., a solution concentrated near one straight line, the axis, but although most of the power is within the region near the axis the solution is nonzero, although very small, at an infinite distance from the axis. Thus, the Gaussian profile is truncated when it passes through any aperture of finite diameter — e.g., a lens mount, an aperture stop, or even the finite-diameter end mirror of a laser resonator — after which it is no longer Gaussian and the above equations are no longer valid! In practice, this need not be a problem, for if the radius of the aperture is 2ω, the complex amplitude is down to 1.8% of its value at the center and the intensity is 0.03% of its value at the center. Thus, it is often assumed that an aperture of radius 2ω has no significant effect on the Gaussian beam, and this assumption is adequate for many purposes, although not all.

Sometimes it is useful to truncate a Gaussian beam deliberately, i.e., turn it into a hard-edged beam, by using an aperture of radius less than, say, ω. In this way an approximation to the Airy pattern is produced at the focus instead of a Gaussian profile waist, and this pattern may be better for certain purposes, e.g., for printers where the spot must be as small as possible for an optical system of given numerical aperture.

Gaussian Beams and Aberrations

In principle, a Gaussian beam is a paraxial beam, from the nature of the approximations made in solving the Helmholtz equation. However, Gaussian beams can be expanded to large diameters simply by letting them propagate a large distance, and they can acquire aberrations by passing through an aberrating lens or mirror system. The beam is then no longer Gaussian, of course, in the strict sense, but we stress that the conventional optical design idea involving balancing and reduction of aberrations can be applied to

systems in which Gaussian beams are to propagate. For example, a *beam expander*, of which one form in shown in Figure 20.5.33, is an afocal system intended to do what its name implies: if it has aberrations



Lens images point at infinity to F'
therefore QA' + A'F' = AF' (Fermat)
AF' − A'F' = AF' − AF = constant (by property of hyperbola)
= BB'
QA' = AA' sinθ
therefore AA' sinθ = BB'

$(AA'/BB')^2 = 1/\sin^2\theta$

FIGURE 20.5.33

as an afocal system, the output beam from a Gaussian input beam will not have truly plane or spherical phase fronts.

Non-Gaussian Beams from Lasers

Not all lasers produce Gaussian beams, even ignoring the inevitable truncation effects of resonator mirrors. Some gas lasers (e.g., helium-neon at any of its lasing wavelengths) produce Gaussian beams when they are in appropriate adjustment, but they can produce off-axis modes with more structure than a Gaussian beam. Other gas lasers (e.g., copper vapor lasers) produce beams with a great many transverse modes covering an angular range of a few milliradians in an output beam perhaps 20 mm across. Some solid-state lasers, e.g., ruby, may produce a very non-Gaussian beam because of optical inhomogeneities in the ruby. Laser diodes, which as already mentioned are becoming increasingly useful as very compact coherent sources, whether cw or pulsed, produce a single strongly divergent transverse mode which is wider across one direction than the other. This mode can be converted into a circular section of approximately Gaussian profile by means of a prism system, as in Figure 20.5.24.

References

Born, M. and Wolf, E. 1959. *Principles of Optics*, Pergamon Press, Elmsford, NY.
Welford, W.T. 1986. *Aberrations of Optical Systems*, Adam Hilger, Bristol.
Welford, W.T. and Winston, R. 1989. *High Collection Nonimaging Optics,* Academic Press, New York.
Mandel, L. and Wolf, E. 1995. *Optical Coherence and Quantum Optics,* Cambridge University Press., New York.

20.6 Water Desalination

Noam Lior

Introduction and Overview

Water desalination is a process that separates water from a saline water solution. The natural water cycle is the best and most prevalent example of water desalination. Ocean waters evaporate due to solar heating and atmospheric influences; the vapor consisting mostly of fresh water (because of the negligible volatility of the salts at these temperatures) rises buoyantly and condenses into clouds in the cooler atmospheric regions, is transported across the sky by cloud motion, and is eventually deposited back on the earth surface as fresh water rain, snow, and hail. The global freshwater supply from this natural cycle is ample, but many regions on Earth do not receive an adequate share. Population growth, rapidly increasing demand for fresh water, and increasing contamination of the available natural fresh water resources render water desalination increasingly attractive. Water desalinaiton has grown over the last four decades to an output of about 20 million m^3 of fresh water per day, by about 10,000 sizeable land-based water desalination plants.

The salt concentration in the waters being desalted ranges from below 100 ppm wt. (essentially fresh water, when ultrapure water is needed), through several thousand parts per million (brackish waters unsuitable for drinking or agricultural use) and seawater with concentrations between 35,000 and 50,000 ppm. Official salt concentration limits for drinkable water are about 1000 ppm, and characteristic water supplies are restricted to well below 500 ppm, with city water in the United States being typically below 100 ppm. Salinity limits for agricultural irrigation waters depend on the type of plant, cultivation, and soil, but are typically below 2000 ppm.

Many ways are availiable for separating water from a saline water solution. The oldest and still prevalent desalination process is distillation. The evaporation of the solution is effected by the addition of heat or by lowering of its vapor pressure, and condensation of these vapors on a cold surface produces fresh water. The three dominant distillation processes are multistage flash (MSF), multieffect (ME), and vapor compression (VC). Until the early 1980s the MSF process was prevalent for desalination. Now membrane processes, especially reverse osmosis (RO), are economical enough to have taken about one third of the market. In all membrane processes separation occurs due to the selective nature of the permeability of a membrane, which permits, under the influence of an external driving force, the passage of either water or salt ions but not of both. The driving force may be pressure (as in RO), electric potential (as in electrodialysis, ED), or heat (as in membrane distillation, MD). A process used for low-salinity solutions is the well-known ion exchange (IE), in which salt ions are preferentially adsorbed onto a material that has the required selective adsorption property and thus reduce the salinity of the water in the solution.

The cost of desalted water is comprised of the capital cost of the plant, the cost of the energy needed for the process, and the cost of operation and maintenance staff and supplies. In large seawater desalination plants the cost of water is about \$1.4 to \$2/m^3, dropping to less than \$1/m^3 for desalting brackish water. A methodology for assessing the economic viability of desalination in comparison with other water supply methods is described by Kasper and Lior (1979). Desalination plants are relatively simple to operate, and progress toward advanced controls and automation is gradually reducing operation expenses. The relative effect of the cost of the energy on the cost of the fresh water produced depends on local conditions, and is up to one half of the total.

The boiling point of a salt solution is elevated as the concentration is increased, and the **boiling point elevation** is a measure of the energy needed for separation. Thermodynamically reversible separation defines the minimal energy requirement for that process. The minimal energy of separation W_{min} in such a process is the change in the Gibbs free energy between the beginning and end of the process, ΔG. The minimal work when the number of moles of the solution changes from n_1 to n_2 is thus

$$W_{\min} = \int_{n_1}^{n_2} (\Delta G)\, dn_W \tag{20.6.1}$$

The minimal energy of separation of water from seawater containing 3.45 wt.% salt, at 25°C, is 2.55 kJ/(kg fresh water) for the case of zero fresh water recovery (infinitesimal concentration change) and 2.91 kJ/(kg fresh water) for the case of 25% freshwater recovery. $W_{\min}$ is, however, severalfold smaller than the energy necessary for water desalination in practice. Improved energy economy can be obtained when desalination plants are integrated with power generation plants (Aschner, 1980). Such dual-purpose plants save energy but also increase the capital cost and complexity of operation.

Two aspects of the basically simple desalination process require special attention. One is the high-corrosivity of seawater, especially pronounced in the higher-temperature destillation processes, which requires the use of corrosion-resistant expensive materials. Typical materials in use are copper–nickel alloys, stainless steel, titanium, and, at lower temperatures, fiber-reinforced polymers (George et al., 1975). Another aspect is scale formation (Glater et al., 1980; Heitman, 1990). Salts in saline water, particularly calcium sulfate, magnesium hydroxide, and calcium carbonate, tend to precipitate when a certain temperature and concentration are exceeded. The precipitate, often mixed with dirt entering with the seawater and with corrosion products, will gradually plug up pipes, and when depositing on heat transfer surfaces reduces heat transfer rates and thus impairs plant performance. While the ambient-temperature operation of membrane processes reduces scaling, membranes are much more susceptible not only to minute amounts of scaling or even dirt, but also to the presence of certain salts and other compounds that reduce their ability to separate salt from water. To reduce corrosion, scaling, and other problems, the water to be desalted is pretreated. The pretreatment consists of filtration, and may inlude removal of air (deaeration), removal of CO_2 (decarbonation), and selective removal of scale-forming salts (softening). It also includes the addition of chemicals that allow operation at higher temperatures without scale deposition, or which retard scale deposition and/or cause the precipitation of scale which does not adhere to solid surfaces, and that prevent foam formation during the desalination process.

Saline waters, including seawater, contain, besides a variety of inorganic salts, also organic materials and various particles. They differ in composition from site to site, and also change with time due to both natural and person-made causes. Design and operation of desalination plants requires good knowledge of the saline water composition and properties (Fabuss, 1980; Heitman, 1991).

The major water desalination processes that are currently in use or in advanced research stages are concisely described below. Information on detailed modeling can be found in the references.

Distillation Processes

Multistage Flash Evaporation (MSF)

Almost all of the large desalination plants use the MSF process shown schematically in Figure 20.6.1. A photo of an operating plant is shown in Figure 20.6.2. The seawater feed is preheated by internal heat recovery from condensing water vapor during passage through a series of stages, and then heated to its top temperature by steam generated by an external heat source. The hot seawater then flows as a horizontal free-surface stream through a series of "stages," created by vertical walls which separate the vapor space of each stage from the others. These walls allow the vapor space of each stage to be maintained at a different pressure, which is gradually decreased along the flow path due to the gradually decreasing temperature in the condenser/seawater-preheater installed above the free stream. The seawater is super-heated by a few degrees celsius relative to the vapor pressure in each stage it enters, and consequently evaporates in each stage along its flow path. The latent heat of the evaporation is supplied by equivalent reduction of the sensible heat of the evaporating water, thus resulting in a gradual lowering of the stream temperature. The evaporation is vigorous, resulting in intensive bubble generation and growth with accompanying stream turbulence, a process known as **flash evaporation** (Lior and Greif, 1980; Miyatake et al., 1992; 1993). One of the primary advantages of the MSF process is the fact that evaporation occurs

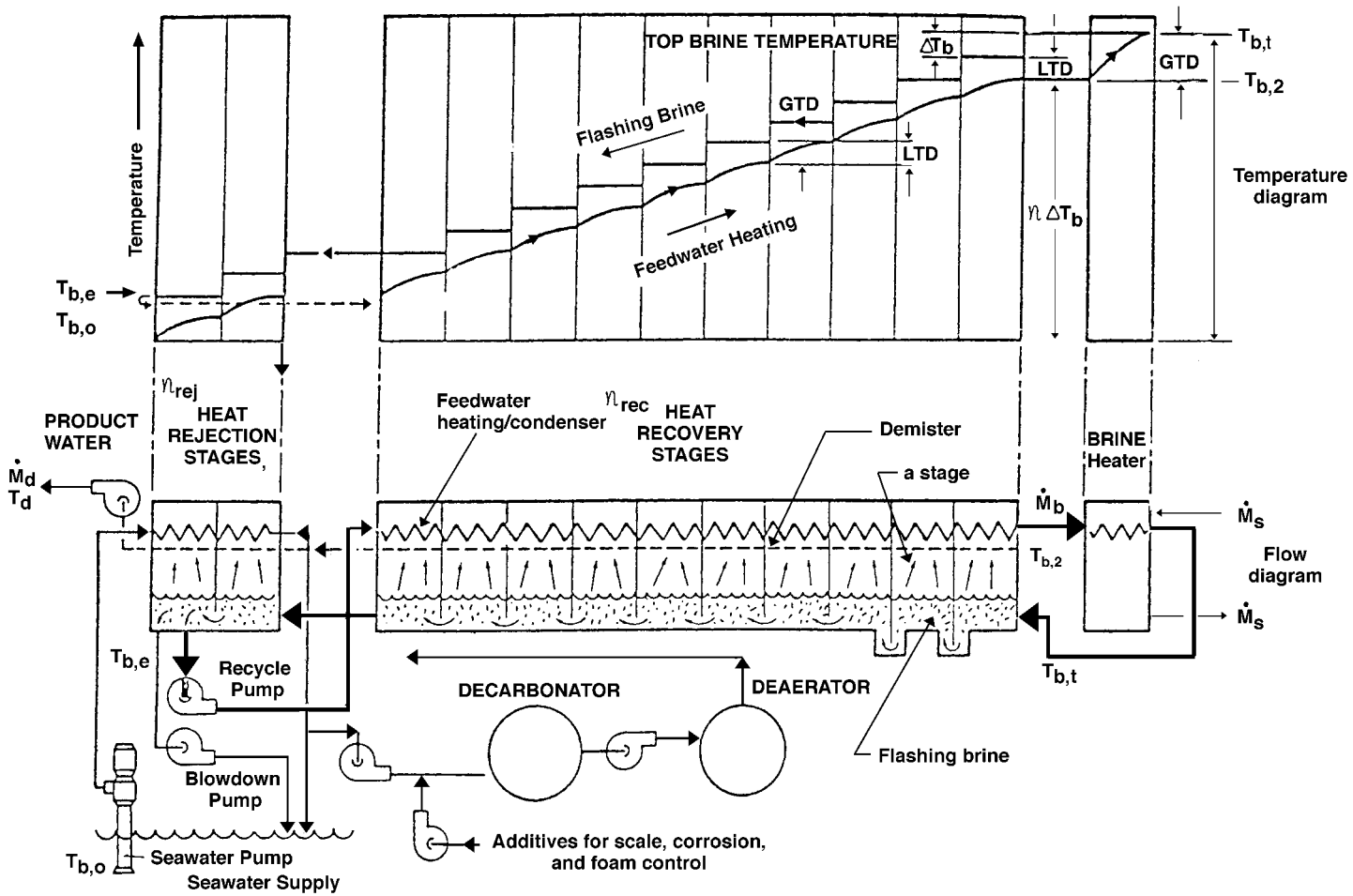


FIGURE 20.6.1 Schematic flow and temperature diagram of the MSF process, for a recirculation type plant.

FIGURE 20.6.2 One of the six units of the 346,000 m^3/day MSF desalination plant Al Taweelah B in Abu Dhabi, United Arab Emirates. (Courtesy of Italimpianti S. p. A.) It is a dual-purpose plant, composed of six identical power and desalination units. Five of the six boilers are seen in the background. The desalination units were in 1996 the largest in the world. They have 17 recovery and 3 reject stages and a performance ratio (PR) of 8.1. The plant also produces 732 MWe of power.

from the saline water stream and not on heated surfaces (as in other distillation processes such as submerged tube and ME evaporation) where evaporation typically causes scale depostition and thus gradual impairment of heat transfer rates. Also, the fact that the sensible heat of water is much smaller than its latent heat of evaporation, where the specific heat $c_p = 4.182$ kJ/kg/°C change of water temperature and the latent heat is $h_{fg} = 2378$ kJ/kg, and the fact that the top temperature is limited by considerations of scaling and corrosion, dictate the requirement for a very large flow rate of the evaporating stream. For example (in the following, the subscripts b, d, and s refer to brine distillate, and steam, respectively), operating between a typical top temperature $T_{b,t}$ of 90°C at the inlet to the evaporator and an exit temperature $T_{b,e}$ of 40°C corresponding to the ambient conditions, the overall temperature drop of the evaporating stream if 50°C. By using these values, the heat balance between the sensible heat of the water stream, flowing at a mass flow rate $\dot{m}_b$, and the latent heat needed for generating water vapor (distillate) at a mass flow rate $\dot{m}_d$ is

$$\left(\dot{m}_b - \dot{m}_d\right) c_p \left(T_{b,t} - T_{b,e}\right) \approx \dot{m}_d h_{fg} \tag{20.6.2}$$

which yields the brine-to-product mass flow ratio as

$$\frac{\dot{m}_b}{\dot{m}_d} = \frac{h_{fg}}{c_p\left(T_{b,t} - T_{b,e}\right)} + 1 = \frac{2378}{(4.182)(50)} + 1 = 12.37 \tag{20.6.3}$$

Therefore, 12.37 kg of saline water are needed to produce 1 kg of distillate. This high flow rate incurs corresponding pumping equipment and energy expenses, sluggish system dynamics, and, since the stream level depth is limited to about 0.3 to 0.5 m for best evaporation rates, also requires large evaporator vessels with their associated expense.

The generated water vapor rises through a screen (“demister”) placed to remove entrained saline water droplets. Rising further, it then condenses on the condenser tube bank, and internal heat recovery is

achieved by transferring its heat of condensation to the seawater feed that is thus being preheated. This internal heat recovery is another of the primary advantages of the MSF process. The energy performance of distillation plants is often evaluated by the *performance ratio*, PR, typically defined as

$$\mathrm{PR} \equiv \frac{\dot{m}_d}{\dot{m}_s} \tag{20.6.4}$$

where $\dot{m}_s$ is the mass flow rate of heating steam. Since the latent heat of evaporation is almost the same for the distillate and the heating steam, PR is also the ratio of the heat energy needed for producing one unit mass of product (distillate) to the external heat actually used for that purpose. Most of the heating of the brine stream to the top temperature $T_{b,t}$ is by internal heat recovery, and as seen in Figure 20.6.1, the external heat input is only the amount of heat needed to elevate the temperature of the preheated brine from its exit from the hottest stage at $T_{b,2}$ to $T_{b,t}$. Following the notation in Figure 20.6.1, and using heat balances similar to that in Equation (20.6.3) for the brine heater and flash evaporator, the PR can thus also be defined as

$$\mathrm{PR} = \frac{\dot{m}_b\left(\overline{c_{p,b}}\right)_{e\to t}\left(T_{b,t} - T_{b,e}\right)\Big/h_{fg,b}}{\dot{m}_b\left(\overline{c_{p,b}}\right)_{2\to t}\left(T_{b,t} - T_{b,2}\right)\Big/h_{fg,s}} \approx \frac{T_{b,t} - T_{b,e}}{T_{b,t} - T_{b,2}} \tag{20.6.5}$$

where $(\overline{c_{p,b}})_{e\to t}$ and $(\overline{c_{p,b}})_{2\to t}$ are the specific heats of brine, the first averaged over the temperature range $T_{b,e} \to T_{b,t}$ and the second over $T_{b,2} \to T_{b,t}$. The rightmost expression in Equation (20.6.5) is nearly correct because the specific heat of the brine does not change much with temperature, and the latent heat of evaporation of the brine is nearly equal to the latent heat of condensation of the heating steam. It is obvious from Equation (20.6.5) that PR increases as the top heat recovery temperature $T_{b,2}$ (at the exit from the condenser/brine-preheater) increases. It is also obvious (even from just examining Figure 20.6.1) that increasing the number of stages (matched with a commensurate increase in condenser heat transfer area and assuming no significant change in the overall heat transfer coefficient) for a given $T_{b,t}$, will raise the flash evaporator inlet temperature $T_{b,3}$, which will lead to a rise in $T_{b,2}$ and thus also in the PR.

Assuming that the temperature drop of the flashing brine, ΔT_b, is the same in each stage, the relationship between the number of stages (n) and the performance ratio is

$$\mathrm{PR} = \frac{1}{\dfrac{\mathrm{LTD}}{T_{b,t} - T_{b,e}} + \dfrac{1}{n}} \tag{20.6.6}$$

where LTD is the lowest temperature difference between the flashed vapor and the heated feedwater, in each stage (Figure 20.6.1). Equation (20.6.6) shows that increasing the number of stages increases the PR. This implies that more heat is then recovered internally, which would thus require a larger condenser/brine-preheater heat transfer area. The required heat transfer area, A, *per unit mass of distillate produced* for the entire heat recovery section (composed of n_{rec} stages), and taking average values of the overall vapor-to-feedwater heat transfer coefficient U and LMTD, is thus

$$A = n_{\mathrm{rec}} A_n = n_{\mathrm{rec}} \frac{h_{b,fg}}{U(\mathrm{LMTD})} \tag{20.6.7}$$

LMTD, the log-mean temperature difference between the vapor condensing on the tubes and the heated brine flowing inside the tubes, for an average stage is

$$\mathrm{LMTD} = \frac{\mathrm{GTD} - \mathrm{LTD}}{\ln \dfrac{\mathrm{GTD}}{\mathrm{LTD}}} = \frac{\left(T_{b,t} - T_{b,2}\right) - \mathrm{LTD}}{\ln\left(\dfrac{T_{b,t} - T_{b,2}}{\mathrm{LTD}}\right)} \tag{20.6.8}$$

where GTD is the greatest temperature difference between the flashing brine and the brine heated in the condenser. The size of the heat transfer area per unit mass of distillate is

$$A = \frac{h_{fg,b}}{U} \frac{n_{\mathrm{rec}}}{\left(T_{b,t} - T_{b,e}\right)} \ln\left(\frac{n_{\mathrm{rec}}}{n_{\mathrm{rec}} - PR}\right) \tag{20.6.9}$$

Examination of this equation will show that the required heat transfer area for the heat recovery section per unit mass of distillate produced, A, increases significantly when PR is increased, and decreases slightly as the number of heat recovery stages, n_{rec}, is increased.

The MSF plant shown in Figure 20.6.1 is of the *recirculation* type, where not all of the brine stream emerging from the last evaporation stage is discharged from the plant (as it would have been in a *once-through* type of plant). A fraction of the emerging brine is mixed with pretreated seawater and recirculated into the condenser of the heat recovery section of the plant. Since only a fraction of the entire stream in this configuration is new seawater, which needs to be pretreated (removal of air and CO_2, i.e., deaeration and decarbonation, and the addition of chemicals that reduce scale deposition, corrosion, and foaming), the overall process cost is reduced. The recirculation plant is also easier to control than the once-through type.

While most of the energy exchange in the plant is internal, steady-state operation requires that energy in an amount equal to all external energy input be also discharged from the plant. Consequently, the heat supplied in the brine heater (plus any pumping energy) is discharged in the heat rejection stages section of the plant (Figure 20.6.1). Assuming an equal temperature drop in each stage, and that the pumping energy can be neglected relative to the heat input in the brine heater, indicates that the ratio of the number of the heat-recovery to heat-rejection stages is approximately equal to the performance ratio PR.

Further detail about MSF desalination can be found in Steinbruchel and Rhinesmith, (1980) and Khan (1986). A detailed design of an MSF plant producing 2.5 million gals. of freshwater per day was published by the U.S. government (Burns and Roe, 1969).

Multi-Effect Distillation (ME)

The principle of the ME distillation process is that the latent heat of condensation of the vapor generated in one effect is used to generate vapor in the next effect, thus obtaining internal heat recovery and good energy efficiency. Several ME plant configurations, most prominently the horizontal tube ME (HTME, shown in Figure 20.6.3) and the vertical tube evaporator (VTE, shown schematically in Figure 20.6.4) are in use. In the HTME, vapor is circulated through a horizontal tube bundle, which is subjected to an external spray of somewhat colder saline water. The vapor flowing in these spray-cooled tubes condenses, and the latent heat of condensation is transferred through the tube wall to the saline water spray striking the exterior of the tube, causing it to evaporate. The vapor generated thereby flows into the tubes in the next effect, and the process is repeated from effect to effect.

In the VTE the saline water typically flows downward inside vertical tubes and evaporates as a result of condensation of vapor coming from a higher temperature effect on the tube exterior. While internal heat recovery is a feature common to both MSF and ME processes, there are at least three important differences between them. One is that evaporation in the ME process occurs on the heat transfer surfaces (tubes), while in the MSF process it takes place in the free stream. This makes the ME process much more susceptible to scale formation. At the same time, the heat transfer coefficient between the vapor and the preheated brine is lower in the MSF process because the heated brine does not boil. In the ME

FIGURE 20.6.3 Two HTME desalination units, each producing 5000 m^3/day, in St. Croix, U.S. Virgin Islands. (Courtesy of I.D.E. Technologies Ltd.)

process it does boil, and it is well known that boiling heat transfer coefficients are significantly higher than those where the heating does not result in boiling. In using direct transfer of latent heat of condensation to latent heat of evaporation, instead of sensible heat reduction to latent heat of evaporation as in MSF, the ME process requires a much smaller brine flow than the MSF. Limiting brine concentration in the last effect to about three times that of the entering seawater, for example, requires a brine flow of only about 1.5 times that of the distillate produced. At the same time, a pump (although much smaller than the two pumps needed in MSF) is needed for each effect.

The PR of ME plants is just sightly lower than the number of effects, which is determined as an optimized compromise between energy efficiency and capital cost. Six effects are typical, although plants with as many as 18 effects have been built.

Further detail about ME desalination can be found in Steinbruchel and Rhinesmith (1980) and Standiford, (1986a).

Vapor Compression Distillation (VC)

As stated earlier, the vapor pressure of saline water is lower than that of pure water at the same temperature, with the pressure difference proportional to the boiling point elevation of the saline water. Desalination is attained here by evaporating the saline water and condensing the vapor on the pure water. Therefore, the pressure of the saline water vapor must be raised by the magnitude of that pressure difference, plus some additional amount to compensate for various losses. This is the principle of the vapor compression desalination method. Furthermore, as shown in Figure 20.6.5, the heat of condensation of the compressed vapor is recovered internally by using it to evaporate the saline water. Additional heat recovery is obtained by transferring heat from the concentrated brine effluent and the produced freshwater (which need to be cooled down to as close to ambient conditions as possible anyway) to the feed saline water which is thus preheated. The schematic flow diagram in Figure 20.5.5 shows a design in which the preheated seawater is sprayed onto a bank of horizontal tubes carrying condensing compressed vapor at a temperature higher than that of the seawater. The spray thus evaporates on contact with the exterior of the tube and provides the cooling needed for the internal condensation. Considering the fact that the energy required for vapor compression over a typical overall temperature difference of 4°C and a vapor compressor efficiency of 0.8 is 34 kJ/kg (easily calculated from an enthalpy balance), and that the latent

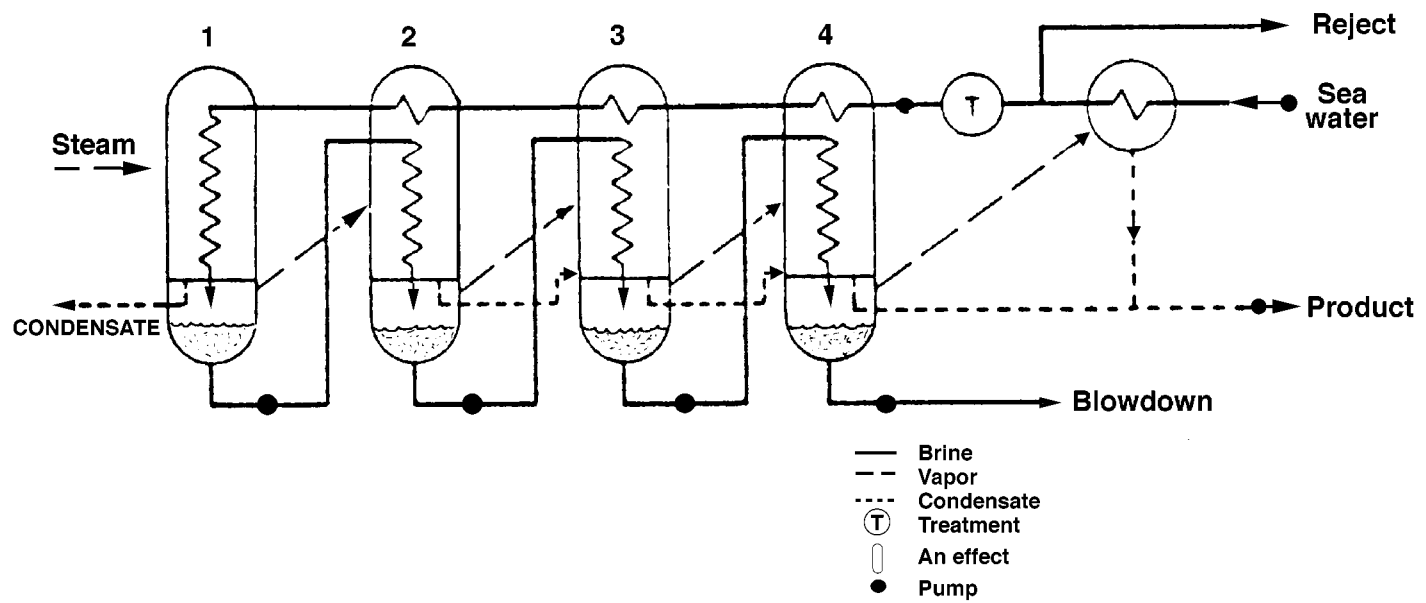


FIGURE 20.6.4 Simplified schematic flow diagram of a typical four-effect VTE desalination plant.

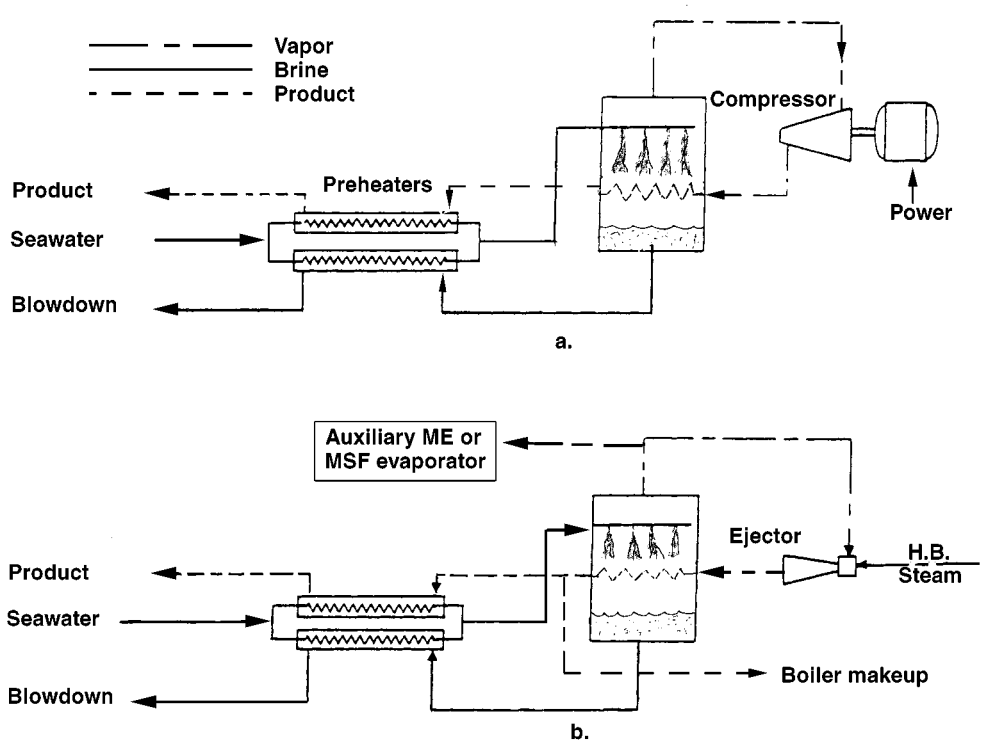


FIGURE 20.6.5 Schematic flow diagram of a basic horizontal-tube VC desalination plant (a) with mechanical, motor-driven compressor; (b) with a thermo-compressor, using an ejector.

heat of condensation is about 2400 kJ/kg, one can see that a small amount of compression energy enables a large amount of heat to be used internally for desalination. One can thus envisage the VC plant as a large flywheel, wheeling a large amount of energy around at the expense of a small amount needed for sustaining its motion.

The compressor can be driven by electric motors, gas or steam turbines, or internal combustion (usually diesel) engines. The compressor can also be a steam-driven ejector (Figure 20.6.5b), which improves plant reliability because of its simplicity and absence of moving parts, but also reduces its efficiency because an ejector is less efficient than a mechanical compressor. In all of the mentioned thermally driven devices, turbines, engines, and the ejector, the exhaust heat can be used for process efficiency improvement, or for desalination by an additional distillation plant.

Figure 20.6.6 shows a multi-effect VC plant. Using more than a single effect reduces the vapor volume that needs to be compressed. Furthermore, the overall required heat transfer area is also decreased because much of the single-phase heat transfer process in the preheater of the single-effect plant is replaced by the high-heat-transfer condensation–evaporation processes in the effects. Although the ME feature also increases the required compression ratio, the cost of produced water is reduced overall.

Further detail about VC desalination can be found in Steinbruchel and Rhinesmith (1980), Khan (1986), and Standiford, (1986b).

Solar Distillation

The benefits of using the nonpolluting and practically inexhaustible energy of the sun for water desalination are obvious. Furthermore, many water-poor regions also have a relatively high solar flux over a large fraction of the time. The major impediment in the use of solar energy is economical: the diffuse nature of solar energy dictates the need for constructing a large solar energy collection area. For example,

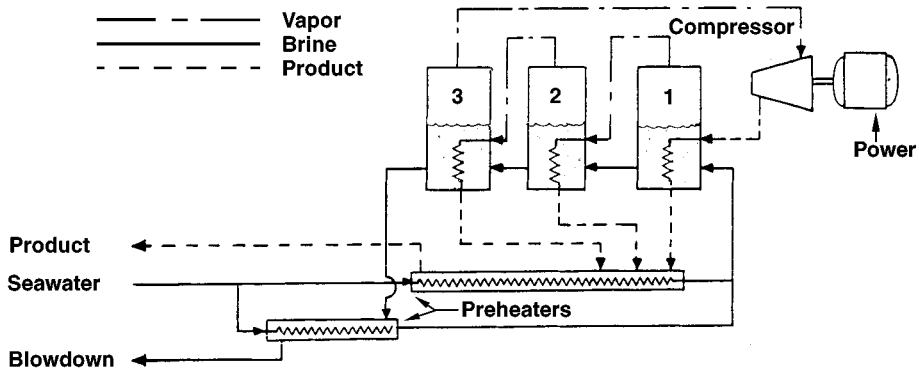


FIGURE 20.6.6 Schematic flow diagram of a ME vapor compression submerged-tube desalinaton plant with three effects.

assuming a single-effect solar still efficiency of 50% (which is the upper practical limit for conventional designs), the still would produce at most about 3.5 to 4.8 kg fresh water per m^2 per day, or a 208 to 286 m^2 solar still would be required to produce 1 m^3 of fresh water per day. More realistic still efficiencies increase the area requirement about twofold.

Shown in Figure 20.6.7, a typical solar still consists of a saline water container in which the water is exposed to the sun and heated by it. The temperature rise to above ambient causes net evaporation of the saline water, thus separating pure water vapor from the solution. The vapor condenses on the colder cover, and this destilled water flows to collection troughs.

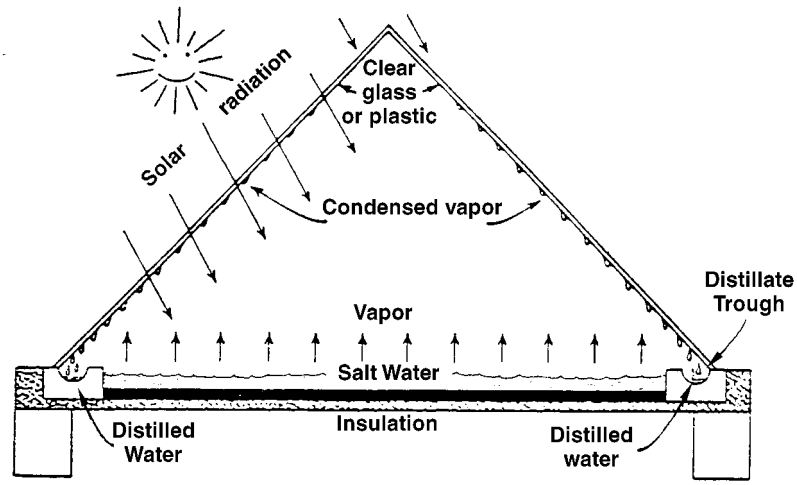


FIGURE 20.6.7 A typical basin-type solar still.

Solar stills of the type depicted in Figure 20.6.7, in many sizes and constructional variants, have been built and used successfully in many countries in the world. They are simple, easy to construct, reliable, and require very little maintenance although in some regions the covers must be cleaned frequently from accumulated dust or sand.

Since the heat of condensation in single-effect stills of the type shown in Figure 20.6.7 is lost to the ambient, more-energy-efficient operation can obviously be achieved in a multi-effect design, where the

heat of condensation is used to evaporate additional saline water. A number of such stills were built and tested successfully, but are not commercially competitive yet.

Solar stills integrate the desalination and solar energy collection processes. Another approach to solar desalination is to use separately a conventional desalination process and a suitable solar energy supply system for it. Any compatible desalination and solar energy collection processes could be used. Distillation, such as MSF or ME, can be used with heat input from solar collectors, concentrators, or solar ponds (Hoffman, 1992; Glueckstern, 1995). Net average solar energy conversion efficiencies of solar collectors (Rabl, 1985; Lior, 1991) are about 25% and of solar ponds (Lior, 1993) about 18%, similar to the efficiencies of solar stills, but the MSF or ME plants can operate at preformance ratios of 10 or more, thus basically increasing the freshwater production rate by at least tenfold, or reducing the required solar collection area by at least tenfold for the same production rate.

Solar or wind energy can also be used for desalination processes that are driven by mechanical or electrical power, such as VC, RO, and ED. The solar energy can be used to generate the required power by a variety of means, or photovoltaic cells can be used to convert solar energy to electricity directly.

Freeze Desalination

It is rather well known that freezing of saline water solutions is an effetive separation process in that it generates ice crystals that are essentially salt-free water, surrounded by saline water of higher concentration. This process requires much less energy than distillation, and the problems of corrosion and scaling are markedly reduced due to the much lower operating temperatures. Several pilot plants were constructed and have proven concept viability. Nevertheless, the process has not yet reached commercial introduction for several reasons, such as the difficulty in developing efficient and economical compressors for vapor with the extremely high specific volume at the low process pressure, and difficulties in maintaining the vacuum system leak free and in effecting reliable washing of the ice crystals. A review of freeze desalination processes is given by Tleimat (1980).

Membrane Separation Processes

Reverse Osmosis (RO)

Separation of particulate matter from a liquid by applying pressure to the liquid and passing it through a porous membrane, whereby particles larger than the pore size remain on the upstream side of the membrane and the liquid flows to its downstream side, is well known as *filtration*. Semipermeable very dense membranes that actually separate salt molecules (ions) from the water, by similarly keeping the salt on the upstream side and allowing the pressurized pure water to flow through the membrane, were developed in the 1950s. The reverse of this process, **osmosis**, is well known: for example, if a membrane is placed to separate water from an aqueous salt solution, and the membrane is semipermeable (here meaning that it permits transfer of water only, not the salt components in the aqueous solution), the water will tend naturally to migrate through this membrane into the salt solution. Osmosis is, for example, the major mass transport phenomenon across living cells. The driving force for this water flux is proportional to the concentration difference between the two sides of the membrane, and is exhibited as the so-called **osmotic pressure**, which is higher by 2.51 MPa on the water side of the membrane for typical seawater at 25°C. If a pressure higher than the osmotic pressure is applied on the saline solution side of the membrane, the water flux can be reversed to move pure water across the membrane from the saline solution side to the pure water one. This process is called *reverse osmosis* (and sometimes *hyperfiltration*), and is the basic principle of RO desalination

Unlike filtration of particulates, the selective "filtration" of the water in RO is not due to the relationship of the membrane pore size to the relative sizes of the salt and water molecules. Rather, one way to explain the process is that the very thin active surface layer of the membrane forms hydrogen bonds with water molecules and thus makes them unavailable for dissolving salt. Salt thus cannnot penetrate through that layer. Water molecules approaching that layer are, however, transported through it by forming

such hydrogen bonds with it and in that process displacing water molecules that were previously hydrogen bonded at these sites. The displaced water molecules then move by capillary action through the pores of the remainder of the membrane, emerging at its other side.

The most prevalent membrane configurations used in RO plants are of the spiral-wound or hollow-fiber types. The basic spiral-wound-type module (Figure 20.6.8) is made of two sheets placed upon each other and rolled together in an increasing diameter spiral around a cylindrical perforated tube. One of the sheets is in the form of a sandwich typically composed of five layers bonded together along three edges. The two outer layers are the semipermeable membranes. Each of them is backed by a porous material layer for mechanical strength, and the very central layer is a thicker porous material layer that takes up the produced fresh water. The second sheet is a porous mesh through which the high-pressure saline water feed is passed in an axial direction. Product water separates from the saline solution and permeates through the two adjacent semipermeable membranes into the central product water–carrying layer, which conducts it spirally to the unbonded edge of the "sandwich" and to the inner perforated tube. The semipermeable membranes are typically made from cellulose acetate, and more recently from composites of several polymers.

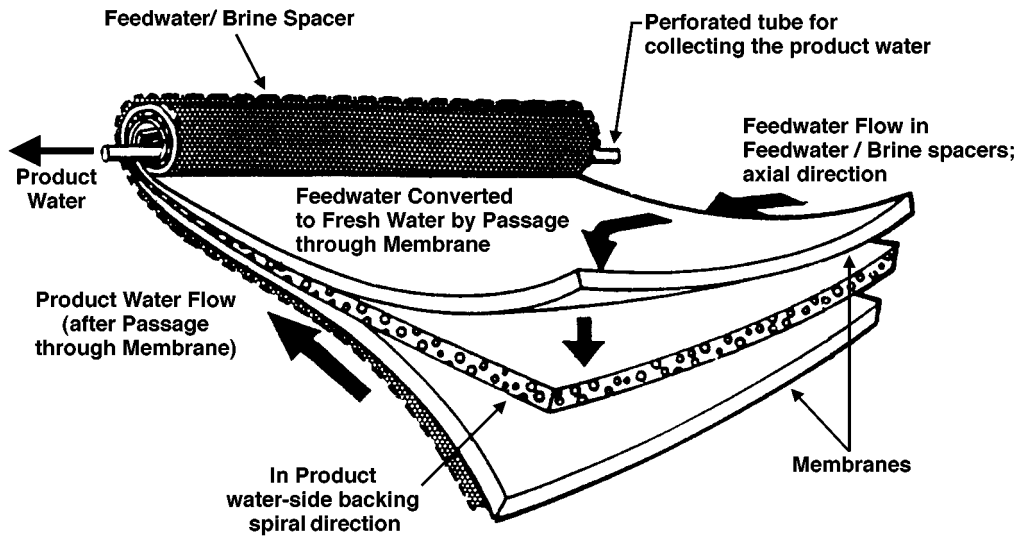


FIGURE 20.6.8 A spiral-wound RO membrane element.

Hollow fiber modules have a configuration similar to a shell-and-tube heat exchanger, with the fibers taking the place of the tubes. A very large number of typicallly 25 to 250 μm outside-diameter semi-permeable hollow fibers (wall thickness typically 5 to 50 μm) are bundled together and placed in a saline water pressure vessel. The hollow core of each fiber is sealed on one end. The pressurized saline water is brought into the module (through a central porous feed tube, Figure 20.6.9) to circulate on the exterior surface of the fibers, and water permeates through the fiber wall into its hollow core, through which it flows to a permeate collection manifold at the open end of the fiber bundle. The increasingly concentrated saline water flows radially and is discharged at the exterior shell of the bundle. The hollow fibers are typically made of polyamide or cellulose triacetate, and offer about 20 fold more surface (separation) area per unit volume than the spiral-wound configuration.

The basic approximate equation for the separation process gives the water flux $\dot{m}''_w$ (kg/m^{2}sec) across an RO membrane, in the absence of fouling, as

$$\dot{m}''_w = K_{pe} K_{cf} \left[\left(P_f - P_p \right) - \left(\pi_f - \pi_p \right) \right] \tag{20.6.10}$$

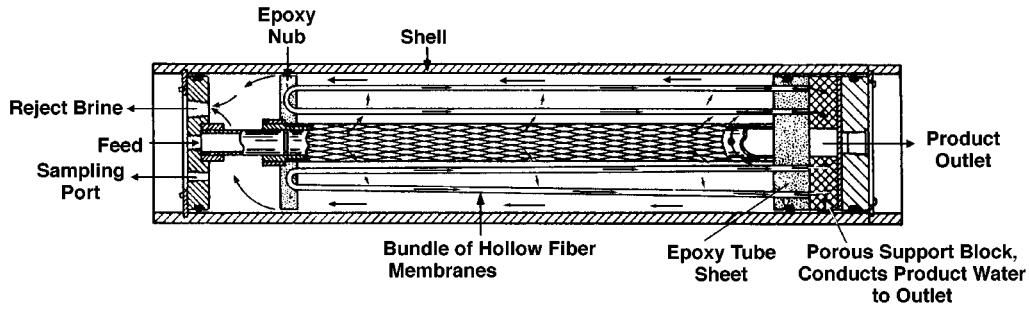


FIGURE 20.6.9 A hollow-fiber RO membrane module. (Du Pont Permasep™.)

where

K_{pe} *water permeability constant* of the membrane (in kg/m^{2}sec Pa), typically increasing strongly as the temperature rises: a plant designed to operate at 20°C may produce up to 24% more water if the water temperature is 28°C,

K_{cf} *compaction correction factor* (dimensionless) which corrects for the fact that the flux is reduced due to densification of the barrier layer (a phenomenon similar to creep) of the membrane, and which increases with the operating pressure and temperature. It is often calculated by the relationship

$$K_{cf} = BC(T)C(P)C(t) \tag{20.6.11}$$

where B is a constant,
$C(T)$ represents the temperature dependence of the Compaction Correction Factor for the particular membrane of interest,
$C(P)$ represents its pressure dependence: while a higher pressure difference across the membrane is shown in Equation (20.6.10) to increase the water flux, higher feed pressure (P_f) also tends to compact the membrane and thus reduce its water flux, typically according to

$$C(P) = P_f^n \tag{20.6.12}$$

where n is a negative number,
and where the time dependence $C(t)$ is represented by

$$C(t) = t^m \tag{20.6.13}$$

where t is the operating time (say, in days) and m is a negative number depending on the membrane.

P water or saline solution pressure (Pa),

π osmotic pressure (Pa),

and the subscripts f and p pertain to the saline feed water and to the desalted product water, respectively.

The required membrane area A can be estimated by

$$A = \frac{\mathfrak{M}_p}{\mathfrak{M}''_p f} \tag{20.6.14}$$

where $\dot{m}_p$ is the freshwater mass production rate of the plant (kg/sec), and f ($0 < f \leq 1.0$) is the *area utilization factor* that corrects for the fact that the membrane surface is incompletely in contact with the saline water feed stream due to the porous mesh and other devices, such as turbulence promoters, placed in the feed stream path; in a good design $f > 0.9$.

Examination of Equation (20.6.10) shows that water separation rate increases with the water permeability constant K_{pe}. Unfortunately, so does the salt flux across the membrane, resulting in a saltier product. An approximation for this salt flow is

$$\dot{m}_s = KK_s\left(C_{fm} - C_p\right) \tag{20.6.15}$$

where

$\dot{m}_s$ salt mass transfer rate across the membrane, kg/sec,
K a proportionality constant, dimensionless,
K_s *salt permeation constant,* kg/sec, which increases with pressure and temperature.

The salinity of the product water (C_p) can be estimated by the formula

$$C_p = K_{cp}\left(1 - \eta\right)\overline{C} \tag{20.6.16}$$

where

K_{cp} *concentration polarization coefficient*, $\equiv C_{fm}/\overline{C}$ is a measure of the increase of the feedwater salinity at the membrane wall beyond that of the bulk solution,
C_{fm} salt concentration at the membrane wall,
$\overline{C}$ bulk salinity of the saline water feed, $\approx (C_f + C_r)/2$,
C_r salt concentration of the reject brine,
η *salt rejection factor*, $\equiv$ (amount of salts rejected by the membrane)/(amount of salts in the brine feed).

The pressure to be used for RO depends on the salinity of the feed water, the type of membrane, and the desired product purity. It ranges form about 1.5 MPa for low feed concentrations or high-flux membranes, through 2.5 to 4 MPa for brackish waters, and to 6 to 8.4 MPa for seawater desalination. In desalination of brackish water, typical product water fluxes through spiral-wound membranes are about 600 to 800 kg/(m^{2}day) at a recovery ratio (RR) of 15% and an average salt rejection of 99.5%, where

$$\text{RR} = \frac{\dot{m}_p}{\dot{m}_f} \cong 1 - \frac{C_f}{C_r} \tag{20.6.17}$$

The fluxes in hollow-fiber membranes used in seawater desalination are 20- to 30-fold smaller, but the overall RO system size does not increase, because the hollow-fiber membranes have a much larger surface area per unit volume. The RR and salt rejection ratio are similar to those of spiral-wound membranes.

Since the concentrated reject brine is still at high pressure, it is possible to recover energy by passing this brine through hydraulic turbines, and thus reduce the overall energy consumption by up to 20%. The energy requirements of seawater RO desalination plants with energy recovery are about 5 to 9 kWh, or 18 to 33 MJ, of mechanical or electric power per m^3 fresh water produced. In comparison, the MSF desalination process requires about 120 to 280 MJ of heat and about 15 MJ of mechanical/electric power (for pumping and auxiliaries) per m^3. The energy requirement of the RO process is thus smaller than that of the MSF process even if the RO energy requirement is multiplied by the thermal-to-mechanical

(or electrical) power conversion factor of 3 to 4. The specific *exergy* consumption of the MSF process using 120°C steam is about 2- to 3-fold higher than that of the RO process, but becomes comparable in magnitude if the steam temperature is lowered to 80°C.

The life of membranes is affected by gradual chemical decomposition or change. For example, cellulose acetate membranes **hydrolyze** with time. The rate of hydrolysis has a steep minimum at a solution pH of 4.5 to 5.0, and increases drastically with temperature.

Membranes are susceptible to plugging by dirt and to deterioration in their selectivity caused by various species present in the saline water. Careful pretreatment of the feed water is therefore necessary. It typically consists of clarification, filtration, chlorination for destroying organic matter and microorganisms, removal of excess chlorine to prevent membrane oxidation, and dosing with additives to prevent calcium sulfate scaling and foam formation. Periodical chemical or mechanical cleaning is also necessary. Pretreatment and cleaning are significant and increasing fractions of the RO plant capital and operating costs.

Further detail about RO desalination can be found in Sourirajan and Matsuura (1985) and Amjad (1993).

Electrodialysis (ED)

In ED, the saline solution is placed between two membranes, one permeable to cations only and the other to anions only. A direct electrical current is passed across this system by means of two electrodes, cathode and anode, exposed to the solution (Figure 20.6.10). It causes the cations in the saline solution to move toward the cathode, and the anions to the anode. As shown in Figure 20.6.10, the anions can leave the compartment in their travel to the anode because the membrane separating them from the anode is permeable to them. Cations would similarly leave the compartment toward the cathode. The exit of these ions from the compartment reduces the salt concentration in it, and increases the salt concentration in the adjacent compartments. Tens to hundreds of such compartments are stacked together in practical ED plants, leading to the creation of alternating compartments of fresh and salt-concentrated water. ED is a continuous-flow process, where saline feed is continuously fed into all compartments and the product water and concentrated brine flow out of alternate compartments. The flow along the membranes also improves the mass transport there, and the separators between the membranes are constructed to provide good flow distribution and mixing on the membrane surfaces. Membrane sizes are roughly 0.5 × 1 m, spaced about 1 mm apart. Many typed of polymers are used to manufacture these ion-exchange selective membranes, which are often reinforced by strong fabrics made from other polymers or glass fibers.

Careful and thorough feed water pretreatment similar to that described in the section on RO is required. Pretreatment needs and operational problems of scaling are diminished in the electrodialysis reversal (EDR) process, in which the electric current flow direction is periodically reversed (say, three to four times per hour), with simultaneous switching of the water flow connections. This also reverses the salt concentration buildup at the membrane and electrode surfaces, and prevents concentrations that cause the precipitation of salts and scale deposition.

The voltage used for ED is about 1 V per membrane pair, and the current flux is of the order of 100 A/m^2 of membrane surface. The total power requirement increases with the feed water salt concentration, amounting to about 10 MW/m^3 product water per 1000 ppm reduction in salinity. About half this power is required for separation and half for pumping. Many plant flow arrangements exist, and their description can be found, along with other details about the process, in Shaffer and Mintz (1980) and Heitman (1991).

Defining Terms

Boiling point elevation: The number of degrees by which the boiling point temperature of a solution is higher than that of the pure solute at the same pressure.

Flash evaporation: An evaporation process that occurs when a liquid with a free surface is exposed to its vapor, where the vapor is below the saturation pressure corresponding to the temperature of the liquid.

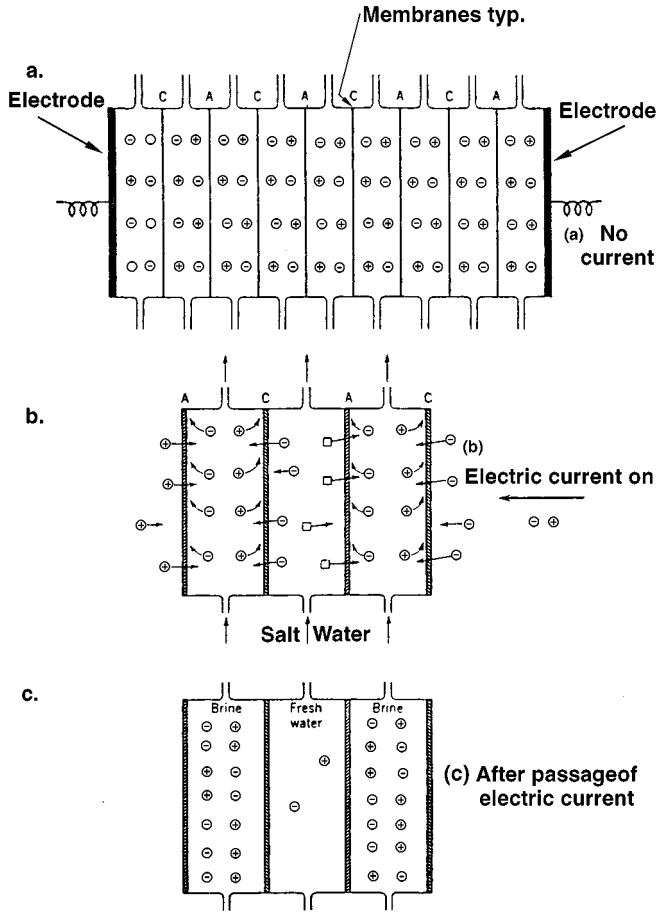


FIGURE 20.6.10 The ED process. *C* and *A* are cation- and anion-permeable membranes, respectively. Application of electric current causes ion transport in a way that salt is depleted in alternate compartments, and enriched in the remaining ones.

The process is typically vigorous, accompanied by rapid growth of bubbles and associated turbulence in the liquid.

Hydrolysis: Decompostion in which a compound is split into other compounds by taking up the elements of water.

Osmosis: The diffusion process of a component of a solution (or mixture) across a semipermeable membrane, driven by the concentration difference (or gradient) of that component across the membrane.

Osmotic pressure: The minimal pressure that has to be applied to the solution (mixture) on the lower concentration side of a membrane permeable to one solution component, for stopping the osmosis of that component through the membrane.

References

Amjad, Z., Ed. 1993. *Reverse Osmosis: Membrane Technology, Water Chemistry and Industrial Applications*. Van Nostrand Reinhold, New York.

Aschner, F.S. 1980. Dual purpose plants, in *Principles of Desalination*, 2nd ed., Part A, K.S. Spiegler and A.D.K. Laird, Eds., Academic Press, New York, chap. 5, 193–256.

Burns and Roe, Inc, 1969. *Universal Design—Report and User's Manual on Design of 2.5 Million Gallon per Day Universal Desalting Plant*, Vols. I–V, U.S. Department of the Interior, O.S.W. Contract No. 14-01-0001-955, Washington, D.C.

Fabuss, B.M. 1980. Properties of seawater, in *Principles of Desalination*, 2nd ed., Part B, K. S. Spiegler and A.D.K. Laird, Eds., Academic Press, New York, Appendix 2, 765–799.

George P.F., Manning, J.A., and Schrieber, C.F. 1975. *Desalinatin Materials Manual*. U.S. Department of the Interior, Office of Saline Water, Washington, D. C.

Glater, J., York, J.L., and Campbell, K.S. 1980. Scale formation and prevention, in *Principles of Desalination*, 2nd ed., Part B, K.S. Spiegler and A.D.K. Laird, Eds., Academic Press, New York, chap. 10, 627–678.

Glueckstern, P. 1995, Potential uses of solar energy for seawater desalination, *Desalination*, 101, 11–20.

Heitman, H.-G. 1990. *Saline Water Processing*, VCH Publications, New York.

Hoffman, D. 1992. The application of solar energy for large scale sea water desalination, *Desalination*, 89, 115–184.

Kasper, S.P. and Lior, N. 1979. A methodology for comparing water desalination to competitive freshwater transportation and treatment, *Desalination*, 30, 541–552.

Khan, A.S. 1986. *Desalination Processes and Multistage Flash Distillation Practice*, Elsevier, Amsterdam.

Lior, N., Ed. 1986. *Measurements and Control in Water Desalinaion*, Elsevier, Amsterdam.

Lior, N. 1991. Thermal theory and modeling of solar collectors, in *Solar Collectors, Energy Storage, and Materials*, F. de Winter, Ed., MIT Press, Cambridge, MA, chap. 4, 99–182.

Lior, N. 1993. Research and new concepts, in *Active Solar Systems*, G.O.G. Löf, Ed., MIT Press, Cambridge, MA, chap. 17, 615–674.

Lior, N. and Greif, R, 1980. Some basic observations on heat transfer and evaporation in the horizontal flash evaporator, *Desalination*, 33, 269–286.

Miyatake, O., Hashimoto, T., and Lior, N. 1992. The liquid flow in multi-stage flash evaporators, *Int. J. Heat Mass Transfer*, 35, 3245–3257.

Miyatake, O., Hashimoto, T., and Lior, N. 1993. The relationship between flow pattern and thermal non-equilibrium in the multi-stage flash evaporation process, *Desalination*, 91, 51–64.

M.W. Kellogg Co. 1975. *Saline Water Data Conversion Engineering Data Book*, 3rd ed., U.S. Department of the Interior, Office of Saline Water Contract No. 14-30-2639, Washington, D.C.

Rabl, A. 1985. *Active Solar Collectors and Their Applications*, Oxford University Press, New York.

Shaffer, L.H. and Mintz, M.S. 1980. Electrodialysis, in *Principles of Desalination*, 2nd ed., Part A, K.S. Spiegler and A.D.K. Laird, Eds., Academic Press, New York, chap. 6, 257–357.

Sourirajan, S. and Matsuura, T., Eds. 1985. *Reverse Osmosis and Ultrafiltration*, ACS Symposium Series 281, American Chemical Society, Washington, D.C.

Spiegler, K.S. and El-Sayed, Y.M. 1994. *A Desalination Primer*. Balaban Desalination Publications, Mario Negri Sud Research Institute, 66030 Santa Maria Imbaro (Ch), Italy.

Spiegler, K.S. and Laird, A.D.K., Eds. 1980. *Principles of Desalination*, 2nd ed., Academic Press, New York.

Standiford, F.C. 1986a. Control in multiple effect desalination plants, in *Measurements and Control in Water Desalination*, N. Lior, Ed., Elsevier, Amsterdam, chap. 2.2, 263–292.

Standiford, F.C. 1986b. Control in vapor compression evaporators, in *Measurements and Control in Water Desalination*, N. Lior, Ed., Elsevier, Amsterdam, chap. 2.3, 293–306.

Steinbruchel, A.B. and Rhinesmith, R.D. 1980. Design of distilling plants, in *Principles of Desalination*, 2nd ed., Part A, K.S. Spiegler and A.D.K. Laird, Eds., Academic Press, New York, chap. 3, 111–165.

Tleimat, B.W. 1980. Freezing methods, in *Principles of Desalination*, 2nd ed., Part B, K.S. Spiegler and A.D.K. Laird, Eds., Academic Press, New York, chap. 7, 359–400.

Further Information

The major texts on water desalination written since the 1980s are Spiegler and Laird (1980), Khan, (1986) (contains many practical design aspects), Lior (1986) (on the measurements and control aspects), Heitman (1990) (on pretreatment and chemistry aspects), and Spiegler and El-Sayed (1994) (an overiew primer). Extensive data sources are provided in George et al. (1975) and M. W. Kellog (1975).

The two major professional journals in this field are *Desalination, The International Journal on the Science and Technology of Desalting and Water Purification* and *Membrane Science*, which often addresses membrane-based desalination processes, both published by Elsevier, Amsterdam.

The major professional society in the field is the International Desalination Association (IDA) headquartered at P.O. Box 387, Topsfield, MA 01983. IDA regularly organizes international conferences, promotes water desalinaton and reuse technology, and is now publishing a trade magazine T*he International Desalination & Water Reuse Quarterly.*

The *Desalination Directory* by M. Balaban Desalination Publications, Mario Negri Sud Research Institute, 66030 Santa Maria Imbaro (Ch), Italy, lists more than 5000 individuals and 2000 companies and institutions in the world of desalinaiton and water reuse.

Two useful (though by now somewhat dated) books on desalination are by Howe, E. D. 1974. *Fundamentals of Water Desalination*, Marcel Dekker, New York, and by Porteous, A. 1975. *Saline Water Distillation Processes*, Longman, London.

Much information on oceans and seawater properties is avaiable in the book by Riley, J. P. and Skinner, Eds. 1975. *Chemical Oceanography*, Academic Press, New York.

20.7 Noise Control

Malcolm J. Crocker

Introduction

Noise is usually defined as unwanted sound. Noise in industry experienced over an extended period can cause hearing loss. Noise in other environments — in buildings, vehicles, and communities from a variety of sources causes speech interference, sleep disturbance, annoyance, and other effects (Crocker, 1997b,d). Noise propagates as sound waves in the atmosphere and as vibration in buildings, machinery, vehicles, and other structures. Noise can be controlled at the *source*, in the *path*, or at the *receiver*. The ear is more sensitive to noise in the mid- to high-frequency range, but fortunately high-frequency is easier to control than low-frequency noise. Several passive methods of noise and vibration control are described. An example of successful noise control is the considerable reduction in passenger jet aircraft noise in the last several years which has made them considerably quieter.

Sound Propagation

Sound waves propagate rather like ripples on a lake when a stone is thrown in (Crocker, 1997c). The ripples spread out from the source of the disturbance as circular waves until they reach a solid body or boundary (such as the lake edge) where reflections occur. The water does not flow from the source, but the disturbance propagates in the form of momentum and energy which is eventually dissipated. Sound waves in air cannot be seen but behave in a similar manner. Sound waves propagating in three dimensions from a source of sound are spherical rather than circular like the two-dimensional water wave propagation. Sound waves propagate at the wave speed (or *sound speed c*) which is dependent only on the absolute temperature *T*. It is 343 m/sec (1120 ft/sec) at a normal atmospheric temperature of 20°C. The *wavelength* λ of sound is inversely proportional to the *frequency f* in cycles/sec (known as hertz or Hz) and is given by $\lambda = c/f$ Hz. The sound waves result in fluctuations in the air pressure as they propagate. The air pressure difference from the mean atmospheric pressure is defined as the *sound pressure*. A logarithmic measure, the sound pressure level SPL or L_p, is usually used with sound and noise and the units are *decibels* (dB). The sound pressure level is $L_p = 10 \log_{10} (p^2 / p_{\text{ref}}^2)$, where p is the rms sound pressure and p_{ref} is the reference sound pressure 20 μPa (or 20×10^{-6} N/m^2). See Figure 20.7.1 (Crocker, 1997c).

Human Hearing

The human ear has a wide frequency response from about 15 or 20 Hz to about 16,000 Hz (Crocker, 1975; Greenberg, 1997). The ear also has a large dynamic range; the ratio of the loudest sound pressure we can tolerate to the quietest sound that we can hear is about ten million (10^7). This is equivalent to 140 dB. The ear can be divided into three main parts: the outer, middle, and inner ear. The outer ear, consisting of the fleshy pinna and ear canal, conducts the sound waves onto the ear drum. The middle ear converts the sound pressure at the ear drum into the mechanical motion of three small bones (named auditory ossicles: malleus, incus, and stapes) which in turn convert the mechanical motion into waves in the inner ear. Hair cells in the inner ear respond to the excitation and send neural impulses along the auditory nerves to the brain (Figure 20.7.2).

The higher the sound pressure level of a sound, the louder it normally sounds, although the frequency content of the sound is important too. The ear is most sensitive to sound in the mid-frequency range and hears sound only poorly at lower frequencies (below 200 or 300 Hz). Most people have a maximum sensitivity to sound at about 4000 Hz (corresponding to a quarter wave resonance in the ear canal, with a pressure maximum at the eardrum). Electrical filters have been produced corresponding approximately to the frequency response of the ear. The A-weighting filter is the one most used and it filters off a considerable amount of the sound energy at low frequencies. The sound pressure level measured with an A-weighting filter is normally known as the A-weighted sound level (or the sound level for short).

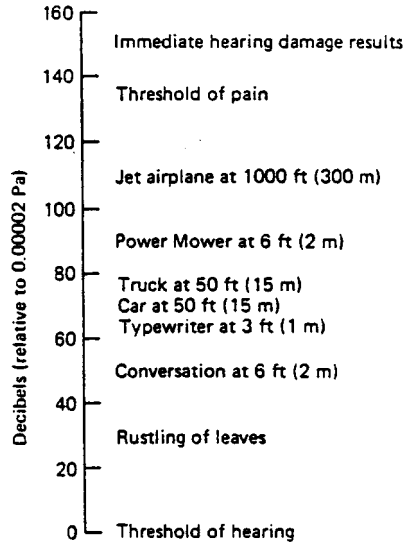


FIGURE 20.7.1 Some typical sound pressure levels.

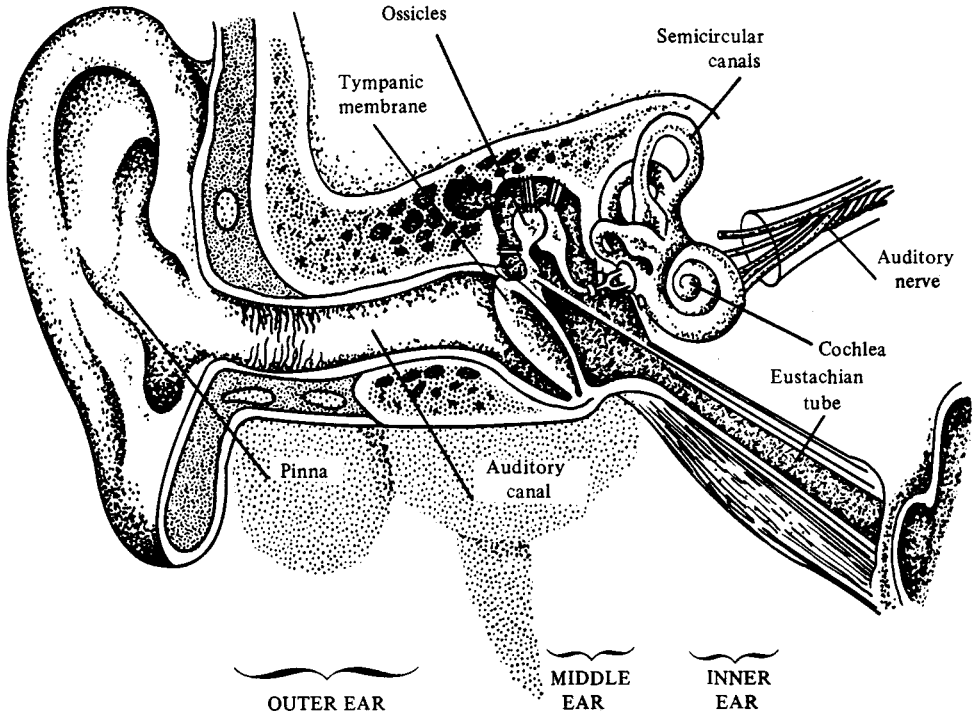


FIGURE 20.7.2 Cross section of the human ear showing the three main parts: outer, middle, and inner ear.

The anatomy and functioning of the ear are described more completely in several books (Crocker, 1997; Greenberg, 1997).

Noise Measures

There are several rating measures and descriptors used to determine human response to noise. Only a few of the most important can be discussed here. The reader will find more such measures discussed in the literature. [1] Criteria derived from such measures can be used to produce regulations or legislation.

The speech interference level (SIL) is a measure used to evaluate the effect of background noise on speech communication. The SIL is the arithmetic average of the sound pressure levels of the interfering background noise in the four octave bands with center frequencies of 500, 1000, 2000 and 4000 Hz.[1,6]

The speech interference level of the background noise is calculated; then this may be used in conjunction with Figure 20.7.3 to determine if communication is possible at various distances for different voice levels. This figure is for male voices. Since the average female voice is normally quieter, for female voices the horizontal scale should be moved to the right by 5 dB. Use of the results obtained with Figure 20.7.3 and criteria for various spaces in buildings enable decisions to be made whether they are suitable for their desired uses.

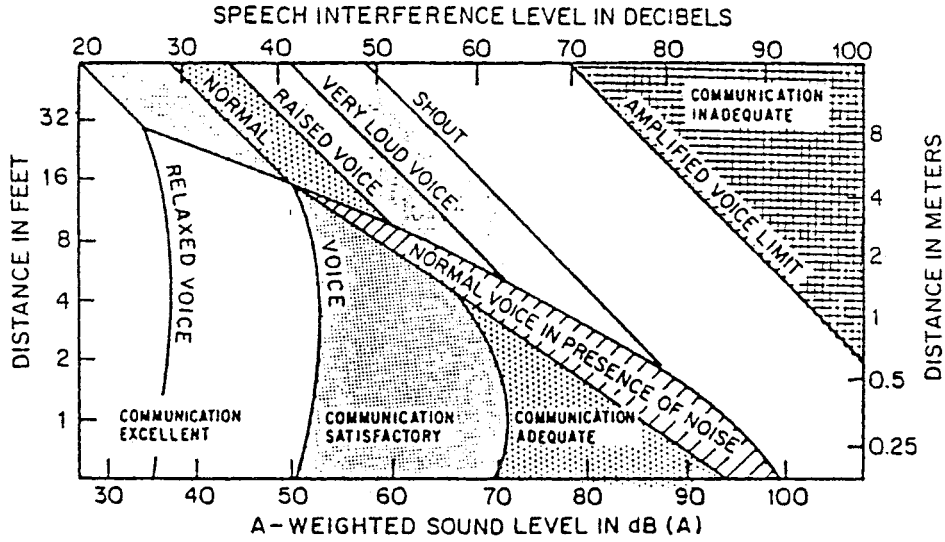


FIGURE 20.7.3 Recommended distances between speaker and listener for various voice levels for just reliable speech communication. (From C.M. Harris *Handbook of Noise Control,* McGraw-Hill, New York, 1979. With permission.)

The equivalent sound level L_{eq} is the A-weighted sound pressure level averaged over a suitable time period T. The averaging time T can be chosen to be a number of minutes, hours or days, as desired.

$$L_{eq} = 10\log_{10}\left[(1/T)\int p_A^2\, dt / p_{ref}^2\right] dB$$

where p_A is the instantaneous sound pressure measured using an A-weighting frequency filter. The L_{eq} is also sometimes known as the *average sound level* L_{AT}.

The *day-night equivalent sound level* (DNL) or L_{dn} is a measure that accounts for the different human response to sound at night. It is defined (Crocker, 1997d) as:

$$L_{\text{dn}} = 10\log_{10}\left\{(1/24)\left[15\left(10^{L_d/10}\right) + 9\left(10^{(L_n+10)/10}\right)\right]\right\}\ dB$$

where L_d is the 15-hr daytime A-weighted equivalent sound level (from 0700 to 2200 hr) and L_n is the 9-hr nighttime equivalent sound level (from 2200 to 0700 hr). The nighttime level is subjected to a 10-dB penalty because noise at night is known to be more disturbing than noise during the day.

There is some evidence that noise that fluctuates markedly in level is more annoying than noise which is steady in level. Several noise measures have been proposed to try to account for the annoying effect of these fluctuations. The percentile levels are used in some measures. The *percentile level* L_n is defined to be the level exceeded n% of the time (Crocker, 1997d). The A-weighted sound level is normally used in Ln.

Response of People to Noise and Noise Criteria and Regulations

In industry, noise is often intense enough to interfere with speech and to create noise conditions that are hazardous to hearing. By using Figure 20.7.3 it is seen that if the SIL is above 50 dB, then the noise will interfere with normal speech communication between male voices at 4 m. If it is above 60 dB, then speech communication even when shouting is barely possible at the same distance. For women the comparable values of SIL are 45 and 55 dB at the same distance. If the SIL is 90 dB, then communication between male voices is only possible at distances less than 0.25 m, even when shouting. A-weighted sound levels are sometimes used instead of SIL values but with somewhat less confidence. It is seen that if one has difficulty in communicating in an industrial situation, then the A-weighted sound level is likely to be above 90 dB. In the United States, OSHA regulations start at this level for an 8-hr period. There is a halving in the allowable exposure time for every 5-dB increase in sound level. See Table 20.7.1. In almost all other countries the allowable exposure time is halved for every 3-dB increase in sound level (Ward, 1997).

TABLE 20.7.1 Maximum A-Weighted Sound Levels Allowed by the U.S. Occupational Safety and Health Administration (OSHA) for Work Periods Shown during a Workday

Duration per Day (hr)	Sound Level in dB(A)
8	90
6	92
4	95
3	97
2	100
1.5	102
1	105
0.5	110
0.25 or less	115

Noise in communities is caused by many different sources. In most countries, the maximum equivalent A-weighted sound level L_{eq} is recommended for evaluating different types of noise source (Gottlob, 1995). In some countries there are regulations which use L_{eq} for road traffic noise and railroad noise, although some countries use L_{10} (e.g., the U.K.) or L_{50} (e.g., Japan) for planning permission in which road traffic noise is of concern (Crocker, 1997d; Gottlob, 1995). In the United States the L_{dn} has been used for community noise situations involving aircraft noise at airports and road traffic noise. Table 20.7.2 presents levels of noise given by the U.S. EPA and several other bodies to protect public health.

TABLE 20.7.2 Guidelines from the U.S. Environmental Protection Agency (EPA), World Health Organization (WHO), Federal Interagency on Community Noise (FICON), and Various European Agencies for Acceptable Noise Levels

Authority	Specified Sound Levels	Criterion
EPA Levels Document	$L_{dn} \leq 55$ dB (outdoors) $L_{dn} \leq 45$ dB (indoors)	Protection of public health and welfare with adequate margin of safety
WHO Document (1995)	$L_{eq} \leq 50/55$ dB (outdoors: day) $L_{eq} \leq 45$ dB (outdoors: night) $L_{eq} \leq 30$ dB (bedroom) $L_{max} \leq 45$ dB (bedroom)	Recommended guideline values (Task Force consensus)
U.S. Interagency Committee (FICON)	$L_{dn} \leq 65$ dB	Considered generally compatible with residential development
	$65 \leq L_{dn} \leq 70$ dB	Residential use discouraged
Various European road traffic regulations	$L_{eq} \geq 65$ or 70 dB (day)	Remedial measures required

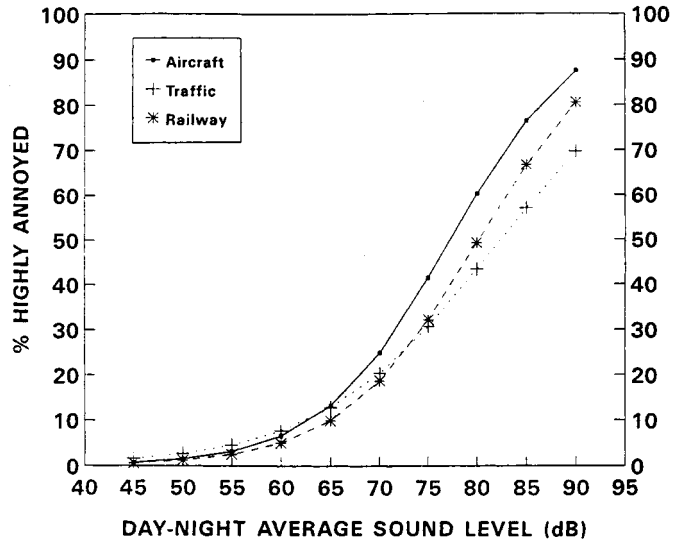


FIGURE 20.7.4 Percentage of survey respondents highly annoyed vs. day-night equivalent sound level for aircraft, road traffic, and railway noise.

Social surveys in several countries have been used to relate the percentage of respondents highly annoyed by noise to the day-night equivalent sound level, L_{dn} (Crocker, 1997d; Gottlob, 1995). See Figure 20.7.4. (Finegold et al., 1994). It is seen that many studies have shown that aircraft noise appears to be more annoying than other sources, perhaps because of the larger fluctuation in levels with time compared with the other sources. However, some other studies suggest that railroad noise is less annoying than traffic noise and this is borne out by the lower levels used for railroad noise than traffic noise regulations in several European countries (Gottlob, 1995).

Various investigations have shown that noise disturbs sleep (Crocker, 1997d). It is well known that there are several stages of sleep and that people progress through these stages as they sleep. Noise can change the progress through these stages and if sufficiently intense can awaken the sleeper. Recently, sleep disturbance data from several analyses have been reanalyzed and the preliminary sleep disturbance curve given in Figure 20.7.5 has been proposed. A regression fit to these sleep disturbance data (Finegold et al., 1994) gave the following expression (which is also shown graphically in Figure 20.7.5) (Crocker, 1997d; Finegold et al., 1994):

$$\% \text{ Awakenings} = 7.1 \times 10^{-6} L_{AE}^{3.5}$$

where L_{AE} is the indoor A-weighted sound exposure level ASEL.

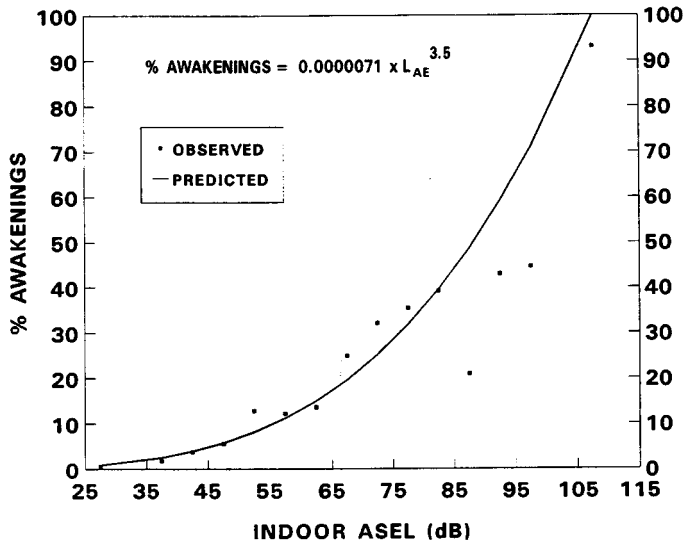


FIGURE 20.7.5 Proposed sleep disturbance curve: percentage of subjects awakened as a function of indoor sound exposure level.

Noise Control Approaches

The main noise control approaches include use of sound absorption, enclosures, barriers, and vibration isolation and damping (Crocker, 1997c). Most porous materials absorb sound and those materials specially made for this purpose include materials such as porous foams and fiberglass. However, ordinary materials such as carpets and drapes are also effective and can be used in building spaces to reduce reverberant sound buildup and noise. Although all these materials are rather ineffective at low frequency, at frequencies above 500 to 1000 Hz they can absorb almost all of the sound energy incident on them and in this case are said to have an absorption coefficient α of one. In industry they are used inside machine enclosures or placed on the walls and inside ceilings of manufacturing buildings to reduce the reverberant noise buildup (Crocker, 1997c).

Enclosures can be used to partially or completely enclose machines (machine enclosures) or to enclose operators of machines (personnel enclosures). The first approach may be regarded as *path* control and the second as *receiver* control. The improvement in noise reduction that these enclosures can achieve is related not only to the so-called transmission loss TL of the enclosure material used, TL = 10 log mf – 34 dB, where m is the mass/unit area of the enclosure walls in kg/m^2 and f is the frequency in Hz, but also to the absorption coefficient α by 10 log($1/\alpha$). Thus, enclosures that have massive walls and absorbing material inside are the most effective at reducing noise both as machine enclosures or personnel enclosures. The enclosures should be well sealed to prevent sound being transmitted through leaks. If it is necessary to have a vent or hole in the enclosure for ventilation or for access, then the vent should be lined with sound-absorbing material and be bent or constructed like a labyrinth to try to reduce the direct transmission of noise through the vent or hole. As the relationship for TL indicates, enclosures are generally more effective at high frequency (Crocker, 1997c).

Barriers are used to shield personnel from sound sources. The effectiveness of a barrier depends not only on the effective height of the barrier in wavelengths, but also how far the receiver point is into the sound shadow. Barriers are thus most effective when they are taller (in wavelengths) and thus best for

high-frequency noise and also best when placed close to the source or close to the receiver, since such placement increases the shadowing effect. Barriers are used in industry where it is desired to shield personnel from machinery noise sources. It is important in such cases to put sound-absorbing material on the ceiling of the building just above the barrier or on walls just behind a barrier where these surfaces could allow the reflection of sound to bypass the barrier and thus severely reduce its effectiveness (Crocker, 1997c).

The vibration isolation of machine sources from their supports can be particularly useful in reducing the noise produced especially if the machine is small compared with a large flexible support or enclosure that can act as a sounding board and radiate the sound. Soft metal springs or elastomeric isolators are often used as isolators. They should be designed so that the natural frequency of the machine mass on its isolators is much less than the forcing frequency, if possible. Care should be taken that such a design condition does not produce excessive static deflection of the system that could interfere with the proper machine operation. Vibrating pipes and ducts can also be vibration-isolated from walls of buildings using pipe hangers or soft rubber isolators. Vibration breaks made of rubber or similar soft material can be built into elements such as walls in buildings or structural elements in vehicles or machine enclosures to prevent vibration being propagated throughout the building or other structure and being reradiated as noise (Crocker, 1997c).

Damping materials can also be effective at reducing noise when applied properly to structures if their vibration is resonant in nature. Damping materials that are viscous, applied with thicknesses two or three times that of the vibrating metal panel, are particularly effective. Constrained damping layers can be very effective even when the damping layer is relatively thin (Crocker, 1997c).

Figure 20.7.6 shows the reduction in the A-weighted sound level that can be expected using these passive noise control approaches discussed above. Often it is insufficient to use one approach, and greater, more-economic noise reduction can be achieved by using two or more approaches in conjunction.

References

Beranek, L.L. and Ver, I.L. 1992. *Noise and Vibration Control Engineering*, John Wiley & Sons, New York.

Crocker, M.J. 1975. in *Noise and Noise Control*, CRC Press, Cleveland, OH, chap. 2.

Crocker, M.J. 1997a. Introduction to linear acoustics, in *Encyclopedia of Acoustics*, M.J. Crocker, Ed., John Wiley & Sons, New York, chap. 1.

Crocker, M.J. 1997b. Noise, in *Handbook of Human Factors and Ergonomics*, 2nd ed., G. Salvendy, Ed., John Wiley & Sons, New York, chap. 24.

Crocker, M.J. 1977c. Noise generation in machinery, its control and source identification, in *Encyclopedia of Acoustics*, M.J. Crocker, Ed., John Wiley & Sons, New York, chap. 83.

Crocker, M.J. 1997d. Rating measures, criteria, and procedures for determining human response to noise, in *Encyclopedia of Acoustics*, M.J. Crocker, Ed., John Wiley & Sons, New York, chap. 80.

Finegold, L.S., Harris, C.S., and von Gierke, H.E. 1994. Community annoyance and sleep disturbance: updated criteria for assessment of the impacts of general transportation noise on people, *Noise Control Eng. J.*, 42(1), 25–30.

Gottlob, D. 1995. Regulations for community noise, *Noise/News Int.*, 3(4), 223–236.

Greenberg, S. 1997. Auditory function, chap. 104; Shaw, E.A.G. Acoustical characteristics of the outer ear, chap. 105; Peake, W.T. Acoustical properties of the middle ear, chap. 106; and Slepecky, N.B. Anatomy of the cochlea and auditory nerve, in *Encyclopedia of Acoustics*, M.J. Crocker, Ed., John Wiley & Sons, New York.

Ward, W.D. 1997. Effects of high-intensity sound, in *Encyclopedia of Acoustics*, M.J. Crocker, Ed., John Wiley & Sons, New York, chap. 119.

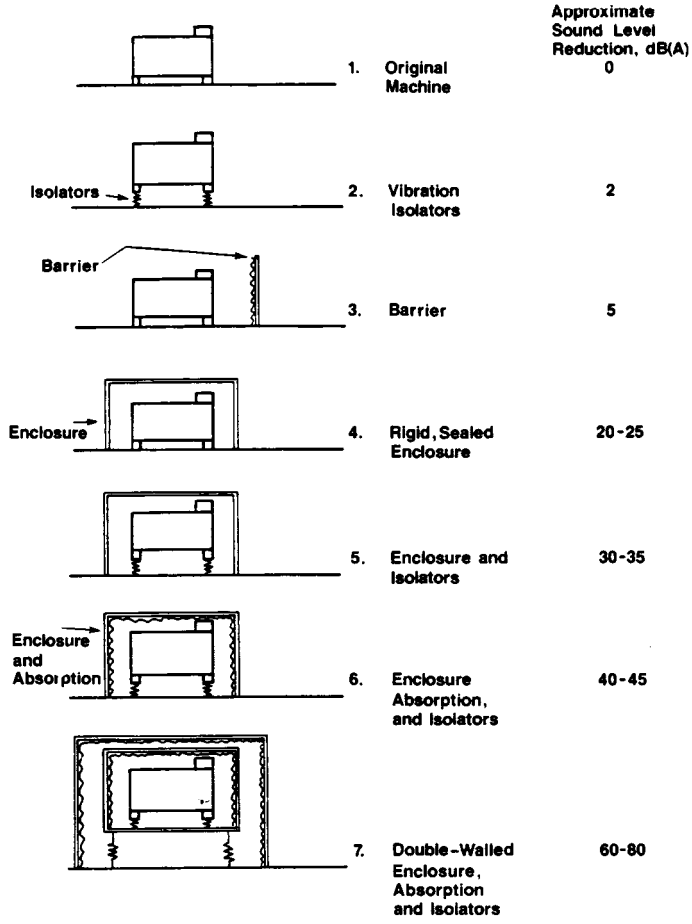


FIGURE 20.7.6 Approximate A-weighted sound level reductions expected from different noise control approaches.

20.8 Lighting Technology*

Barbara Atkinson, Andrea Denver, James E. McMahon, Leslie Shown, Robert Clear, and Craig B. Smith

In this section, we describe the general categories of lamps, ballasts, and fixtures in use today. In addition, we briefly discuss several techniques for improving the energy-efficiency of lighting systems.

Because the purpose of a lamp is to produce light, and not just radiated power, there is no direct measure of lamp efficiency. Instead, a lamp is rated in terms of its **efficacy**, which is the ratio of the amount of light emitted (lumens) to the power (watts) drawn by the lamp. The unit used to express lamp efficacy is lumens per watt (LPW). The theoretical limit of efficacy is 683 LPW and would be produced by an ideal light source emitting monochromatic radiation with a wavelength of 555 nm. The lamps that are currently on the market produce from a few percent to almost 25% of the maximum possible efficacy.** The efficacies of various light sources are depicted in Figure 20.8.1. Lamps also differ in terms of their cost; size; color; lifetime; optical controllability; dimmability; **lumen maintenance***; reliability; simplicity and convenience in use, maintenance, and disposal; and environmental effects (e.g., emission of noise, radio interference, and ultraviolet (UV) light).

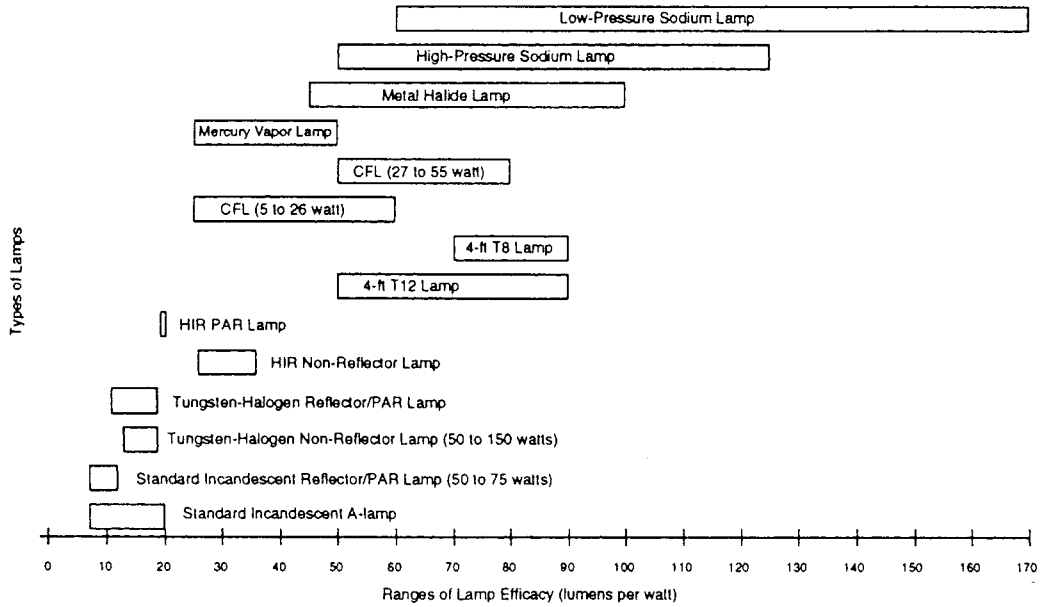


FIGURE 20.8.1 Ranges of lamp efficacy. (Ballast losses are included for all discharge lamps.) (Compiled from U.S. DOE, 1993a; Morse, personal communication, 1994; and manufacturer's catalogs.)

* The contents of this section have been abstracted from Chapers 10 (Electrical Energy Management in Buildings by Craig B. Smith) and 12B (Energy Efficient Lighting Technologies and Their Applications in the Commercial and Residential Sectors by Barbara Atkinson, Andrea Denver, James E. McMahon, Leslie Shown, and Robert Clear) published in the CRC *Handbook of Energy Efficiency*, Frank Kreith and Ronald E. West, Eds., 1997.

** The efficacies of fluorescent lamp/ballast combinations reported in this section are based on data compiled by Oliver Morse from tests by the National Electrical Manufacturer's Association; all comparisons assume SP41 lamps. Efficacies of incandescent lamps are based on manufacturer catalogs. Efficacies of high-intensity discharge lamps are based on manufacturer catalogs and an assumed ballast factor of 0.95.

*** Over time, most lamps continue to draw the same amount of power but produce fewer lumens. The lumen maintenance of a lamp refers to the extent to which the lamp sustains its lumen output, and therefore efficacy, over time.

The color properties of a lamp are described by its color temperature and its color rendering index. **Color temperature**, expressed in degrees Kelvin (K), is a measure of the color appearance of the light of a lamp. The concept of color temperature is based on the fact that the emitted radiation spectrum of a blackbody radiator depends on temperature alone. The color temperature of a lamp is the temperature at which an ideal blackbody radiator would emit light that is closest in color to the light of the lamp. Lamps with low color temperatures (3000 K and below) emit "warm" white light that appears yellowish or reddish in color. Incandescent and warm-white fluorescent lamps have a low color temperature. Lamps with high color temperatures (3500 K and above) emit "cool" white light that appears bluish in color. Cool-white fluorescent lamps have a high color temperature.

The **color rendering index (CRI)** of a lamp is a measure of how surface colors appear when illuminated by the lamp compared to how they appear when illuminated by a reference source of the same color temperature. For color temperatures above 5000 K, the reference source is a standard daylight condition of the same color temperature; below 5000 K, the reference source is a blackbody radiator. The CRI of a lamp indicates the difference in the perceived color of objects viewed under the lamp and under the reference source. There are 14 differently colored test samples, 8 of which are used in the calculation of the general CRI index.

The CRI is measured on a scale that has a maximum value of 100 and is an average of the results for the 8 colors observed. A CRI of 100 indicates that there is no difference in perceived color for any of the test objects; a lower value indicates that there are differences. CRIs of 70 and above are generally considered to be good, while CRIs of 20 and below are considered to be quite poor. Most incandescent lamps have CRIs equal to or approaching 100. Low-pressure sodium lamps have the lowest CRI of any common lighting source (–44); their light is essentially monochromatic.

The optical controllability of a lamp describes the extent to which a user can direct the light of the lamp to the area where it is desired. Optical controllability depends on the size of the light-emitting area, which determines the beam spread of the light emitted. In addition, controllability depends on the fixture in which the lamp is used. Incandescent lamps emit light from a small filament area: they are almost point sources of light, and their optical controllability is excellent. In contrast, fluorescent lamps emit light from their entire phosphored area: their light is extremely diffuse, and their controllability is poor.

Because of the many different characteristics and the variety of applications by which a lamp can be judged, no one type of source dominates the lighting market. The types of lamps that are commonly available include incandescent, fluorescent, and high-intensity discharge (HID).

Lamps

The *incandescent lamp* was invented independently by Thomas Edison in the United States and Joseph Swan in England in the late 1800s. An incandescent lamp produces light when electricity heats the lamp filament to the point of incandescence. In modern lamps the filament is made of tungsten. Because 90% or more of an incandescent's emissions are in the infrared (thermal) rather than the visible range of the electromagnetic spectrum, incandescent lamps are less efficacious than other types of lamps.

The two primary types of standard incandescent lamps are general service and reflector/PAR (parabolic aluminized reflector) lamps. General-service lamps (also known as A-lamps) are the pear-shaped, common household lamps. Reflector lamps, such as flood or spotlights, are generally used to illuminate outdoor areas or highlight indoor retail displays and artwork. They are also commonly used to improve the optical efficiency of downlights (discussed later). Downlights are used where controlling glare or hiding the light source is important.

In spite of the fact that they are the least efficacious lamps on the market today, standard incandescent lamps are used for almost all residential lighting in the U.S. and are also common in the commercial sector. They have excellent CRIs and a warm color; they are easily dimmed, inexpensive, small, lightweight, and can be used with inexpensive fixtures; and, in a properly designed fixture, they permit

excellent optical control. In addition, incandescent lamps make no annoying noises and contain essentially no toxic chemicals. They are simple to install, maintain, and dispose of.

Although they account for less than 3% of the incandescent market, *tungsten-halogen* and now *tungsten-halogen infrared-reflecting (HIR) lamps* are the incandescents that offer the greatest opportunity for energy savings. Halogen lamps produce bright white light and have color temperatures and CRIs that are similar to, or slightly higher than, those of standard incandescents. In addition, they have longer lives, can be much more compact, are slightly more efficacious, and have better lumen maintenance than standard incandescent lamps.

HIR lamps have been promoted to residential- and commercial-sector customers primarily as low-wattage reflector lamps. In general, HIR lamps have a small market share due to their high cost.

Fluorescent lamps came into general use in the 1950s. In a fluorescent lamp, gaseous mercury atoms within a phosphor-coated lamp tube are excited by an electric discharge. As the mercury atoms return to their ground state, ultraviolet radiation is emitted. This UV radiation excites the phosphor coating on the lamp tube and causes it to flouresce, thus producing visible light. Fluorescent lamps are far more efficacious than incandescent lamps. The efficacy of a fluorescent lamp system depends upon the lamp length and diameter, the type of phosphor used to coat the lamp, the type of **ballast** used to drive the lamp, the number of lamps per ballast, the temperature of the lamp (which depends on the fixture and its environment), and a number of lesser factors.

Fluorescent lamps have long lives and fairly good lumen maintanance. While the standard-phosphor (cool-white (CW) and warm-white (WW)), lamps have CRIs of 50 to 60, the new rare earth phosphor lamps have CRIs of 70 to 80. The majority of lighting used in the commercial sector is fluorescent. Fluorescent lighting is also common in the industrial sector. The small amount of full-size fluorescent lighting in the residential sector is primarily found in kitchens, bathrooms, and garages. The most common fluorescent lamps are tubular and 4 ft (1.2 m) in length (see Figure 20.8.2).

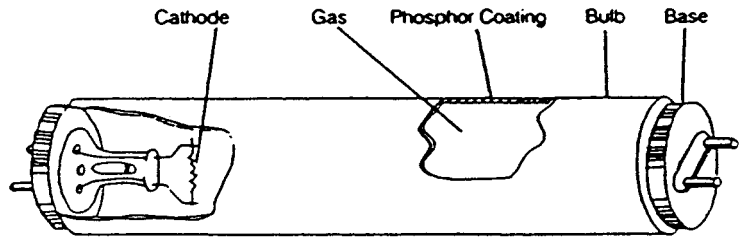


FIGURE 20.8.2 Typical full-size fluorescent lamp. (From Atkinson, B. et al., *Analysis of Federal Policy Options for Improving U.S. Lighting Energy Efficiency: Commercial and Residential Buildings*, Lawrence Berkeley National Laboratory, Berkeley, CA, 1992. With permission.)

Lamp tubes with a diameter of 1.5 in. (38 mm) are called T12s, and tubes that are 1 in. (26 mm) in diameter are called T8s. The 8 and 12 refer to the number of eights of an inch in the diameter of the lamp tube. Lamp tubes are available in other diameters as well. Four-foot T12s are available in 32, 34, and 40 W. The specified wattage of a lamp refers to the power draw of the lamp alone even if lamp operation requires a ballast, as do all fluorescent and high-intensity discharge light sources. The ballast typically adds another 10 to 20% to the power draw, thus reducing system efficacy.

Even the smallest, least efficacious fluorescent lamps (≈30 LPW for a 4-W lamp) are more efficacious than the most efficacious incandescent lamps (≈20 LPW). Of the full-size fluorescent lamps available today, *rare earth phosphor lamps* are the most efficacious. In these lamps, rare earth phosphor compounds are used to coat the inside of the fluorescent lamp tube. Rare earth phosphor lamps are also called tri-phosphor lamps because they are made with a mixture of three rare earth phosphors that produce visible light of the wavelengths to which the red, green, and blue retinal sensors of the human eye are most sensitive. These lamps have improved color rendition as well as efficacy. Fluorescent lamps with

diameters of 1 in. (26 mm) and smaller use tri-phosphors almost exclusively. Rare earth coatings can also be used for lamps of larger diameter.

The most efficacious of the fluorescent lamps available today are T8 lamps operating with electronic ballasts. The efficacy of two 32-W T8 lamps operating with a single electronic ballast is about 90 LPW, approximately 30% more efficacious than the more standard lighting system consisting of two 40-W T12 lamps and a high-efficiency magnetic ballast. T12 lamps are also available with rare earth phosphors, and can attain over 80 LPW when used with a two-lamp electronic ballast. The characteristics and applications of 4-ft T8 lamps are summarized in Table 20.8.1.

TABLE 20.8.1 Characteristics and Applications of 4-ft Full-Size Fluorescent T8 Lamps

Available wattages	32 W
Efficacy	For two T8s and a single electronic ballast, ≈90 LPW
Rated lifetime	Typically 12,000 to 20,000 hr
Color rendition	Typically 75–84
Color temperature	2800–7500 K
Lumen maintenance	Very good (light output typically declines by 10 to 12% over rated lamp life)
Optical controllability	Poor (very diffuse light but slightly better than a T12)
Typical uses	The majority of lighting used in the commercial sector is fluorescent. Fluorescent lighting is also used in the industrial sector. In the residential sector, full-size fluorescent lighting is used in some kitchens, bathrooms, and garages
Technologies for which these lamps are energy-efficient alternatives	These lamps are most often replacements for less-efficient fluorescents (T12s)
Notes	Some T8 lamps are now available with CRIs above 90, especially those with high color temperature; however, these lamps are less efficacious than lamps with lower CRIs

In spite of their much greater efficiency, fluorescent lamps have several disadvantages when compared to incandescent lamps. Fluorescent lamps can be dimmed, but only with special equipment that costs much more than the dimming controls used for incandescent lamps. Standard fluorescent lamps are much bigger than incandescent lamps of equivalent output and are much harder to control optically. In addition, fluorescent lamps contain trace amounts of mercury, a toxic metal, and emit more UV light than incandescent lamps. The ballast equipment that drives the lamps is sometimes noisy and may emit radio interference.

Circular fluorescent lamps in 20- to 40-W sizes have been available for many years, but have always had a fairly small market. Essentially, a circular lamp is a standard fluorescent lamp tube (as described earlier) that has been bent into a circle. Although they have a more compact geometry than a straight tube, circular lamps are still moderately large (16.5 to 41 cm in diameter).

Compact fluorescent lamps (CFLs), which are substantially smaller than standard fluorescent lamps, were introduced to the U.S. market in the early 1980s. In a CFL, the lamp tube is smaller in diameter and is bent into two to six sections (see Figure 20.8.3). CFLs have much higher power densities per phosphor area than standard fluorescents, and their design was therefore dependent on the development of rare earth phosphors, which could hold up much better than standard phosphors at high power loadings. CFLs are available as both screw-in replacements for incandescent lamps and as pin-base lamps for hard-wired fixtures. They may be operated with separate ballasts or purchased as integral lamp/ballast units.

High-intensity discharge lamps produce light by discharging an electrical arc through a mixture of gases. In contrast to fluorescent lamps, HID lamps use a compact "arc tube" in which both temperature and pressure are very high. Compared to a fluorescent lamp, the arc tube in an HID lamp is small enough to permit compact reflector designs with good light control. Consequently, HID lamps are both compact and powerful. There are currently three common types of HID lamps available: mercury vapor (MV), metal halide (MH), and high-pressure sodium (HPS).

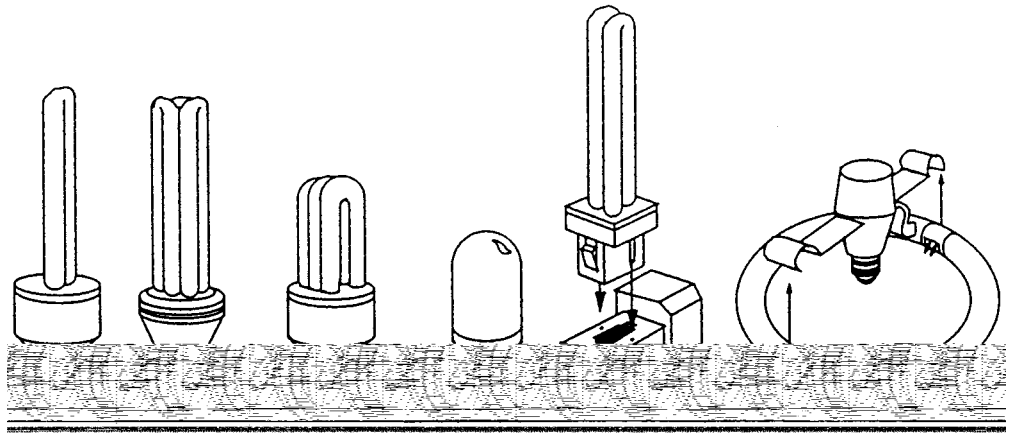


FIGURE 20.8.3 A variety of compact fluorescent and circular lamps: (a) twin-tube integral, (b and c) triple-tube integrals, (d) integral model with glare-reducing casing, (e) modular quad tube and ballast, and (f) modular circline and ballast. (Courtesy of Katherine Falk and *Home Energy* magazine, 1995.)

Because of their very high light levels, except in the smallest lamps, and their substantial cost ($15 to 80 for lamps up to 400 W), HID lamps are most often used for exterior applications such as street lighting and commercial, industrial, and residential floodlighting (sports and security lighting). They are also used in large, high-ceilinged, interior spaces such as industrial facilitites and warehouses, where good color is not typically a priority. Occasionally, HID lamps are used for indirect lighting in commercial offices. Interior residential applications are rare because of high cost, high light level, and the fact that HID lamps take several minutes to warm up to full light output. If there is a momentary power outage, the lamps must cool down before they will restrike. Some HID lamps are now available with dual arc tubes or parallel filaments. Dual arc tubes eliminate the restrike problem and a parallel filament gives instantaneous light output both initially and on restrike, but at a cost of a high initial power draw.

The *mercury vapor lamp* was the first HID lamp developed. Including ballast losses, the efficacies of MV lamps range from about 25 to 50 LPW. Uncoated lamps have a bluish tint and very poor color rendering (CRI ≈ 15). Phosphor-coated lamps emit more red, but are still bluish, and have a CRI of about 50. Because of their poor color rendition, these lamps are only used where good color is not a priority. Although MV lamps generally have rated lifetimes in excess of 24,000 hr, light output can diminish significantly after only 12,000 hr (Audin et al., 1994). Both metal halide and high-pressure sodium HIDD lamps have higher efficacies than mercury vapor lamps and have consequently replaced them in most markets.

Including ballast losses, *metal halide lamps* range in efficacy from 46 to 100 LPW. They produce a white to slightly bluish-white light and have CRIs ranging from 65 to 70. Lamp lifetimes range from only 3500 to 20,000 hr, depending on the type of MH lamp. Lower-wattage metal halides (particularly the 70-W and 100-W) are now available with CRIs of 65 to 75 and color temperatures of 3200 to 4200 K. Good lumen maintenance, longer life, reduced maintenance costs, and the fact that they blend more naturally with fluorescent sources have made MH lamps a very good replacement in the commercial sector for 300-W and 500-W PAR lamps. New fixtures utilizing these lamps, particularly 1-ft by 1-ft recessed lensed troffers (downlights), are becoming common in lobbies, shopping malls, and retail stores.

Including ballast losses, *high-pressure sodium lamps* have efficacies ranging from 50 LPW for the smallest lamps to 124 LPW for the largest lamps. Standard HPS lamps emit a yellow light and have very poor color rendition. Like MV lamps, HPS lamps are only used where good color is not a priority. The rated lifetimes of HPS lamps rival those of MV lamps and typically exceed 24,000 hr.

Ballasts

Because both fluorescent and HID lamps (discharge lamps) have a low resistance to the flow of electric current once the discharge arc is struck, they require some type of device to limit current flow. A lamp ballast is an electrical device used to control the current provided to the lamp. In most discharge lamps, a ballast also provides the high voltage necessary to start the lamp. Older "preheat" fluorescent lamps require a separate starter, but these lamps are becoming increasingly uncommon. In many HID ballasts, the ignitor used for starting the lamp is a replaceable module.

The most common types of ballasts are magnetic core-coil and electronic high-frequency ballasts. A *magnetic core-coil ballast* uses a transformer with a magnetic core coiled in copper or aluminum wire to control the current provided to a lamp. Magnetic ballasts operate at an input frequency of 60 Hz and operate lamps at the same 60 Hz. An *electronic high-frequency ballast* uses electronic circuitry rather than magnetic components to control current. Electronic ballasts use standard 60 Hz power but operate lamps at a much higher frequency (20,000 to 60,000 Hz). Both magnetic and electronic ballasts are available for most fluorescent lamp types.

Of the ballasts that are currently available for fluorescent lamps, the most efficient options are the electronic ballast and the cathode cutout ballast. Because an electronic ballast is more efficient than a standard core-coil magnetic ballast in transforming the input power to lamp requirements, and because fluorescent lamps are more efficient when operated at frequencies of 20,000 Hz or more, a lamp/ballast system using an electronic rather than magnetic ballast is more efficacious. Where there are two lamps per ballast, electronic ballast systems are approximately 20% more efficacious than magnetic ballast systems; where there is only one lamp per ballast, electronic ballast systems are almost 40% more efficacious than magnetic ballast systems.

In addition, electronic ballasts eliminate flicker, weigh less than magnetic ballasts, and operate more quietly. Since electronic ballasts are packaged in "cans" that are the same size as magnetic ballasts, they can be placed in fixtures designed to be used with magnetic ballasts. Electronic ballasts are widely available, but their use is limited because of their high cost and some technological limitations for certain applications. Fluorescent electronic ballasts are available for standard commercial-sector applications.

The *cathode cut-out (hybrid) ballast* is a modified magnetic ballast. It uses an electronic circuit to remove the filament power after the discharge has been initiated for rapid-start lamps. Cathode cut-out ballasts use approximately 5 to 10% less energy than energy-efficient magnetic ballasts. Almost all ballasts used for HID lamps are magnetic, and a number of different types are available. The various types differ primarily in how well they tolerate voltage swings and, in the case of HPS lamps, the increased voltage required to operate the lamp as it ages.

Lighting Fixtures

A lighting fixture is a housing for securing lamp(s) and ballast(s) and for controlling light distribution to a specific area. The function of the fixture is to distribute light to the desired area without causing glare or discomfort. The distribution of light is determined by the geometric design of the fixture as well as the material of which the reflector and/or lens is made. The more efficient a fixture is, the more light it emits from the lamp(s) within it. Although a lighting fixture is sometimes referred to as a luminaire, the term *luminaire* is most commonly used to refer to a complete lighting system including a lamp, ballast, and fixture.

Types of fluorescent lighting fixtures that are commonly used in the nonresidential sectors include recessed troffers, pendant-mounted indirect fixtures and indirect/direct fixtures, and surface-mounted commercial fixtures such as wraparound, strip, and industrial fixtures.

Most offices are equipped with *recessed troffers*, which are direct (downward) fixtures and emphasize horizontal surfaces. Many forms of optical control are possible with recessed luminaires. In the past, prismatic lenses were the preferred optical control because they offer high luminaire efficiency and uniform illuminance in the work space. Electronic offices have become increasingly common, however,

and the traditional direct lighting fixtures designed for typing and other horizontal tasks have become less useful because they tend to cause reflections on video display terminal (VDT) screens.

No lighting system reduces glare entirely, but some fixtures and/or components can reduce the amount of glare significantly. Because the glossy, vertical VDT screen can potentially reflect bright spots on the ceiling, and because VDT work is usually done with the head up, existing fixtures are sometimes replaced with indirect or direct/indirect fixtures, which produce light that is considered more visually comfortable. Most indirect lighting systems are suspended from the ceiling. They direct light toward the ceiling where the light is then reflected downward to provide a calm, diffuse light. Some people describe the indirect lighting as similar to the light on an overcast day, with no shadows or highlights. Generally, indirect lighting does not cause bright reflections on VDT screens. A *direct/indirect fixture* is suspended from the ceiling and provides direct light as well as indirect. These fixtures combine the high efficiency of direct lighting systems with the uniformity of light and lack of glare produced by indirect lighting systems.

A *wraparound fixture* has a prismatic lens that wraps around the bottom and sides of the lamp, and is always surface mounted rather than recessed. Wraparound fixtures are less expensive than other commercial fixtures and are typically used in areas where lighting control and distribution are not a priority. *Strip and industrial fixtures* are even less expensive and are typically used in places where light distribution is less important, such as large open areas (grocery stores, for example) and hallways. These are open fixtures in which the lamp is not hidden from view. The most common incandescent fixture in the nonresidential sector is the *downlight*, also known as a recessed can fixture. Fixtures designed for CFLs are available to replace downlight fixtures in areas where lighting control is less critical.

Lighting Efficiency*

There are seven techniques for improving the efficiency of lighting systems:

- Delamping
- Relamping
- Improved controls
- More efficient lamps and devices
- Task-oriented lighting
- Increased use of daylight
- Room color changes, lamp maintenance

The first two techniques and possibly the third are low in cost and may be considered operational changes. The last four items generally involve retrofit or new designs.

The first step in reviewing lighting electricity use is to perform a lighting survey. An inexpensive hand-held light meter can be used as a first approximation; however, distinction must be made between raw intensities (lux or footcandles) recorded in this way and *equivalent sphere illumination* (ESI) values.

Many variables can affect the "correct" lighting values for a particular task: task complexity, age of employee, glare, and so on. For reliable results, consult a lighting specialist or refer to the literature and publications of the Illuminating Engineering Society.

The lighting survey indicates those areas of the building where lighting is potentially inadequate or excessive. Deviations from adequate illumination levels can occur for several reasons: overdesign, building changes, change of occupancy, modified layout of equipment or personnel, more efficient lamps,

* It should be noted that the "Lighting Efficiency" discussion was written by a different author than the earlier material in the chapter, and that there are some discrepancies between the lamp efficacy ranges reported in the two sections. These discrepancies are likely the result of the different data sources used to calculate efficacy ranges, the different time periods during which they were calculated, and different assumptions regarding which lamp types are most typical.

improper use of equipment, dirt buildup, and so on. Once the building manager has identified areas with potentially excessive illumination levels, he or she can apply one or more of the seven techniques listed earlier. Each of these will be described briefly.

Delamping refers to the removal of lamps to reduce illumination to acceptable levels. With incandescent lamps, bulbs are removed. With fluorescent or HID lamps, ballasts account for 10 to 20% of total energy use and should be disconnected after lamps are removed.

Fluorescent lamps often are installed in luminaires in groups of two or more lamps where it is impossible to remove only one lamp. In such cases an artificial load (called a "phantom tube") can be installed in place of the lamp that has been removed.

Relamping refers to the replacement of existing lamps by lamps of lower wattage or increased efficiency. Low-wattage fluorescent tubes are available that require 15 to 20% less wattage (but produce

TABLE 20.8.2 Typical Relamping Opportunities

Change Office Lamps (2700 hr per year) From	To	Energy Savings/Cost Savings kWh To save annually	GJe	5¢ kWh
1 300-W incandescent	1 100-W mercury vapor	486	5.25	$24.30
2 100-W incandescent	1 40-W fluorescent	400	4.32	20.00
7 150-W incandescent	1 150-W sodium vapor	2360	25.5	$118.00
Change industrial lamps (3000 hr per year)				
1 300-W incandescent	2 40-W fluorescent	623	6.73	$31.15
1 100-W incandescent	2 215-W fluorescent	1617	17.5	80.85
3 300-W incandescent	1 250-W sodium vapor	1806	19.5	90.30
Change store lamps (3300 hr per year)				
1 300-W incandescent	2 40-W fluorescent	685	7.40	$34.25
1 200-W incandescent	1 100-W mercury vapor	264	2.85	13.20
2 200-W incandescent	1 175-W mercury vapor	670	7.24	33.50

10 to 15% less light). In some types of HID lamps, a more efficient lamp can be substituted directly. However, in most cases, ballasts must also be changed. Table 20.8.2 shows typical savings by relamping.

Improved controls permit lamps to be used only when and where needed. For example, certain office buildings have all lights for one floor on a single contactor. These lamps will be switched on at 6 A.M., before work begins, and are not turned off until 10 P.M., when maintenance personnel finish their cleanup duties. Energy usage can be cut by as much as 50% by installing individual switches for each office or work area, installing time clocks, installing occupancy sensors, using photocell controls, or instructing custodial crews to turn lights on as needed and turn them off when work is complete.

There is a great variation in the efficacy (a measure of light output per electricity input) of various lamps. Since incandescent lamps have the lowest efficacy, typically 8 to 20 LPW, wherever possible, fluorescent lamps should be substituted. This not only saves energy but also offers economic savings, since fluorescent lamps last 10 to 50 times longer. Fluorescent lamps have efficacies in the range of 30 to 90 LPW.

Compact fluorescent lamps are available as substitutes for a wide range of incandescent lamps. They range in wattage from 5 to 25 W with efficacies of 26 to 58 lm/W and will replace 25- to 100-W incandescent lamps. In addition to the energy savings, they have a 10,000-hr rated life and do not need to be replaced as often as incandescent lamps. Exit lights are good candidates for compact fluorescent lamps. Conversion kits are available to replace the incandescent lamps. Payback is rapid because of the lower energy use and lower maintenance cost, since these lamps are normally on 24 hr a day. There are also light-emitting diode exit lights that are very energy efficient.

Still greater improvements are possible with HID lamps such as mercury vapor, metal halide, and high-pressure sodium lamps. Although they are generally not suited to residential use high light output and high capital cost) they are increasingly used in commercial buildings. They have efficacies in the range of 25 to 124 LPW.

Improved ballasts are another way of saving lighting energy. A comparison of the conventional magnetic ballast with improved ballasts shows the difference:

Type Lamp	**2 Lamp 40 W**	**F40 T-12 CW**	**2 Lamp 32 W**	**F32 T-8**
Ballast type	Standard magnetic	Energy-efficient magnetic	Electronic	Electronic
Input watts	96	88	73	64
Efficacy	60 lm/W	65 lm/W	78 lm/W	90 lm/W

The best performance comes from electronic ballasts, which operate at higher frequency. In addition to the lighting energy savings, there are additional savings from the reduced air-conditioning load due to less heat output from the ballasts. The environmental benefit for the electronic ballast described earlier, as estimated by the U.S. Environmental Protection Agency, is a reduction in CO_2 production of 150 lb/year, a reduction of 0.65 lb/year of SO_2, and a reduction of 0.4 lb/year of NO_x.

In certain types of buildings and operations, daylighting can be utilized to reduce (if not replace) electric lighting. Techniques include windows, an atrium, skylights, and so on. There are obvious limitations such as those imposed by the need for privacy, 24-hr operation, and building core locations with no access to natural light. Also, the use of light colors can substantially enhance illumination without modifying existing lamps.

An effective lamp maintenance program can also have important benefits. Light output gradually decreases over lamp lifetime. This should be considered in the initial design and when deciding on lamp replacement. Dirt can substantially reduce light output; simply cleaning lamps and luminaries more frequently can gain up to 5 to 10% greater illumination, permitting some lamps to be removed.

Reflectors are available to insert in fluorescent lamp fixtures. These direct and focus the light onto the work area, yielding a greater degree of illumination. Alternatively, in multiple lamp fixtures, one lamp can be removed and the reflector keeps the level of illumination at about 75% of the previous value.

Defining Terms

Ballast: A lamp ballast is an electrical device used to control the current provided to the lamp. In most discharge lamps, a ballast also provides the high voltage necessary to start the lamp.

Color Rendering Index (CRI): A measure of how surface colors appear when illuminated by a lamp compared to how they appear when illuminated by a reference source of the same color temperature. For color temperature above 5000 K, the reference source is a standard daylight condition of the same color temperature; below 5000 K, the reference source is a blackbody radiator.

Color temperature: The color of a lamp's light is described by its color temperature, expressed in degrees Kelvin (K). The concept of color temperature is based on the fact that the emitted radiation spectrum of a blackbody radiator depends on temperature alone. The color temperature of a lamp is the temperature at which an ideal blackbody radiator would emit light that is the same color as the light of the lamp.

Efficacy: The ratio of the amount of light emitted (lumens) to the power (watts) drawn by a lamp. The unit used to express lamp efficacy is lumens per watt (LPW).

Lumen maintenance: Refers to the extent to which a lamp sustains its lumen output, and therefore efficacy, over time.

References

Atkinson, B., McMahon, J., Mills, E., Chan, P., Chan, T., Eto, J., Jennings, J., Koomey, J., Lo, K., Lecar, M., Price, L., Rubinstein, F., Sezgen, O., and Wenzel, T. 1992. *Analysis of Federal Policy Options for Improving U.S. Lighting Energy Efficiency: Commercial and Residential Buildings*. Lawrence Berkeley National Laboratory, Berkeley, CA. LBL-31469.

Audin, L., Houghton, D., Shepard, M., and Hawthorne, W. 1994. *Lighting Technology Atlas*. E-Source, Snowmass, CO.

Illuminating Engineering Society of North Amercia. 1993. *Lighting Handbook*, 8th ed., M. Rhea, Ed., Illuminating Engineering Society of North America, New York.

Leslie, R. and Conway, K. 1993. *The Lighting Pattern Book for Homes*, Lighting Research Center, Rensselaer Polytechnic Institute, Troy, NY.

Turiel, I., Atkinson, B., Boghosian, S., Chan, P., Jennings, J., Lutz, J., McMahon, J., and Roenquist, G. 1995. *Evaluation of Advanced Technologies for Residential Appliances and Residential and Commercial Lighting*. Lawrence Berkely National Laboratory, Berkeley, CA. LBL-35982.

Further Information

For additional information on performance characteristics of lamps, ballasts, lighting fixtures, and controls, the reader is referred to the CRC *Handbook of Energy Efficiency* from which this section has been extracted. If interested in more discussions of energy-efficient lighting design strategies and technologies, consult the references and request information from the Green Lights program: Green Lights U.S. EPA, Air and Radiation (6202-J), Washington, D.C., 20460.

Norton, P. “Appendices”
Mechanical Engineering Handbook
Ed. Frank Kreith
Boca Raton: CRC Press LLC, 1999

Appendices

Paul Norton

National Renewable Energy Laboratory

Appendix A. Properties of Gases and Vapors

TABLE A.1 Properties of Dry Air at Atmospheric Pressure

Symbols and Units:

K = absolute temperature, degrees Kelvin
deg C = temperature, degrees Celsius
deg F = temperature, degrees Fahrenheit
ρ = density, kg/m^3
c_p = specific heat capacity, kJ/kg·K
c_p/c_v = specific heat capacity ratio, dimensionless
μ = viscosity, N·s/m^2 × 10^6 (For N·s/m^2 (= kg/m·s) multiply tabulated values by 10^{-6})
k = thermal conductivity, W/m·k × 10^3 (For W/m·K multiply tabulated values by 10^{-3})
Pr = Prandtl number, dimensionless
h = enthalpy, kJ/kg
V_s = sound velocity, m/s

Temperature			*Properties*							
K	*deg C*	*deg F*	ρ	c_p	c_p/c_v	μ	k	Pr	h	V_s
100	−173.15	−280	3.598	1.028		6.929	9.248	.770	98.42	198.4
110	−163.15	−262	3.256	1.022	1.420 2	7.633	10.15	.768	108.7	208.7
120	−153.15	−244	2.975	1.017	1.416 6	8.319	11.05	.766	118.8	218.4
130	−143.15	−226	2.740	1.014	1.413 9	8.990	11.94	.763	129.0	227.6
140	−133.15	−208	2.540	1.012	1.411 9	9.646	12.84	.761	139.1	236.4
150	−123.15	−190	2.367	1.010	1.410 2	10.28	13.73	.758	149.2	245.0
160	−113.15	−172	2.217	1.009	1.408 9	10.91	14.61	.754	159.4	253.2
170	−103.15	−154	2.085	1.008	1.407 9	11.52	15.49	.750	169.4	261.0
180	−93.15	−136	1.968	1.007	1.407 1	12.12	16.37	.746	179.5	268.7
190	−83.15	−118	1.863	1.007	1.406 4	12.71	17.23	.743	189.6	276.2
200	−73.15	−100	1.769	1.006	1.405 7	13.28	18.09	.739	199.7	283.4
205	−68.15	−91	1.726	1.006	1.405 5	13.56	18.52	.738	204.7	286.9
210	−63.15	−82	1.684	1.006	1.405 3	13.85	18.94	.736	209.7	290.5
215	−58.15	−73	1.646	1.006	1.405 0	14.12	19.36	.734	214.8	293.9
220	−53.15	−64	1.607	1.006	1.404 8	14.40	19.78	.732	219.8	297.4
225	−48.15	−55	1.572	1.006	1.404 6	14.67	20.20	.731	224.8	300.8
230	−43.15	−46	1.537	1.006	1.404 4	14.94	20.62	.729	229.8	304.1
235	−38.15	−37	1.505	1.006	1.404 2	15.20	21.04	.727	234.9	307.4
240	−33.15	−28	1.473	1.005	1.404 0	15.47	21.45	.725	239.9	310.6
245	−28.15	−19	1.443	1.005	1.403 8	15.73	21.86	.724	244.9	313.8
250	−23.15	−10	1.413	1.005	1.403 6	15.99	22.27	.722	250.0	317.1
255	−18.15	−1	1.386	1.005	1.403 4	16.25	22.68	.721	255.0	320.2
260	−13.15	8	1.359	1.005	1.403 2	16.50	23.08	.719	260.0	323.4
265	−8.15	17	1.333	1.005	1.403 0	16.75	23.48	.717	265.0	326.5
270	−3.15	26	1.308	1.006	1.402 9	17.00	23.88	.716	270.1	329.6
275	+1.85	35	1.285	1.006	1.402 6	17.26	24.28	.715	275.1	332.6
280	6.85	44	1.261	1.006	1.402 4	17.50	24.67	.713	280.1	335.6
285	11.85	53	1.240	1.006	1.402 2	17.74	25.06	.711	285.1	338.5
290	16.85	62	1.218	1.006	1.402 0	17.98	25.47	.710	290.2	341.5
295	21.85	71	1.197	1.006	1.401 8	18.22	25.85	.709	295.2	344.4
300	26.85	80	1.177	1.006	1.401 7	18.46	26.24	.708	300.2	347.3
305	31.85	89	1.158	1.006	1.401 5	18.70	26.63	.707	305.3	350.2
310	36.85	98	1.139	1.007	1.401 3	18.93	27.01	.705	310.3	353.1
315	41.85	107	1.121	1.007	1.401 0	19.15	27.40	.704	315.3	355.8
320	46.85	116	1.103	1.007	1.400 8	19.39	27.78	.703	320.4	358.7

*Condensed and computed from: "Tables of Thermal Properties of Gases", National Bureau of Standards Circular 564, U.S. Government Printing Office, November 1955.

TABLE A.1 (continued) Properties of Dry Air at Atmospheric Pressure

Temperature			*Properties*							
K	*deg C*	*deg F*	ρ	c_p	c_p/c_v	μ	k	*Pr*	h	V_s
325	51.85	125	1.086	1.008	1.400 6	19.63	28.15	.702	325.4	361.4
330	56.85	134	1.070	1.008	1.400 4	19.85	28.53	.701	330.4	364.2
335	61.85	143	1.054	1.008	1.400 1	20.08	28.90	.700	335.5	366.9
340	66.85	152	1.038	1.008	1.399 9	20.30	29.28	.699	340.5	369.6
345	71.85	161	1.023	1.009	1.399 6	20.52	29.64	.698	345.6	372.3
350	76.85	170	1.008	1.009	1.399 3	20.75	30.03	.697	350.6	375.0
355	81.85	179	0.994 5	1.010	1.399 0	20.97	30.39	.696	355.7	377.6
360	86.85	188	0.980 5	1.010	1.398 7	21.18	30.78	.695	360.7	380.2
365	91.85	197	0.967 2	1.010	1.398 4	21.38	31.14	.694	365.8	382.8
370	96.85	206	0.953 9	1.011	1.398 1	21.60	31.50	.693	370.8	385.4
375	101.85	215	0.941 3	1.011	1.397 8	21.81	31.86	.692	375.9	388.0
380	106.85	224	0.928 8	1.012	1.397 5	22.02	32.23	.691	380.9	390.5
385	111.85	233	0.916 9	1.012	1.397 1	22.24	32.59	.690	386.0	393.0
390	116.85	242	0.905 0	1.013	1.396 8	22.44	32.95	.690	391.0	395.5
395	121.85	251	0.893 6	1.014	1.396 4	22.65	33.31	.689	396.1	398.0
400	126.85	260	0.882 2	1.014	1.396 1	22.86	33.65	.689	401.2	400.4
410	136.85	278	0.860 8	1.015	1.395 3	23.27	34.35	.688	411.3	405.3
420	146.85	296	0.840 2	1.017	1.394 6	23.66	35.05	.687	421.5	410.2
430	156.85	314	0.820 7	1.018	1.393 8	24.06	35.75	.686	431.7	414.9
440	166.85	332	0.802 1	1.020	1.392 9	24.45	36.43	.684	441.9	419.6
450	176.85	350	0.784 2	1.021	1.392 0	24.85	37.10	.684	452.1	424.2
460	186.85	368	0.767 7	1.023	1.391 1	25.22	37.78	.683	462.3	428.7
470	196.85	386	0.750 9	1.024	1.390 1	25.58	38.46	.682	472.5	433.2
480	206.85	404	0.735 1	1.026	1.389 2	25.96	39.11	.681	482.8	437.6
490	216.85	422	0.720 1	1.028	1.388 1	26.32	39.76	.680	493.0	442.0
500	226.85	440	0.705 7	1.030	1.387 1	26.70	40.41	.680	503.3	446.4
510	236.85	458	0.691 9	1.032	1.386 1	27.06	41.06	.680	513.6	450.6
520	246.85	476	0.678 6	1.034	1.385 1	27.42	41.69	.680	524.0	454.9
530	256.85	494	0.665 8	1.036	1.384 0	27.78	42.32	.680	534.3	459.0
540	266.85	512	0.653 5	1.038	1.382 9	28.14	42.94	.680	544.7	463.2
550	276.85	530	0.641 6	1.040	1.381 8	28.48	43.57	.680	555.1	467.3
560	286.85	548	0.630 1	1.042	1.380 6	28.83	44.20	.680	565.5	471.3
570	296.85	566	0.619 0	1.044	1.379 5	29.17	44.80	.680	575.9	475.3
580	306.85	584	0.608 4	1.047	1.378 3	29.52	45.41	.680	586.4	479.2
590	316.85	602	0.598 0	1.049	1.377 2	29.84	46.01	.680	596.9	483.2
600	326.85	620	0.588 1	1.051	1.376 0	30.17	46.61	.680	607.4	486.9
620	346.85	656	0.569 1	1.056	1.373 7	30.82	47.80	.681	628.4	494.5
640	366.85	692	0.551 4	1.061	1.371 4	31.47	48.96	.682	649.6	502.1
660	386.85	728	0.534 7	1.065	1.369 1	32.09	50.12	.682	670.9	509.4
680	406.85	764	0.518 9	1.070	1.366 8	32.71	51.25	.683	692.2	516.7
700	426.85	800	0.504 0	1.075	1.364 6	33.32	52.36	.684	713.7	523.7
720	446.85	836	0.490 1	1.080	1.362 3	33.92	53.45	.685	735.2	531.0
740	466.85	872	0.476 9	1.085	1.360 1	34.52	54.53	.686	756.9	537.6
760	486.85	908	0.464 3	1.089	1.358 0	35.11	55.62	.687	778.6	544.6
780	506.85	944	0.452 4	1.094	1.355 9	35.69	56.68	.688	800.5	551.2
800	526.85	980	0.441 0	1.099	1.354	36.24	57.74	.689	822.4	557.8
850	576.85	1 070	0.415 2	1.110	1.349	37.63	60.30	.693	877.5	574.1
900	626.85	1 160	0.392 0	1.121	1.345	38.97	62.76	.696	933.4	589.6
950	676.85	1 250	0.371 4	1.132	1.340	40.26	65.20	.699	989.7	604.9
1 000	726.85	1 340	0.352 9	1.142	1.336	41.53	67.54	.702	1 046	619.5
1 100	826.85	1 520	0.320 8	1.161	1.329	43.96			1 162	648.0
1 200	926.85	1 700	0.294 1	1.179	1.322	46.26			1 279	675.2
1 300	1 026.85	1 880	0.271 4	1.197	1.316	48.46			1 398	701.0
1 400	1 126.85	2 060	0.252 1	1.214	1.310	50.57			1 518	725.9
1 500	1 220.85	2 240	0.235 3	1.231	1.304	52.61			1 640	749.4
1 600	1 326.85	2 420	0.220 6	1.249	1.299	54.57			1 764	772.6
1 800	1 526.85	2 780	0.196 0	1.288	1.288	58.29			2 018	815.7
2 000	1 726.85	3 140	0.176 4	1.338	1.274				2 280	855.5
2 400	2 126.85	3 860	0.146 7	1.574	1.238				2 853	924.4
2 800	2 526.85	4 580	0.124 5	2.259	1.196				3 599	983.1

TABLE A.2 Ideal Gas Properties of Nitrogen, Oxygen, and Carbon Dioxide

Symbols and Units:

T = absolute temperature, degrees Kelvin
$\bar{h}$ = enthalpy, kJ/kmol
$\bar{u}$ = internal energy, kJ/kmol
$\bar{s}°$ = absolute entropy at standard reference pressure, kJ/kmol K
[$\bar{h}$ = enthalpy of formation per mole at standard state = 0 kJ/kmol]

Part a. Ideal Gas Properties of Nitrogen, N_2

T	$\bar{h}$	$\bar{u}$	$\bar{s}°$	T	$\bar{h}$	$\bar{u}$	$\bar{s}°$
0	0	0	0	600	17,563	12,574	212.066
220	6,391	4,562	182.639	610	17,864	12,792	212.564
230	6,683	4,770	183.938	620	18,166	13,011	213.055
240	6,975	4,979	185.180	630	18,468	13,230	213.541
250	7,266	5,188	186.370	640	18,772	13,450	214.018
260	7,558	5,396	187.514	650	19,075	13,671	214.489
270	7,849	5,604	188.614	660	19,380	13,892	214.954
280	8,141	5,813	189.673	670	19,685	14,114	215.413
290	8,432	6,021	190.695	680	19,991	14,337	215.866
298	8,669	6,190	191.502	690	20,297	14,560	216.314
300	8,723	6,229	191.682	700	20,604	14,784	216.756
310	9,014	6,437	192.638	710	20,912	15,008	217.192
320	9,306	6,645	193.562	720	21,220	15,234	217.624
330	9,597	6,853	194.459	730	21,529	15,460	218.059
340	9,888	7,061	195.328	740	21,839	15,686	218.472
350	10,180	7,270	196.173	750	22,149	15,913	218.889
360	10,471	7,478	196.995	760	22,460	16,141	219.301
370	10,763	7,687	197.794	770	22,772	16,370	219.709
380	11,055	7,895	198.572	780	23,085	16,599	220.113
390	11,347	8,104	199.331	790	23,398	16,830	220.512
400	11,640	8,314	200.071	800	23,714	17,061	220.907
410	11,932	8,523	200.794	810	24,027	17,292	221.298
420	12,225	8,733	201.499	820	24,342	17,524	221.684
430	12,518	8,943	202.189	830	24,658	17,757	222.067
440	12,811	9,153	202.863	840	24,974	17,990	222.447
450	13,105	9,363	203.523	850	25,292	18,224	222.822
460	13,399	9,574	204.170	860	25,610	18,459	223.194
470	13,693	9,786	204.803	870	25,928	18,695	223.562
480	13,988	9,997	205.424	880	26,248	18,931	223.927
490	14,285	10,210	206.033	890	26,568	19,168	224.288
500	14,581	10,423	206.630	900	26,890	19,407	224.647
510	14,876	10,635	207.216	910	27,210	19,644	225.002
520	15,172	10,848	207.792	920	27,532	19,883	225.353
530	15,469	11,062	208.358	930	27,854	20,122	225.701
540	15,766	11,277	208.914	940	28,178	20,362	226.047
550	16,064	11,492	209.461	950	28,501	20,603	226.389
560	16,363	11,707	209.999	960	28,826	20,844	226.728
570	16,662	11,923	210.528	970	29,151	21,086	227.064
580	16,962	12,139	211.049	980	29,476	21,328	227.398
590	17,26_	12,356	211.562	990	29,803	21,571	227.728

Source: Adapted from M.J. Moran and H.N. Shapiro, *Fundamentals of Engineering Thermodynamics*, 3rd. ed., Wiley, New York, 1995, as presented in K. Wark. *Thermodynamics*, 4th ed., McGraw-Hill, New York, 1983, based on the *JANAF Thermochemical Tables*, NSRDS-NBS-37, 1971.

TABLE A.2 (continued) Ideal Gas Properties of Nitrogen, Oxygen, and Carbon Dioxide

T	$\bar{h}$	$\bar{u}$	$\bar{s}^\circ$	T	n	$\bar{u}$	$\bar{s}^\circ$
1000	30,129	21,815	228.057	1760	56,227	41.594	247.396
1020	30,784	22,304	228.706	1780	56,938	42,139	247.798
1040	31,442	22,795	229.344	1800	57,651	42,685	248.195
1060	32,101	23,288	229.973	1820	58,363	43,231	248.589
1080	32,762	23,782	230.591	1840	59,075	43,777	248.979
1100	33,426	24,280	231.199	1860	59,790	44,324	249.365
1120	34,092	24,780	231.799	1880	60,504	44,873	249.748
1140	34,760	25,282	232.391	1900	61,220	45,423	250.128
1160	35,430	25,786	232.973	1920	61,936	45,973	250.502
1180	36,104	26,291	233.549	1940	62,654	46,524	250.874
1200	36,777	26,799	234.115	1960	63,381	47,075	251.242
1220	37,452	27,308	234.673	1980	64,090	47,627	251.607
1240	38,129	27,819	235.223	2000	64,810	48,181	251.969
1260	38,807	28,331	235.766	2050	66,612	49,567	252.858
1280	39,488	28,845	236.302	2100	68,417	50,957	253.726
1300	40,170	29,361	236.831	2150	70,226	52,351	254.578
1320	40,853	29,878	237.353	2200	72,040	53,749	255.412
1340	41,539	30,398	237.867	2250	73,856	55,149	256.227
1360	42,227	30,919	238.376	2300	75,676	56,553	257.027
1380	42,915	31,441	238.878	2350	77,496	57,958	257.810
1400	43,605	31,964	239.375	2400	79,320	59,366	258.580
1420	44,295	32,489	239.865	2450	81,149	60,779	259.332
1440	44,988	33,014	240.350	2500	82,981	62,195	260.073
1460	45,682	33,543	240.827	2550	84,814	63,613	260.799
1480	46,377	34,071	241.301	2600	86,650	65,033	261.512
1500	47,073	34,601	241.768	2650	88,488	66,455	262.213
1520	47,771	35,133	242.228	2700	90,328	67,880	262.902
1540	48,470	35,665	242.685	2750	92,171	69,306	263.577
1560	49,168	36,197	243.137	2800	94,014	70,734	264.241
1580	49,869	36,732	243.585	2850	95,859	72,163	264.895
1600	50,571	37,268	244.028	2900	97,705	73,593	265.538
1620	51,275	37,806	244.464	2950	99,556	75,028	266.170
1640	51,980	38,344	244.896	3000	101,407	76,464	266.793
1660	52,686	38,884	245.324	3050	103,260	77,902	267.404
1680	53,393	39,424	245.747	3100	105,115	79,341	268.007
1700	54,099	39,965	246.166	3150	106,972	80,782	268.601
1720	54,807	40,507	246.580	3200	108,830	82,224	269.186
1740	55,516	41,049	246.990	3250	110,690	83,668	269.763

TABLE A.2 (continued) Ideal Gas Properties of Nitrogen, Oxygen, and Carbon Dioxide

Part b. Ideal Gas Properties of Oxygen, O_2

T	$\bar{h}$	$\bar{u}$	$\bar{s}^\circ$	T	$\bar{h}$	$\bar{u}$	$\bar{s}^\circ$
0	0	0	0	600	17,929	12,940	226.346
220	6,404	4,575	196.171	610	18,250	13,178	226.877
230	6,694	4,782	197.461	620	18,572	13,417	227.400
240	6,984	4,989	198.696	630	18,895	13,657	227.918
250	7,275	5,197	199.885	640	19,219	13,898	228.429
260	7,566	5,405	201.027	650	19,544	14,140	228.932
270	7,858	5,613	202.128	660	19,870	14,383	229.430
280	8,150	5,822	203.191	670	20,197	14,626	229.920
290	8,443	6,032	204.218	680	20,524	14,871	230.405
298	8,682	6,203	205.033	690	20,854	15,116	230.885
300	8,736	6,242	205.213	700	21,184	15,364	231.358
310	9,030	6,453	206.177	710	21,514	15,611	231.827
320	9,325	6,664	207.112	720	21,845	15,859	232.291
330	9,620	6,877	208.020	730	22,177	16,107	232.748
340	9,916	7,090	208.904	740	22,510	16,357	233.201
350	10,213	7,303	209.765	750	22,844	16,607	233.649
360	10,511	7,518	210.604	760	23,178	16,859	234.091
370	10,809	7,733	211.423	770	23,513	17,111	234.528
380	11,109	7,949	212.222	780	23,850	17,364	234.960
390	11,409	8,166	213.002	790	24,186	17,618	235.387
400	11,711	8,384	213.765	800	24,523	17,872	235.810
410	12,012	8,603	214.510	810	24,861	18,126	236.230
420	12,314	8,822	215.241	820	25,199	18,382	236.644
430	12,618	9,043	215.955	830	25,537	18,637	237.055
440	12,923	9,264	216.656	840	25,877	18,893	237.462
450	13,228	9,487	217.342	850	26,218	19,150	237.864
460	13,535	9,710	218.016	860	26,559	19,408	238.264
470	13,842	9,935	218.676	870	26,899	19,666	238.660
480	14,151	10,160	219.326	880	27,242	19,925	239.051
490	14,460	10,386	219.963	890	27,584	20,185	239.439
500	14,770	10,614	220.589	900	27,928	20,445	239.823
510	15,082	10,842	221.206	910	28,272	20,706	240.203
520	15,395	11,071	221.812	920	28,616	20,967	240.580
530	15,708	11,301	222.409	930	28,960	21,228	240.953
540	16,022	11,533	222.997	940	29,306	21,491	241.323
550	16,338	11,765	223.576	950	29,652	21,754	241.689
560	16,654	11,998	224.146	960	29,999	22,017	242.052
570	16,971	12,232	224.708	970	30,345	22,280	242.411
580	17,290	12,467	225.262	980	30,692	22,544	242.768
590	17,609	12,703	225.808	990	31,041	22,809	243.120

TABLE A.2 (continued) Ideal Gas Properties of Nitrogen, Oxygen, and Carbon Dioxide

T	$\bar{h}$	$\bar{u}$	$\bar{s}°$	T	$\bar{h}$	$\bar{u}$	$\bar{s}°$
1000	31,389	23,075	243.471	1760	58,880	44,247	263.861
1020	32,088	23,607	244.164	1780	59,624	44,825	264.283
1040	32,789	24,142	244.844	1800	60,371	45,405	264.701
1060	33,490	24,677	245.513	1820	61,118	45,986	265.113
1080	34,194	25,214	246.171	1840	61,866	46,568	265.521
1100	34,899	25,753	246.818	1860	62,616	47,151	265.925
1120	35,606	26,294	247.454	1880	63,365	47,734	266.326
1140	36,314	26,836	248.081	1900	64,116	48,319	266.722
1160	37,023	27,379	248.698	1920	64,868	48,904	267.115
1180	37,734	27,923	249.307	1940	65,620	49,490	267.505
1200	38,447	28,469	249.906	1960	66,374	50,078	267.891
1220	39,162	29,018	250.497	1980	67,127	50,665	268.275
1240	39,877	29,568	251.079	2000	67,881	51,253	268.655
1260	40,594	30,118	251.653	2050	69,772	52,727	269.588
1280	41,312	30,670	252.219	2100	71,668	54,208	270.504
1300	42,033	31,224	252.776	2150	73,573	55,697	271.399
1320	42,753	31,778	253.325	2200	75,484	57,192	272.278
1340	43,475	32,334	253.868	2250	77,397	58,690	273.136
1360	44,198	32,891	254.404	2300	79,316	60,193	273.981
1380	44,923	33,449	254.932	2350	81,243	61,704	274.809
1400	45,648	34,008	255.454	2400	83,174	63,219	275.625
1420	46,374	34,567	255.968	2450	85,112	64,742	276.424
1440	47,102	35,129	256.475	2500	87,057	66,271	277.207
1460	47,831	35,692	256.978	2550	89,004	67,802	277.979
1480	48,561	36,256	257.474	2600	90,956	69,339	278.738
1500	49,292	36,821	257.965	2650	92,916	70,883	279.485
1520	50,024	37,387	258.450	2700	94,881	72,433	280.219
1540	50,756	37,952	258.928	2750	96,852	73,987	280.942
1560	51,490	38,520	259.402	2800	98,826	75,546	281.654
1580	52,224	39,088	259.870	2850	100,808	77,112	282.357
1600	52,961	39,658	260.333	2900	102,793	78,682	283.048
1620	53,696	40,227	260.791	2950	104,785	80,258	283.728
1640	54,434	40,799	261.242	3000	106,780	81,837	284.399
1660	55,172	41,370	261.690	3050	108,778	83,419	285.060
1680	55,912	41,944	262.132	3100	110,784	85,009	285.713
1700	56,652	42,517	262.571	3150	112,795	86,601	286.355
1720	57,394	43,093	263.005	3200	114,809	88,203	286.989
1740	58,136	43,669	263.435	3250	116,827	89,804	287.614

TABLE A.2 (continued) Ideal Gas Properties of Nitrogen, Oxygen, and Carbon Dioxide

Part c. Ideal Gas Properties of Carbon Dioxide, CO_2

T	$\bar{h}$	$\bar{u}$	$\bar{s}^\circ$	T	$\bar{h}$	$\bar{u}$	$\bar{s}^\circ$
0	0	0	0	600	22,280	17,291	243.199
220	6,601	4,772	202.966	610	22,754	17,683	243.983
230	6,938	5,026	204.464	620	23,231	18,076	244.758
240	7,280	5,285	205.920	630	23,709	18,471	245.524
250	7,627	5,548	207.337	640	24,190	18,869	246.282
260	7,979	5,817	208.717	650	24,674	19,270	247.032
270	8,335	6,091	210.062	660	25,160	19,672	247.773
280	8,697	6,369	211.376	670	25,648	20,078	248.507
290	9,063	6,651	212.660	680	26,138	20,484	249.233
298	9,364	6,885	213.685	690	26,631	20,894	249.952
300	9,431	6,939	213.915	700	27,125	21,305	250.663
310	9,807	7,230	215.146	710	27,622	21,719	251.368
320	10,186	7,526	216.351	720	28,121	22,134	252.065
330	10,570	7,826	217.534	730	28,622	22,552	252.755
340	10,959	8,131	218.694	740	29,124	22,972	253.439
350	11,351	8,439	219.831	750	29,629	23,393	254.117
360	11,748	8,752	220.948	760	30,135	23,817	254.787
370	12,148	9,068	222.044	770	30,644	24,242	255.452
380	12,552	9,392	223.122	780	31,154	24,669	256.110
390	12,960	9,718	224.182	790	31,665	25,097	256.762
400	13,372	10,046	225.225	800	32,179	25,527	257.408
410	13,787	10,378	226.250	810	32,694	25,959	258.048
420	14,206	10,714	227.258	820	33,212	26,394	258.682
430	14,628	11,053	228.252	830	33,730	26,829	259.311
440	15,054	11,393	229.230	840	34,251	27,267	259.934
450	15,483	11,742	230.194	850	34,773	27,706	260.551
460	15,916	12,091	231.144	860	35,296	28,125	261.164
470	16,351	12,444	232.080	870	35,821	28,588	261.770
480	16,791	12,800	233.004	880	36,347	29,031	262.371
490	17,232	13,158	233.916	890	36,876	29,476	262.968
500	17,678	13,521	234.814	900	37,405	29,922	263.559
510	18,126	13,885	235.700	910	37,935	30,369	264.146
520	18,576	14,253	236.575	920	38,467	30,818	264.728
530	19,029	14,622	237.439	930	39,000	31,268	265.304
540	19,485	14,996	238.292	940	39,535	31,719	265.877
550	19,945	15,372	239.135	950	40,070	32,171	266.444
560	20,407	15,751	239.962	960	40,607	32,625	267.007
570	20,870	16,131	240.789	970	41,145	33,081	267.566
580	21,337	16,515	241.602	980	41,685	33,537	268.119
590	21,807	16,902	242.405	990	42,226	33,995	268.670

TABLE A.2 (continued) Ideal Gas Properties of Nitrogen, Oxygen, and Carbon Dioxide

T	$\bar{h}$	$\bar{u}$	$\bar{s}^\circ$	T	$\bar{h}$	$\bar{u}$	$\bar{s}^\circ$
1000	42,769	34,455	269.215	1760	86,420	71,787	301.543
1020	43,859	35,378	270.293	1780	87,612	72,812	302.271
1040	44,953	36,306	271.354	1800	88,806	73,840	302.884
1060	46,051	37,238	272.400	1820	90,000	74,868	303.544
1080	47,153	38,174	273.430	1840	91,196	75,897	304.198
1100	48,258	39,112	274.445	1860	92,394	76,929	304.845
1120	49,369	40,057	275.444	1880	93,593	77,962	305.487
1140	50,484	41,006	276.430	1900	94,793	78,996	306.122
1160	51,602	41,957	277.403	1920	95,995	80,031	306.751
1180	52,724	42,913	278.362	1940	97,197	81,067	307.374
1200	53,848	43,871	279.307	1960	98,401	82,105	307.992
1220	54,977	44,834	280.238	1980	99,606	83,144	308.604
1240	56,108	45,799	281.158	2000	100,804	84,185	309.210
1260	57,244	46,768	282.066	2050	103,835	86,791	310.701
1280	58,381	47,739	282.962	2100	106,864	89,404	312.160
1300	59,522	48,713	283.847	2150	109,898	92,023	313.589
1320	60,666	49,691	284.722	2200	112,939	94,648	314.988
1340	61,813	50,672	285.586	2250	115,984	97,277	316.356
1360	62,963	51,656	286.439	2300	119,035	99,912	317.695
1380	64,116	52,643	287.283	2350	122,091	102,552	319.011
1400	65,271	53,631	288.106	2400	125,152	105,197	320.302
1420	66,427	54,621	288.934	2450	128,219	107,849	321.566
1440	67,586	55,614	289.743	2500	131,290	110,504	322.808
1460	68,748	56,609	290.542	2550	134,368	113,166	324.026
1480	69,911	57,606	291.333	2600	137,449	115,832	325.222
1500	71,078	58,606	292.114	2650	140,533	118,500	326.396
1520	72,246	59,609	292.888	2700	143,620	121,172	327.549
1540	73,417	60,613	292.654	2750	146,713	123,849	328.684
1560	74,590	61,620	294.411	2800	149,808	126,528	329.800
1580	76,767	62,630	295.161	2850	152,908	129,212	330.896
1600	76,944	63,741	295.901	2900	156,009	131,898	331.975
1620	78,123	64,653	296.632	2950	159,117	134,589	333.037
1640	79,303	65,668	297.356	3000	162,226	137,283	334.084
1660	80,486	66,592	298.072	3050	165,341	139,982	335.114
1680	81,670	67,702	298.781	3100	168,456	142,681	336.126
1700	82,856	68,721	299.482	3150	171,576	145,385	337.124
1720	84,043	69,742	300.177	3200	174,695	148,089	338.109
1740	85,231	70,764	300.863	3250	177,822	150,801	339.069

TABLE A.3 Psychrometric Table: Properties of Moist Air at 101 325 N/m²

Symbols and Units:

P_s = pressure of water vapor at saturation, N/m²
W_s = humidity ratio at saturation, mass of water vapor associated with unit mass of dry air
V_a = specific volume of dry air, m³/kg
V_s = specific volume of saturated mixture, m³/kg dry air
h_a^a = specific enthalpy of dry air, kJ/kg
h_s = specific enthalpy of saturated mixture, kJ/kg dry air
s_s = specific entropy of saturated mixture, J/K·kg dry air

Temperature			*Properties*						
C	*K*	*F*	P_s	W_s	V_a	V_s	h_a	h_s	s_s
−40	233.15	−40	12.838	0.000 079 25	0.659 61	0.659 68	−22.35	−22.16	−90.659
−30	243.15	−22	37.992	0.000 234 4	0.688 08	0.688 33	−12.29	−11.72	−46.732
−25	248.15	−13	63.248	0.000 390 3	0.702 32	0.702 75	−7.265	−6.306	−24.706
−20	253.15	−4	103.19	0.000 637 1	0.716 49	0.717 24	−2.236	−0.6653	−2.2194
−15	258.15	+5	165.18	0.001 020	0.730 72	0.731 91	+2.794	5.318	21.189
−10	263.15	14	259.72	0.001 606	0.744 95	0.746 83	7.823	11.81	46.104
−5	268.15	23	401.49	0.002 485	0.759 12	0.762 18	12.85	19.04	73.365
0	273.15	32	610.80	0.003 788	0.773 36	0.778 04	17.88	27.35	104.14
5	278.15	41	871.93	0.005 421	0.787 59	0.794 40	22.91	36.52	137.39
10	283.15	50	1 227.2	0.007 658	0.801 76	0.811 63	27.94	47.23	175.54
15	288.15	59	1 704.4	0.010 69	0.816 00	0.829 98	32.97	59.97	220.22
20	293.15	68	2 337.2	0.014 75	0.830 17	0.849 83	38.00	75.42	273.32
25	298.15	77	3 167.0	0.020 16	0.844 34	0.871 62	43.03	94.38	337.39
30	303.15	86	4 242.8	0.027 31	0.858 51	0.896 09	48.07	117.8	415.65
35	308.15	95	5 623.4	0.036 73	0.872 74	0.924 06	53.10	147.3	512.17
40	313.15	104	7 377.6	0.049 11	0.886 92	0.956 65	58.14	184.5	532.31
45	318.15	113	9 584.8	0.065 36	0.901 15	0.995 35	63.17	232.0	783.06
50	323.15	122	12 339	0.086 78	0.915 32	1.042 3	68.21	293.1	975.27
55	328.15	131	15 745	0.115 2	0.929 49	1.100 7	73.25	372.9	1 221.5
60	333.15	140	19 925	0.153 4	0.943 72	1.174 8	78.29	478.5	1 543.5
65	338.15	149	25 014	0.205 5	0.957 90	1.272 1	83.33	621.4	1 973.6
70	343.15	158	31 167	0.278 8	0.972 07	1.404 2	88.38	820.5	2 564.8
75	348.15	167	38 554	0.385 8	0.986 30	1.592 4	93.42	1 110	3 412.8
80	353.15	176	47 365	0.551 9	1.000 5	1.879 1	98.47	1 557	4 710.9
85	358.15	185	57 809	0.836 3	1.014 6	2.363 2	103.5	2 321	6 892.6
90	363.15	194	70 112	1.416	1.028 8	3.340 9	108.6	3 876	11 281

Note: The P_s column in this table gives the vapor pressure of pure water at temperature intervals of five degrees Celsius. For the latest data on vapor pressures at intervals of 0.1 deg C, from 0–100 deg C, see "Vapor Pressure Equation for Water", A. Wexler and L. Greenspan, *J. Res. Nat. Bur. Stand.*, 75A(3):213–229, May–June 1971.

[a]For very low barometric pressures and high wet-bulb temperatures, the values of h_a in this table are somewhat low; for corrections see "ASHRAE Handbook of Fundamentals".

*Computed from: Psychrometric Tables, in "ASHRAE Handbook of Fundamentals", American Society of Heating, Refrigerating and Air-Conditioning Engineers, 1972.

TABLE A.4 Water Vapor at Low Pressures: Perfect Gas Behavior pv/T = R = 0.461 51 kJ/kg·K

Symbols and Units:

t = thermodynamic temperature, deg C
T = thermodynamic temperature, K
pv = RT, kJ/kg
u_o = specific internal energy at zero pressure, kJ/kg
h_o = specific enthalpy at zero pressure, kJ/kg
s_l = specific entropy of semiperfect vapor at 0.1 MN/m^2, kJ/kg·K
ψ_l = specific Helmholtz free energy of semiperfect vapor at 0.1 MN/m^2, kJ/kg
ψ_l = specific Helmholtz free energy of semiperfect vapor at 0.1 MN/m^2, kJ/kg
ζ_l = specific Gibbs free energy of semiperfect vapor at 0.1 MN/m^2, kJ/kg
p_r = relative pressure, pressure of semiperfect vapor at zero entropy, TN/m^2
v_r = relative specific volume, specific volume of semiperfect vapor at zero entropy, mm^3/kg
c_{po} = specific heat capacity at constant pressure for zero pressure, kJ/kg·K
c_{vo} = specific heat capacity at constant volume for zero pressure, kJ/kg·K
k = c_{po}/c_{vo} = isentropic exponent, $-(\partial \log p/\partial \log v)_s$

t	T	pv	u_o	h_o	s_l	ψ_l	ζ_l	p_r	v_r	c_{po}	c_{vo}	k
0	273.15	126.06	2 375.5	2 501.5	6.804 2	516.9	643.0	.252 9	498.4	1.858 4	1.396 9	1.330 4
10	283.15	130.68	2 389.4	2 520.1	6.871 1	443.9	574.6	.292 3	447.0	1.860 1	1.398 6	1.330 0
20	293.15	135.29	2 403.4	2 538.7	6.935 7	370.2	505.5	.336 3	402.4	1.862 2	1.400 7	1.329 5
30	303.15	139.91	2 417.5	2 557.4	6.998 2	296.0	435.9	.385 0	363.4	1.864 7	1.403 1	1.328 9
40	313.15	144.52	2 431.5	2 576.0	7.058 7	221.1	365.6	.439 0	329.2	1.867 4	1.405 9	1.328 3
50	323.15	149.14	2 445.6	2 594.7	7.117 5	145.6	294.7	.498 6	299.1	1.870 5	1.409 0	1.327 5
60	333.15	153.75	2 459.7	2 613.4	7.174 5	69.5	223.2	.564 2	272.5	1.873 8	1.412 3	1.326 8
70	343.15	158.37	2 473.8	2 632.2	7.230 0	−7.2	151.2	.636 3	248.9	1.877 4	1.415 9	1.325 9
80	353.15	162.98	2 488.0	2 651.0	7.284 0	−84.3	78.6	.715 2	227.9	1.881 2	1.419 7	1.325 1
90	363.15	167.60	2 502.2	2 669.8	7.336 6	−162.1	5.5	.801 5	209.1	1.885 2	1.423 7	1.324 2
100	373.15	172.21	2 516.5	2 688.7	7.387 8	−240.3	−68.1	.895 7	192.26	1.889 4	1.427 9	1.323 2
120	393.15	181.44	2 545.1	2 726.6	7.486 7	−398.3	−216.8	1.109 7	163.50	1.898 3	1.436 7	1.321 2
140	413.15	190.67	2 573.9	2 764.6	7.581 1	−558.2	−367.5	1.361 7	140.03	1.907 7	1.446 2	1.319 1
160	433.15	199.90	2 603.0	2 802.9	7.671 5	−720.0	−520.1	1.656 4	120.69	1.917 7	1.456 2	1.316 9
180	453.15	209.13	2 632.2	2 841.3	7.758 3	−883.5	−674.4	1.999 1	104.61	1.928 1	1.466 6	1.314 7
200	473.15	218.4	2 661.6	2 880.0	7.841 8	−1 048.7	−830.4	2.396	91.15	1.938 9	1.477 4	1.312 4
300	573.15	264.5	2 812.3	3 076.8	8.218 9	−1 898.4	−1 633.9	5.423	48.77	1.997 5	1.536 0	1.300 5
400	673.15	310.7	2 969.0	3 279.7	8.545 1	−2 783.1	−2 472.5	10.996	28.25	2.061 4	1.599 9	1.288 5
500	773.15	356.8	3 132.4	3 489.2	8.835 2	−3 699	−3 342	20.61	17.310	2.128 7	1.667 2	1.276 8
600	873.15	403.0	3 302.5	3 705.5	9.098 2	−4 642	−4 239	36.45	11.056	2.198 0	1.736 5	1.265 8
700	973.15	449.1	3 479.7	3 928.8	9.340 3	−5 610	−5 161	61.58	7.293	2.268 3	1.806 8	1.255 4
800	1 073.15	495.3	3 663.9	4 159.2	9.565 5	−6 601	−6 106	100.34	4.936	2.338 7	1.877 1	1.245 9
900	1 173.15	541.4	3 855.1	4 396.5	9.776 9	−7 615	−7 073	158.63	3.413	2.407 8	1.946 2	1.237 1
1 000	1 273.15	587.6	4 053.1	4 640.6	9.976 6	−8 649	−8 061	244.5	2.403	2.474 4	2.012 8	1.299 3
1 100	1 373.15	633.7	4 257.5	4 891.2	10.166 1	−9 702	−9 068	368.6	1.719	2.536 9	2.075 4	1.222 4
1 200	1 473.15	679.9	4 467.9	5 147.8	10.346 4	−10 774	−10 094	544.9	1.248	2.593 8	2.132 3	1.216 4
1 300	1 573.15	726.0	4 683.7	5 409.7	10.518 4	−11 863	−11 137	791.0	.918	2.643 1	2.181 6	1.211 5

*Adapted from: "Steam Tables", J.H. Keenan, F.G. Keyes, P.G. Hill, and J.G. Moore, John Wiley & Sons, Inc., 1969 (International Edition—Metric Units).

REFERENCE

For other steam tables in metric units, see "Steam Tables in SI Units", Ministry of Technology, London, 1970.

TABLE A.5 Properties of Saturated Water and Steam

Part a. Temperature Table

Temp. °C	Press. bars	Specific Volume m^3/kg		Internal Energy kJ/kg		Enthalpy kJ/kg			Entropy kJ/kg · K		Temp. °C
		Sat. Liquid $v_f \times 10^3$	Sat. Vapor v_g	Sat. Liquid u_f	Sat. Vapor u_g	Sat. Liquid h_f	Evap. h_{fg}	Sat. Vapor h_g	Sat. Liquid s_f	Sat. Vapor s_g	
.01	0.00611	1.0002	206.136	0.00	2375.3	0.01	2501.3	2501.4	0.0000	9.1562	.01
4	0.00813	1.0001	157.232	16.77	2380.9	16.78	2491.9	2508.7	0.0610	9.0514	4
5	0.00872	1.0001	147.120	20.97	2382.3	20.98	2489.6	2510.6	0.0761	9.0257	5
6	0.00935	1.0001	137.734	25.19	2383.6	25.20	2487.2	2512.4	0.0912	9.0003	6
8	0.01072	1.0002	120.917	33.59	2386.4	33.60	2482.5	2516.1	0.1212	8.9501	8
10	0.01228	1.0004	106.379	42.00	2389.2	42.01	2477.7	2519.8	0.1510	8.9008	10
11	0.01312	1.0004	99.857	46.20	2390.5	46.20	2475.4	2521.6	0.1658	8.8765	11
12	0.01402	1.0005	93.784	50.41	2391.9	50.41	2473.0	2523.4	0.1806	8.8524	12
13	0.01497	1.0007	88.124	54.60	2393.3	54.60	2470.7	2525.3	0.1953	8.8285	13
14	0.01598	1.0008	82.848	58.79	2394.7	58.80	2468.3	2527.1	0.2099	8.8048	14
15	0.01705	1.0009	77.926	62.99	2396.1	62.99	2465.9	2528.9	0.2245	8.7814	15
16	0.01818	1.0011	73.333	67.18	2397.4	67.19	2463.6	2530.8	0.2390	8.7582	16
17	0.01938	1.0012	69.044	71.38	2398.8	71.38	2461.2	2532.6	0.2535	8.7351	17
18	0.02064	1.0014	65.038	75.57	2400.2	75.58	2458.8	2534.4	0.2679	8.7123	18
19	0.02198	1.0016	61.293	79.76	2401.6	79.77	2456.5	2536.2	0.2823	8.6897	19
20	0.02339	1.0018	57.791	83.95	2402.9	83.96	2454.1	2538.1	0.2966	8.6672	20
21	0.02487	1.0020	54.514	88.14	2404.3	88.14	2451.8	2539.9	0.3109	8.6450	21
22	0.02645	1.0022	51.447	92.32	2405.7	92.33	2449.4	2541.7	0.3251	8.6229	22
23	0.02810	1.0024	48.574	96.51	2407.0	96.52	2447.0	2543.5	0.3393	8.6011	23
24	0.02985	1.0027	45.883	100.70	2408.4	100.70	2444.7	2545.4	0.3534	8.5794	24
25	0.03169	1.0029	43.360	104.88	2409.8	104.89	2442.3	2547.2	0.3674	8.5580	25
26	0.03363	1.0032	40.994	109.06	2411.1	109.07	2439.9	2549.0	0.3814	8.5367	26
27	0.03567	1.0035	38.774	113.25	2412.5	113.25	2437.6	2550.8	0.3954	8.5156	27
28	0.03782	1.0037	36.690	117.42	2413.9	117.43	2435.2	2552.6	0.4093	8.4946	28
29	0.04008	1.0040	34.733	121.60	2415.2	121.61	2432.8	2554.5	0.4231	8.4739	29
30	0.04246	1.0043	32.894	125.78	2416.6	125.79	2430.5	2556.3	0.4369	8.4533	30
31	0.04496	1.0046	31.165	129.96	2418.0	129.97	2428.1	2558.1	0.4507	8.4329	31
32	0.04759	1.0050	29.540	134.14	2419.3	134.15	2425.7	2559.9	0.4644	8.4127	32
33	0.05034	1.0053	28.011	138.32	2420.7	138.33	2423.4	2561.7	0.4781	8.3927	33
34	0.05324	1.0056	26.571	142.50	2422.0	142.50	2421.0	2563.5	0.4917	8.3728	34
35	0.05628	1.0060	25.216	146.67	2423.4	146.68	2418.6	2565.3	0.5053	8.3531	35
36	0.05947	1.0063	23.940	150.85	2424.7	150.86	2416.2	2567.1	0.5188	8.3336	36
38	0.06632	1.0071	21.602	159.20	2427.4	159.21	2411.5	2570.7	0.5458	8.2950	38
40	0.07384	1.0078	19.523	167.56	2430.1	167.57	2406.7	2574.3	0.5725	8.2570	40
45	0.09593	1.0099	15.258	188.44	2436.8	188.45	2394.8	2583.2	0.6387	8.1648	45

TABLE A.5 (continued) Properties of Saturated Water and Steam

Temp. °C	Press. bars	Specific Volume m³/kg		Internal Energy kJ/kg		Enthalpy kJ/kg			Entropy kJ/kg · K		Temp. °C
		Sat. Liquid $v_f \times 10^3$	Sat. Vapor v_g	Sat. Liquid u_f	Sat. Vapor u_g	Sat. Liquid h_f	Evap. h_{fg}	Sat. Vapor h_g	Sat. Liquid s_f	Sat. Vapor s_g	
50	.1235	1.0121	12.032	209.32	2443.5	209.33	2382.7	2592.1	.7038	8.0763	50
55	.1576	1.0146	9.568	230.21	2450.1	230.23	2370.7	2600.9	.7679	7.9913	55
60	.1994	1.0172	7.671	251.11	2456.6	251.13	2358.5	2609.6	.8312	7.9096	60
65	.2503	1.0199	6.197	272.02	2463.1	272.06	2346.2	2618.3	.8935	7.8310	65
70	.3119	1.0228	5.042	292.95	2469.6	292.98	2333.8	2626.8	.9549	7.7553	70
75	.3858	1.0259	4.131	313.90	2475.9	313.93	2321.4	2635.3	1.0155	7.6824	75
80	.4739	1.0291	3.407	334.86	2482.2	334.91	2308.8	2643.7	1.0753	7.6122	80
85	.5783	1.0325	2.828	355.84	2488.4	355.90	2296.0	2651.9	1.1343	7.5445	85
90	.7014	1.0360	2.361	376.85	2494.5	376.92	2283.2	2660.1	1.1925	7.4791	90
95	.8455	1.0397	1.982	397.88	2500.6	397.96	2270.2	2668.1	1.2500	7.4159	95
100	1.014	1.0435	1.673	418.94	2506.5	419.04	2257.0	2676.1	1.3069	7.3549	100
110	1.433	1.0516	1.210	461.14	2518.1	461.30	2230.2	2691.5	1.4185	7.2387	110
120	1.985	1.0603	0.8919	503.50	2529.3	503.71	2202.6	2706.3	1.5276	7.1296	120
130	2.701	1.0697	0.6685	546.02	2539.9	546.31	2174.2	2720.5	1.6344	7.0269	130
140	3.613	1.0797	0.5089	588.74	2550.0	589.13	2144.7	2733.9	1.7391	6.9299	140
150	4.758	1.0905	0.3928	631.68	2559.5	632.20	2114.3	2746.5	1.8418	6.8379	150
160	6.178	1.1020	0.3071	674.86	2568.4	675.55	2082.6	2758.1	1.9427	6.7502	160
170	7.917	1.1143	0.2428	718.33	2576.5	719.21	2049.5	2768.7	2.0419	6.6663	170
180	10.02	1.1274	0.1941	762.09	2583.7	763.22	2015.0	2778.2	2.1396	6.5857	180
190	12.54	1.1414	0.1565	806.19	2590.0	807.62	1978.8	2786.4	2.2359	6.5079	190
200	15.54	1.1565	0.1274	850.65	2595.3	852.45	1940.7	2793.2	2.3309	6.4323	200
210	19.06	1.1726	0.1044	895.53	2599.5	897.76	1900.7	2798.5	2.4248	6.3585	210
220	23.18	1.1900	0.08619	940.87	2602.4	943.62	1858.5	2802.1	2.5178	6.2861	220
230	27.95	1.2088	0.07158	986.74	2603.9	990.12	1813.8	2804.0	2.6099	6.2146	230
240	33.44	1.2291	0.05976	1033.2	2604.0	1037.3	1766.5	2803.8	2.7015	6.1437	240
250	39.73	1.2512	0.05013	1080.4	2602.4	1085.4	1716.2	2801.5	2.7927	6.0730	250
260	46.88	1.2755	0.04221	1128.4	2599.0	1134.4	1662.5	2796.6	2.8838	6.0019	260
270	54.99	1.3023	0.03564	1177.4	2593.7	1184.5	1605.2	2789.7	2.9751	5.9301	270
280	64.12	1.3321	0.03017	1227.5	2586.1	1236.0	1543.6	2779.6	3.0668	5.8571	280
290	74.36	1.3656	0.02557	1278.9	2576.0	1289.1	1477.1	2766.2	3.1594	5.7821	290
300	85.81	1.4036	0.02167	1332.0	2563.0	1344.0	1404.9	2749.0	3.2534	5.7045	300
320	112.7	1.4988	0.01549	1444.6	2525.5	1461.5	1238.6	2700.1	3.4480	5.5362	320
340	145.9	1.6379	0.01080	1570.3	2464.6	1594.2	1027.9	2622.0	3.6594	5.3357	340
360	186.5	1.8925	0.006945	1725.2	2351.5	1760.5	720.5	2481.0	3.9147	5.0526	360
374.14	220.9	3.155	0.003155	2029.6	2029.6	2099.3	0	2099.3	4.4298	4.4298	374.14

TABLE A.5 (continued) Properties of Saturated Water and Steam

Part b. Pressure Table

Press. bars	Temp. °C	Specific Volume m^3/kg		Internal Energy kJ/kg		Enthalpy kJ/kg			Entropy kJ/kg · K		Press. bars
		Sat. Liquid $v_f \times 10^3$	Sat. Vapor v_g	Sat. Liquid u_f	Sat. Vapor u_g	Sat. Liquid h_f	Evap. h_{fg}	Sat. Vapor h_g	Sat. Liquid s_f	Sat. Vapor s_g	
0.04	28.96	1.0040	34.800	121.45	2415.2	121.46	2432.9	2554.4	0.4226	8.4746	0.04
0.06	36.16	1.0064	23.739	151.53	2425.0	151.53	2415.9	2567.4	0.5210	8.3304	0.06
0.08	41.51	1.0084	18.103	173.87	2432.2	173.88	2403.1	2577.0	0.5926	8.2287	0.08
0.10	45.81	1.0102	14.674	191.82	2437.9	191.83	2392.8	2584.7	0.6493	8.1502	0.10
0.20	60.06	1.0172	7.649	251.38	2456.7	251.40	2358.3	2609.7	0.8320	7.9085	0.20
0.30	69.10	1.0223	5.229	289.20	2468.4	289.23	2336.1	2625.3	0.9439	7.7686	0.30
0.40	75.87	1.0265	3.993	317.53	2477.0	317.58	2319.2	2636.8	1.0259	7.6700	0.40
0.50	81.33	1.0300	3.240	340.44	2483.9	340.49	2305.4	2645.9	1.0910	7.5939	0.50
0.60	85.94	1.0331	2.732	359.79	2489.6	359.86	2293.6	2653.5	1.1453	7.5320	0.60
0.70	89.95	1.0360	2.365	376.63	2494.5	376.70	2283.3	2660.0	1.1919	7.4797	0.70
0.80	93.50	1.0380	2.087	391.58	2498.8	391.66	2274.1	2665.8	1.2329	7.4346	0.80
0.90	96.71	1.0410	1.869	405.06	2502.6	405.15	2265.7	2670.9	1.2695	7.3949	0.90
1.00	99.63	1.0432	1.694	417.36	2506.1	417.46	2258.0	2675.5	1.3026	7.3594	1.00
1.50	111.4	1.0528	1.159	466.94	2519.7	467.11	2226.5	2693.6	1.4336	7.2233	1.50
2.00	120.2	1.0605	0.8857	504.49	2529.5	504.70	2201.9	2706.7	1.5301	7.1271	2.00
2.50	127.4	1.0672	0.7187	535.10	2537.2	535.37	2181.5	2716.9	1.6072	7.0527	2.50
3.00	133.6	1.0732	0.6058	561.15	2543.6	561.47	2163.8	2725.3	1.6718	6.9919	3.00
3.50	138.9	1.0786	0.5243	583.95	2546.9	584.33	2148.1	2732.4	1.7275	6.9405	3.50
4.00	143.6	1.0836	0.4625	604.31	2553.6	604.74	2133.8	2738.6	1.7766	6.8959	4.00
4.50	147.9	1.0882	0.4140	622.25	2557.6	623.25	2120.7	2743.9	1.8207	6.8565	4.50
5.00	151.9	1.0926	0.3749	639.68	2561.2	640.23	2108.5	2748.7	1.8607	6.8212	5.00
6.00	158.9	1.1006	0.3157	669.90	2567.4	670.56	2086.3	2756.8	1.9312	6.7600	6.00
7.00	165.0	1.1080	0.2729	696.44	2572.5	697.22	2066.3	2763.5	1.9922	6.7080	7.00
8.00	170.4	1.1148	0.2404	720.22	2576.8	721.11	2048.0	2769.1	2.0462	6.6628	8.00
9.00	175.4	1.1212	0.2150	741.83	2580.5	742.83	2031.1	2773.9	2.0946	6.6226	9.00
10.0	179.9	1.1273	0.1944	761.68	2583.6	762.81	2015.3	2778.1	2.1387	6.5863	10.0
15.0	198.3	1.1539	0.1318	843.16	2594.5	844.84	1947.3	2792.2	2.3150	6.4448	15.0
20.0	212.4	1.1767	0.09963	906.44	2600.3	908.79	1890.7	2799.5	2.4474	6.3409	20.0
25.0	224.0	1.1973	0.07998	959.11	2603.1	962.11	1841.0	2803.1	2.5547	6.2575	25.0
30.0	233.9	1.2165	0.06668	1004.8	2604.1	1008.4	1795.7	2804.2	2.6457	6.1869	30.0
35.0	242.6	1.2347	0.05707	1045.4	2603.7	1049.8	1753.7	2803.4	2.7253	6.1253	35.0
40.0	250.4	1.2522	0.04978	1082.3	2602.3	1087.3	1714.1	2801.4	2.7964	6.0701	40.0
45.0	257.5	1.2692	0.04406	1116.2	2600.1	1121.9	1676.4	2798.3	2.8610	6.0199	45.0
50.0	264.0	1.2859	0.03944	1147.8	2597.1	1154.2	1640.1	2794.3	2.9202	5.9734	50.0
60.0	275.6	1.3187	0.03244	1205.4	2589.7	1213.4	1571.0	2784.3	3.0267	5.8892	60.0
70.0	285.9	1.3513	0.02737	1257.6	2580.5	1267.0	1505.1	2772.1	3.1211	5.8133	70.0
80.0	295.1	1.3842	0.02352	1305.6	2569.8	1316.6	1441.3	2758.0	3.2068	5.7432	80.0
90.0	303.4	1.4178	0.02048	1350.5	2557.8	1363.3	1378.9	2742.1	3.2858	5.6772	90.0
100.	311.1	1.4524	0.01803	1393.0	2544.4	1407.6	1317.1	2724.7	3.3596	5.6141	100.
110.	318.2	1.4886	0.01599	1433.7	2529.8	1450.1	1255.5	2705.6	3.4295	5.5527	110.
120.	324.8	1.5267	0.01426	1473.0	2513.7	1491.3	1193.6	2684.9	3.4962	5.4924	120.
130.	330.9	1.5671	0.01278	1511.1	2496.1	1531.5	1130.7	2662.2	3.5606	5.4323	130.
140.	336.8	1.6107	0.01149	1548.6	2476.8	1571.1	1066.5	2637.6	3.6232	5.3717	140.
150.	342.2	1.6581	0.01034	1585.6	2455.5	1610.5	1000.0	2610.5	3.6848	5.3098	150.
160.	347.4	1.7107	0.009306	1622.7	2431.7	1650.1	930.6	2580.6	3.7461	5.2455	160.
170.	352.4	1.7702	0.008364	1660.2	2405.0	1690.3	856.9	2547.2	3.8079	5.1777	170.
180.	357.1	1.8397	0.007489	1698.9	2374.3	1732.0	777.1	2509.1	3.8715	5.1044	180.
190.	361.5	1.9243	0.006657	1739.9	2338.1	1776.5	688.0	2464.5	3.9388	5.0228	190.
200.	365.8	2.036	0.005834	1785.6	2293.0	1826.3	583.4	2409.7	4.0139	4.9269	200.
220.9	374.1	3.155	0.003155	2029.6	2029.6	2099.3	0	2099.3	4.4298	4.4298	220.9

Source: Adapted from M.J. Moran and H.N. Shapiro, *Fundamentals of Engineering Thermodynamics*, 3rd. ed., Wiley, New York, 1995, as extracted from J.H. Keenan, F.G. Keyes, P.G. Hill, and J.G. Moore, *Steam Tables*, Wiley, New York, 1969.

TABLE A.6 Properties of Superheated Steam

Symbols and Units:

T = temperature, °C
T_{sat} = Saturation temperature, °C
ν = Specific volume, m^3/kg
u = internal energy, kJ/kg
h = enthalpy, kJ/kg
S = entropy, kJ/kg·K
p = pressure, bar and μPa

T °C	v m^3/kg	u kJ/kg	h kJ/kg	s kJ/kg · K	v m^3/kg	u kJ/kg	h kJ/kg	s kJ/kg · K
	p = 0.06 bar = 0.006 MPa (T_{sat} = 36.16°C)				p = 0.35 bar = 0.035 MPa (T_{sat} = 72.69°C)			
Sat.	23.739	2425.0	2567.4	8.3304	4.526	2473.0	2631.4	7.7158
80	27.132	2487.3	2650.1	8.5804	4.625	2483.7	2645.6	7.7564
120	30.219	2544.7	2726.0	8.7840	5.163	2542.4	2723.1	7.9644
160	33.302	2602.7	2802.5	8.9693	5.696	2601.2	2800.6	8.1519
200	36.383	2661.4	2879.7	9.1398	6.228	2660.4	2878.4	8.3237
240	39.462	2721.0	2957.8	9.2982	6.758	2720.3	2956.8	8.4828
280	42.540	2781.5	3036.8	9.4464	7.287	2780.9	3036.0	8.6314
320	45.618	2843.0	3116.7	9.5859	7.815	2842.5	3116.1	8.7712
360	48.696	2905.5	3197.7	9.7180	8.344	2905.1	3197.1	8.9034
400	51.774	2969.0	3279.6	9.8435	8.872	2968.6	3279.2	9.0291
440	54.851	3033.5	3362.6	9.9633	9.400	3033.2	3362.2	9.1490
500	59.467	3132.3	3489.1	10.1336	10.192	3132.1	3488.8	9.3194
	p = 0.70 bar = 0.07 MPa (T_{sat} = 89.95°C)				p = 1.0 bar = 0.10 MPa (T_{sat} = 99.63°C)			
Sat.	2.365	2494.5	2660.0	7.4797	1.694	2506.1	2675.5	7.3594
100	2.434	2509.7	2680.0	7.5341	1.696	2506.7	2676.2	7.3614
120	2.571	2539.7	2719.6	7.6375	1.793	2537.3	2716.6	7.4668
160	2.841	2599.4	2798.2	7.8279	1.984	2597.8	2796.2	7.6597
200	3.108	2659.1	2876.7	8.0012	2.172	2658.1	2875.3	7.8343
240	3.374	2719.3	2955.5	8.1611	2.359	2718.5	2954.5	7.9949
280	3.640	2780.2	3035.0	8.3162	2.546	2779.6	3034.2	8.1445
320	3.905	2842.0	3115.3	8.4504	2.732	2841.5	3114.6	8.2849
360	4.170	2904.6	3196.5	8.5828	2.917	2904.2	3195.9	8.4175
400	4.434	2968.2	3278.6	8.7086	3.103	2967.9	3278.2	8.5435
440	4.698	3032.9	3361.8	8.8286	3.288	3032.6	3361.4	8.6636
500	5.095	3131.8	3488.5	8.9991	3.565	3131.6	3488.1	8.8342
	p = 1.5 bars = 0.15 MPa (T_{sat} = 111.37°C)				p = 3.0 bars = 0.30 MPa (T_{sat} = 133.55°C)			
Sat.	1.159	2519.7	2693.6	7.2233	0.606	2543.6	2725.3	6.9919
120	1.188	2533.3	2711.4	7.2693				
160	1.317	2595.2	2792.8	7.4665	0.651	2587.1	2782.3	7.1276
200	1.444	2656.2	2872.9	7.6433	0.716	2650.7	2865.5	7.3115
240	1.570	2717.2	2952.7	7.8052	0.781	2713.1	2947.3	7.4774
280	1.695	2778.6	3032.8	7.9555	0.844	2775.4	3028.6	7.6299
320	1.819	2840.6	3113.5	8.0964	0.907	2838.1	3110.1	7.7722
360	1.943	2903.5	3195.0	8.2293	0.969	2901.4	3192.2	7.9061
400	2.067	2967.3	3277.4	8.3555	1.032	2965.6	3275.0	8.0330
440	2.191	3032.1	3360.7	8.4757	1.094	3030.6	3358.7	8.1538
500	2.376	3131.2	3487.6	8.6466	1.187	3130.0	3486.0	8.3251
600	2.685	3301.7	3704.3	8.9101	1.341	3300.8	3703.2	8.5892

TABLE A.6 (continued) Properties of Superheated Steam

Symbols and Units:

T = temperature, °C
T_{sat} = Saturation temperature, °C
ν = Specific volume, m^3/kg
u = internal energy, kJ/kg
h = enthalpy, kJ/kg
S = entropy, kJ/kg·K
p = pressure, bar and μPa

T °C	v m^3/kg	u kJ/kg	h kJ/kg	s kJ/kg · K	v m^3/kg	u kJ/kg	h kJ/kg	s kJ/kg . k
	p = 5.0 bars = 0.50 MPa (T_{sat} = 151.86°C)				p = 7.0 bars = 0.70 MPa (T_{sat} = 164.97°C)			
Sat.	0.3749	2561.2	2748.7	6.8213	0.2729	2572.5	2763.5	6.7080
180	0.4045	2609.7	2812.0	6.9656	0.2847	2599.8	2799.1	6.7880
200	0.4249	2642.9	2855.4	7.0592	0.2999	2634.8	2844.8	6.8865
240	0.4646	2707.6	2939.9	7.2307	0.3292	2701.8	2932.2	7.0641
280	0.5034	2771.2	3022.9	7.3865	0.3574	2766.9	3017.1	7.2233
320	0.5416	2834.7	3105.6	7.5308	0.3852	2831.3	3100.9	7.3697
360	0.5796	2898.7	3188.4	7.6660	0.4126	2895.8	3184.7	7.5063
400	0.6173	2963.2	3271.9	7.7938	0.4397	2960.9	3268.7	7.6350
440	0.6548	3028.6	3356.0	7.9152	0.4667	3026.6	3353.3	7.7571
500	0.7109	3128.4	3483.9	8.0873	0.5070	3126.8	3481.7	7.9299
600	0.8041	3299.6	3701.7	8.3522	0.5738	3298.5	3700.2	8.1956
700	0.8969	3477.5	3925.9	8.5952	0.6403	3476.6	3924.8	8.4391
	p = 10.0 bars = 1.0 MPa (T_{sat} = 179.91°C)				p = 15.0 bars = 1.5 MPa (T_{sat} = 198.32°C)			
Sat.	0.1944	2583.6	2778.1	6.5865	0.1318	2594.5	2792.2	6.4448
200	0.2060	2621.9	2827.9	6.6940	0.1325	2598.1	2796.8	6.4546
240	0.2275	2692.9	2920.4	6.8817	0.1483	2676.9	2899.3	6.6628
280	0.2480	2760.2	3008.2	7.0465	0.1627	2748.6	2992.7	6.8381
320	0.2678	2826.1	3093.9	7.1962	0.1765	2817.1	3081.9	6.9938
360	0.2873	2891.6	3178.9	7.3349	0.1899	2884.4	3169.2	7.1363
400	0.3066	2957.3	3263.9	7.4651	0.2030	2951.3	3255.8	7.2690
440	0.3257	3023.6	3349.3	7.5883	0.2160	3018.5	3342.5	7.3940
500	0.3541	3124.4	3478.5	7.7622	0.2352	3120.3	3473.1	7.5698
540	0.3729	3192.6	3565.6	7.8720	0.2478	3189.1	3560.9	7.6805
600	0.4011	3296.8	3697.9	8.0290	0.2668	3293.9	3694.0	7.8385
640	0.4198	3367.4	3787.2	8.1290	0.2793	3364.8	3783.8	7.9391
	p = 20.0 bars = 2.0 MPa (T_{sat} = 212.42°C)				p = 30.0 bars = 3.0 MPa (T_{sat} = 233.90°C)			
Sat.	0.0996	2600.3	2799.5	6.3409	0.0667	2604.1	2804.2	6.1869
240	0.1085	2659.6	2876.5	6.4952	0.0682	2619.7	2824.3	6.2265
280	0.1200	2736.4	2976.4	6.6828	0.0771	2709.9	2941.3	6.4462
320	0.1308	2807.9	3069.5	6.8452	0.0850	2788.4	3043.4	6.6245
360	0.1411	2877.0	3159.3	6.9917	0.0923	2861.7	3138.7	6.7801
400	0.1512	2945.2	3247.6	7.1271	0.0994	2932.8	3230.9	6.9212
440	0.1611	3013.4	3335.5	7.2540	0.1062	3002.9	3321.5	7.0520
500	0.1757	3116.2	3467.6	7.4317	0.1162	3108.0	3456.5	7.2338
540	0.1853	3185.6	3556.1	7.5434	0.1227	3178.4	3546.6	7.3474
600	0.1996	3290.9	3690.1	7.7024	0.1324	3285.0	3682.3	7.5085
640	0.2091	3362.2	3780.4	7.8035	0.1388	3357.0	3773.5	7.6106
700	0.2232	3470.9	3917.4	7.9487	0.1484	3466.5	3911.7	7.7571

TABLE A.6 (continued) Properties of Superheated Steam

Symbols and Units:

T = temperature, °C
T_{sat} = Saturation temperature, °C
ν = Specific volume, m^3/kg
u = internal energy, kJ/kg
h = enthalpy, kJ/kg
S = entropy, kJ/kg·K
p = pressure, bar and μPa

T °C	v m^3/kg	u kJ/kg	h kJ/kg	s kJ/kg · K	v m^3/kg	u kJ/kg	h kJ/kg	s kJ/kg · K
	p = 40 bars = 4.0 MPa (T_{sat} = 250.4°C)				p = 60 bars = 6.0 MPa (T_{sat} = 275.64°C)			
Sat.	0.04978	2602.3	2801.4	6.0701	0.03244	2589.7	2784.3	5.8892
280	0.05546	2680.0	2901.8	6.2568	0.03317	2605.2	2804.2	5.9252
320	0.06199	2767.4	3015.4	6.4553	0.03876	2720.0	2952.6	6.1846
360	0.06788	2845.7	3117.2	6.6215	0.04331	2811.2	3071.1	6.3782
400	0.07341	2919.9	3213.6	6.7690	0.04739	2892.9	3177.2	6.5408
440	0.07872	2992.2	3307.1	6.9041	0.05122	2970.0	3277.3	6.6853
500	0.08643	3099.5	3445.3	7.0901	0.05665	3082.2	3422.2	6.8803
540	0.09145	3171.1	3536.9	7.2056	0.06015	3156.1	3517.0	6.9999
600	0.09885	3279.1	3674.4	7.3688	0.06525	3266.9	3658.4	7.1677
640	0.1037	3351.8	3766.6	7.4720	0.06859	3341.0	3752.6	7.2731
700	0.1110	3462.1	3905.9	7.6198	0.07352	3453.1	3894.1	7.4234
740	0.1157	3536.6	3999.6	7.7141	0.07677	3528.3	3989.2	7.5190
	p = 80 bars = 8.0 MPa (T_{sat} = 295.06°C)				p = 100 bars = 10.0 MPa (T_{sat} = 311.06°C)			
Sat.	0.02352	2569.8	2758.0	5.7432	0.01803	2544.4	2724.7	5.6141
320	0.02682	2662.7	2877.2	5.9489	0.01925	2588.8	2781.3	5.7103
360	0.03089	2772.7	3019.8	6.1819	0.02331	2729.1	2962.1	6.0060
400	0.03432	2863.8	3138.3	6.3634	0.02641	2832.4	3096.5	6.2120
440	0.03742	2946.7	3246.1	6.5190	0.02911	2922.1	3213.2	6.3805
480	0.04034	3025.7	3348.4	6.6586	0.03160	3005.4	3321.4	6.5282
520	0.04313	3102.7	3447.7	6.7871	0.03394	3085.6	3425.1	6.6622
560	0.04582	3178.7	3545.3	6.9072	0.03619	3164.1	3526.0	6.7864
600	0.04845	3254.4	3642.0	7.0206	0.03837	3241.7	3625.3	6.9029
640	0.05102	3330.1	3738.3	7.1283	0.04048	3318.9	3723.7	7.0131
700	0.05481	3443.9	3882.4	7.2812	0.04358	3434.7	3870.5	7.1687
740	0.05729	3520.4	3978.7	7.3782	0.04560	3512.1	3968.1	7.2670
	p = 120 bars = 12.0 MPa (T_{sat} = 324.75°C)				p = 140 bars = 14.0 MPa (T_{sat} = 336.75°C)			
Sat.	0.01426	2513.7	2684.9	5.4924	0.01149	2476.8	2637.6	5.3717
360	0.01811	2678.4	2895.7	5.8361	0.01422	2617.4	2816.5	5.6602
400	0.02108	2798.3	3051.3	6.0747	0.01722	2760.9	3001.9	5.9448
440	0.02355	2896.1	3178.7	6.2586	0.01954	2868.6	3142.2	6.1474
480	0.02576	2984.4	3293.5	6.4154	0.02157	2962.5	3264.5	6.3143
520	0.02781	3068.0	3401.8	6.5555	0.02343	3049.8	3377.8	6.4610
560	0.02977	3149.0	3506.2	6.6840	0.02517	3133.6	3486.0	6.5941
600	0.03164	3228.7	3608.3	6.8037	0.02683	3215.4	3591.1	6.7172
640	0.03345	3307.5	3709.0	6.9164	0.02843	3296.0	3694.1	6.8326
700	0.03610	3425.2	3858.4	7.0749	0.03075	3415.7	3846.2	6.9939
740	0.03781	3503.7	3957.4	7.1746	0.03225	3495.2	3946.7	7.0952

TABLE A.7 Chemical, Physical, and Thermal Properties of Gases: Gases and Vapors, Including Fuels and Refrigerants, English and Metric Units

Common name(s)	*Acetylene (Ethyne)*	*Air [mixture]*	*Ammonia, anhyd.*	*Argon*
Chemical formula	C_2H_2		NH_3	*Ar*
Refrigerant number	—	*729*	*717*	*740*
CHEMICAL AND PHYSICAL PROPERTIES				
Molecular weight	26.04	28.966	17.02	39.948
Specific gravity, air = 1	0.90	1.00	0.59	1.38
Specific volume, ft^3/lb	14.9	13.5	23.0	9.80
Specific volume, m^3/kg	0.93	0.842	1.43	0.622
Density of liquid (at atm bp), lb/ft^3	43.0	54.6	42.6	87.0
Density of liquid (at atm bp), kg/m^3	693.	879.	686.	1 400.
Vapor pressure at 25 deg C, psia			145.4	
Vapor pressure at 25 deg C, MN/m^2			1.00	
Viscosity (abs), lbm/ft·sec	6.72×10^{-6}	12.1×10^{-6}	6.72×10^{-6}	13.4×10^{-6}
Viscosity (abs), centipoises[a]	0.01	0.018	0.010	0.02
Sound velocity in gas, m/sec	343	346	415	322
THERMAL AND THERMODYNAMIC PROPERTIES				
Specific heat, c_p, Btu/lb·deg F or cal/g·deg C	0.40	0.240 3	0.52	0.125
Specific heat, c_p, J/kg·K	1 674.	1 005.	2 175.	523.
Specific heat ratio, c_p/c_v	1.25	1.40	1.3	1.67
Gas constant R, ft-lb/lb·deg R	59.3	53.3	90.8	38.7
Gas constant R, J/kg·deg C	319	286.8	488.	208.
Thermal conductivity, Btu/hr·ft·deg F	0.014	0.015 1	0.015	0.010 2
Thermal conductivity, W/m·deg C	0.024	0.026	0.026	0.017 2
Boiling point (sat 14.7 psia), deg F	−103	−320	−28.	−303.
Boiling point (sat 760 mm), deg C	−75	−195	−33.3	−186
Latent heat of evap (at bp), Btu/lb	264	88.2	589.3	70.
Latent heat of evap (at bp), J/kg	614 000	205 000.	1 373 000	163 000
Freezing (melting) point, deg F (1 atm)	−116	−357.2	−107.9	−308.5
Freezing (melting) point, deg C (1 atm)	−82.2	−216.2	−77.7	−189.2
Latent heat of fusion, Btu/lb	23.	10.0	143.0	
Latent heat of fusion, J/kg	53 500	23 200	332 300	
Critical temperature, deg F	97.1	−220.5	271.4	−187.6
Critical temperature, deg C	36.2	−140.3	132.5	−122
Critical pressure, psia	907.	550.	1 650.	707.
Critical pressure, MN/m^2	6.25	3.8	11.4	4.87
Critical volume, ft^3/lb		0.050	0.068	0.029 9
Critical volume, m^3/kg		0.003	0.004 24	0.001 86
Flammable (yes or no)	Yes	No	No	No
Heat of combustion, Btu/ft^3	1 450	—	—	—
Heat of combustion, Btu/lb	21 600	—	—	—
Heat of combustion, kJ/kg	50 200	—	—	—

[a]For $N \cdot sec/m^2$ divide by 1 000.

Note: The properties of pure gases are given at 25°C (77°F, 298 K) and atmospheric pressure (except as stated).

TABLE A.7 (continued) Chemical, Physical, and Thermal Properties of Gases: Gases and Vapors, Including Fuels and Refrigerants, English and Metric Units

Common name(s)	*Butadiene*	*n-Butane*	*Isobutane (2-Methyl propane)*	*1-Butene (Butylene)*
Chemical formula	C_4H_6	C_4H_{10}	C_4H_{10}	C_4H_8
Refrigerant number	—	*600*	*600a*	—
CHEMICAL AND PHYSICAL PROPERTIES				
Molecular weight	54.09	58.12	58.12	56.108
Specific gravity, air = 1	1.87	2.07	2.07	1.94
Specific volume, ft^3/lb	7.1	6.5	6.5	6.7
Specific volume, m^3/kg	0.44	0.405	0.418	0.42
Density of liquid (at atm bp), lb/ft^3		37.5	37.2	
Density of liquid (at atm bp), kg/m^3		604.	599.	
Vapor pressure at 25 deg C, psia		35.4	50.4	
Vapor pressure at 25 deg C, MN/m^2		0.024 4	0.347	
Viscosity (abs), lbm/ft·sec		4.8×10^{-6}		
Viscosity (abs), centipoises[a]		0.007		
Sound velocity in gas, m/sec	226	216	216	222
THERMAL AND THERMO-DYNAMIC PROPERTIES				
Specific heat, c_p, Btu/lb·deg F or cal/g·deg C	0.341	0.39	0.39	0.36
Specific heat, c_p, J/kg·K	1 427.	1 675.	1 630.	1 505.
Specific heat ratio, c_p/c_v	1.12	1.096	1.10	1.112
Gas constant R, ft-lb/lb·deg F	28.55	26.56	26.56	27.52
Gas constant R, J/kg·deg C	154.	143.	143.	148.
Thermal conductivity, Btu/hr·ft·deg F		0.01	0.01	
Thermal conductivity, W/m·deg C		0.017	0.017	
Boiling point (sat 14.7 psia), deg F	24.1	31.2	10.8	20.6
Boiling point (sat 760 mm), deg C	−4.5	−0.4	−11.8	−6.3
Latent heat of evap (at bp), Btu/lb		165.6	157.5	167.9
Latent heat of evap (at bp), J/kg		386 000	366 000	391 000
Freezing (melting) point, deg F (1 atm)	−164.	−217.	−229	−301.6
Freezing (melting) point, deg C (1 atm)	−109.	−138	−145	−185.3
Latent heat of fusion, Btu/lb		19.2		16.4
Latent heat of fusion, J/kg		44 700		38 100
Critical temperature, deg F		306	273.	291.
Critical temperature, deg C	171.	152.	134.	144.
Critical pressure, psia	652.	550.	537.	621.
Critical pressure, MN/m^2		3.8	3.7	4.28
Critical volume, ft^3/lb		0.070		0.068
Critical volume, m^3/kg		0.004 3		0.004 2
Flammable (yes or no)	Yes	Yes	Yes	Yes
Heat of combustion, Btu/ft^3	2 950	3 300	3 300	3 150
Heat of combustion, Btu/lb	20 900	21 400	21 400	21 000
Heat of combustion, kJ/kg	48 600	49 700	49 700	48 800

[a]For $N \cdot sec/m^2$ divide by 1 000.

TABLE A.7 (continued) Chemical, Physical, and Thermal Properties of Gases: Gases and Vapors, Including Fuels and Refrigerants, English and Metric Units

Common name(s) *Chemical formula* *Refrigerant number*	*cis-2-Butene* C_4H_8 –	*trans-2-Butene* C_4H_8 —	*Isobutene* C_4H_8 —	*Carbon dioxide* CO_2 *744*
CHEMICAL AND PHYSICAL PROPERTIES				
Molecular weight	56.108	56.108	56.108	44.01
Specific gravity, air = 1	1.94	1.94	1.94	1.52
Specific volume, ft^3/lb	6.7	6.7	6.7	8.8
Specific volume, m^3/kg	0.42	0.42	0.42	0.55
Density of liquid (at atm bp), lb/ft^3				—
Density of liquid (at atm bp), kg/m^3				—
Vapor pressure at 25 deg C, psia				931.
Vapor pressure at 25 deg C, MN/m^2				6.42
Viscosity (abs), lbm/ft·sec				9.4×10^{-6}
Viscosity (abs), centipoises[a]				0.014
Sound velocity in gas, m/sec	223.	221.	221.	270.
THERMAL AND THERMODYNAMIC PROPERTIES				
Specific heat, c_p, Btu/lb·deg F or cal/g·deg C	0.327	0.365	0.37	0.205
Specific heat, c_p, J/kg·K	1 368.	1 527.	1 548.	876.
Specific heat ratio, c_p/c_v	1.121	1.107	1.10	1.30
Gas constant R, ft-lb/lb·deg F				35.1
Gas constant R, J/kg·deg C				189.
Thermal conductivity, Btu/hr·ft·deg F				0.01
Thermal conductivity, W/m·deg C				0.017
Boiling point (sat 14.7 psia), deg F	38.6	33.6	19.2	−109.4[b]
Boiling point (sat 760 mm), deg C	3.7	0.9	−7.1	−78.5
Latent heat of evap (at bp), Btu/lb	178.9	174.4	169.	246.
Latent heat of evap (at bp), J/kg	416 000.	406 000.	393 000.	572 000.
Freezing (melting) point, deg F (1 atm)	−218.	−158.		
Freezing (melting) point, deg C (1 atm)	−138.9	−105.5		
Latent heat of fusion, Btu/lb	31.2	41.6	25.3	—
Latent heat of fusion, J/kg	72 600.	96 800.	58 800.	—
Critical temperature, deg F				88.
Critical temperature, deg C	160.	155.		31.
Critical pressure, psia	595.	610.		1 072.
Critical pressure, MN/m^2	4.10	4.20		7.4
Critical volume, ft^3/lb				
Critical volume, m^3/kg				
Flammable (yes or no)	Yes	Yes	Yes	No
Heat of combustion, Btu/ft^3	3 150.	3 150.	3 150.	—
Heat of combustion, Btu/lb	21 000.	21 000.	21 000.	—
Heat of combustion, kJ/kg	48 800.	48 800.	48 800.	—

[a]For N·sec/m^2 divide by 1 000.
[b]Sublimes.

TABLE A.7 (continued) Chemical, Physical, and Thermal Properties of Gases: Gases and Vapors, Including Fuels and Refrigerants, English and Metric Units

Common name(s)	*Carbon monoxide*	*Chlorine*	*Deuterium*	*Ethane*
Chemical formula	CO	Cl_2	D_2	C_2H_6
Refrigerant number	—	—	—	*170*
CHEMICAL AND PHYSICAL PROPERTIES				
Molecular weight	28.011	70.906	2.014	30.070
Specific gravity, air = 1	0.967	2.45	0.070	1.04
Specific volume, ft^3/lb	14.0	5.52	194.5	13.025
Specific volume, m^3/kg	0.874	0.344	12.12	0.815
Density of liquid (at atm bp), lb/ft^3		97.3		28.
Density of liquid (at atm bp), kg/m^3		1 559.		449.
Vapor pressure at 25 deg C, psia			0.756	
Vapor pressure at 25 deg C, MN/m^2			0.005 2	
Viscosity (abs), lbm/ft·sec	12.1×10^{-6}	9.4×10^{-6}	8.75×10^{-6}	$64. \times 10^{-6}$
Viscosity (abs), centipoises[a]	0.018	0.014	0.013	0.095
Sound velocity in gas, m/sec	352.	215.	930.	316.
THERMAL AND THERMODYNAMIC PROPERTIES				
Specific heat, c_p, Btu/lb·deg F or cal/g·deg C	0.25	0.114	1.73	0.41
Specific heat, c_p, J/kg·K	1 046.	477.	7 238.	1 715.
Specific heat ratio, c_p/c_v	1.40	1.35	1.40	1.20
Gas constant R, ft-lb/lb·deg F	55.2	21.8	384.	51.4
Gas constant R, J/kg·deg C	297.	117.	2 066.	276.
Thermal conductivity, Btu/hr·ft·deg F	0.014	0.005	0.081	0.010
Thermal conductivity, W/m·deg C	0.024	0.008 7	0.140	0.017
Boiling point (sat 14.7 psia), deg F	−312.7	−29.2		−127.
Boiling point (sat 760 mm), deg C	−191.5	−34.		−88.3
Latent heat of evap (at bp), Btu/lb	92.8	123.7		210.
Latent heat of evap (at bp), J/kg	216 000.	288 000.		488 000.
Freezing (melting) point, deg F (1 atm)	−337.	−150.		−278.
Freezing (melting) point, deg C (1 atm)	−205.	−101.		−172.2
Latent heat of fusion, Btu/lb	12.8	41.0		41.
Latent heat of fusion, J/kg		95 400.		95 300.
Critical temperature, deg F	−220.	291.	−390.6	90.1
Critical temperature, deg C	−140.	144.	−234.8	32.2
Critical pressure, psia	507.	1 120.	241.	709.
Critical pressure, MN/m^2	3.49	7.72	1.66	4.89
Critical volume, ft^3/lb	0.053	0.028	0.239	0.076
Critical volume, m^3/kg	0.003 3	0.001 75	0.014 9	0.004 7
Flammable (yes or no)	Yes	No		Yes
Heat of combustion, Btu/ft^3	310.	—		
Heat of combustion, Btu/lb	4 340.	—		22 300.
Heat of combustion, kJ/kg	10 100.	—		51 800.

[a]For $N \cdot sec/m^2$ divide by 1 000.

TABLE A.7 (continued) Chemical, Physical, and Thermal Properties of Gases: Gases and Vapors, Including Fuels and Refrigerants, English and Metric Units

Common name(s)	*Ethyl chloride*	*Ethylene (Ethene)*	*Fluorine*
Chemical formula	C_2H_5Cl	C_2H_4	F_2
Refrigerant number	*160*	*1 150*	
CHEMICAL AND PHYSICAL PROPERTIES			
Molecular weight	64.515	28.054	37.996
Specific gravity, air = 1	2.23	0.969	1.31
Specific volume, ft^3/lb	6.07	13.9	10.31
Specific volume, m^3/kg	0.378	0.87	0.706
Density of liquid (at atm bp), lb/ft^3	56.5	35.5	
Density of liquid (at atm bp), kg/m^3	905.	569.	
Vapor pressure at 25 deg C, psia			
Vapor pressure at 25 deg C, MN/m^2			
Viscosity (abs), lbm/ft·sec		6.72×10^{-6}	16.1×10^{-6}
Viscosity (abs), centipoises[a]		0.010	0.024
Sound velocity in gas, m/sec	204.	331.	290.
THERMAL AND THERMODYNAMIC PROPERTIES			
Specific heat, c_p, Btu/lb·deg F or cal/g·deg C	0.27	0.37	0.198
Specific heat, c_p, J/kg·K	1 130.	1 548.	828.
Specific heat ratio, c_p/c_v	1.13	1.24	1.35
Gas constant R, ft-lb/lb·deg F	24.0	55.1	40.7
Gas constant R, J/kg·deg C	129.	296.	219.
Thermal conductivity, Btu/hr·ft·deg F		0.010	0.016
Thermal conductivity, W/m·deg C		0.017	0.028
Boiling point (sat 14.7 psia), deg F	54.	−155.	−306.4
Boiling point (sat 760 mm), deg C	12.2	−103.8	−188.
Latent heat of evap (at bp), Btu/lb	166.	208.	74.
Latent heat of evap (at bp), J/kg	386 000.	484 000.	172 000.
Freezing (melting) point, deg F (1 atm)	−218.	−272.	−364.
Freezing (melting) point, deg C (1 atm)	−138.9	−169.	−220.
Latent heat of fusion, Btu/lb	29.3	51.5	11.
Latent heat of fusion, J/kg	68 100.	120 000.	25 600.
Critical temperature, deg F	368.6	49.	−200
Critical temperature, deg C	187.	9.5	−129.
Critical pressure, psia	764.	741.	810.
Critical pressure, MN/m^2	5.27	5.11	5.58
Critical volume, ft^3/lb	0.049	0.073	
Critical volume, m^3/kg	0.003 06	0.004 6	
Flammable (yes or no)	No	Yes	
Heat of combustion, Btu/ft^3	—	1 480.	
Heat of combustion, Btu/lb	—	20 600.	
Heat of combustion, kJ/kg	—	47 800.	

[a]For $N \cdot sec/m^2$ divide by 1 000.

TABLE A.7 (continued) Chemical, Physical, and Thermal Properties of Gases: Gases and Vapors, Including Fuels and Refrigerants, English and Metric Units

Common name(s)	*Fluorocarbons*			
Chemical formula	CCl_3F	CCl_2F_2	$CClF_3$	$CBrF_3$
Refrigerant number	*11*	*12*	*13*	*13B1*
CHEMICAL AND PHYSICAL PROPERTIES				
Molecular weight	137.37	120.91	104.46	148.91
Specific gravity, air = 1	4.74	4.17	3.61	5.14
Specific volume, ft^3/lb	2.74	3.12	3.58	2.50
Specific volume, m^3/kg	0.171	0.195	0.224	0.975
Density of liquid (at atm bp), lb/ft^3	92.1	93.0	95.0	124.4
Density of liquid (at atm bp), kg/m^3	1 475.	1 490.	1 522.	1 993.
Vapor pressure at 25 deg C, psia		94.51	516.	234.8
Vapor pressure at 25 deg C, MN/m^2		0.652	3.56	1.619
Viscosity (abs), lbm/ft·sec	7.39×10^{-6}	8.74×10^{-6}		
Viscosity (abs), centipoises[a]	0.011	0.013		
Sound velocity in gas, m/sec				
THERMAL AND THERMODYNAMIC PROPERTIES				
Specific heat, c_p, Btu/lb·deg F or cal/g·deg C	0.14	0.146	0.154	
Specific heat, c_p, J/kg·K	586.	611.	644.	
Specific heat ratio, c_p/c_v	1.14	1.14	1.145	
Gas constant R, ft-lb/lb·deg F				
Gas constant R, J/kg·deg C				
Thermal conductivity, Btu/hr·ft·deg F	0.005	0.006		
Thermal conductivity, W/m·deg C	0.008 7	0.010 4		
Boiling point (sat 14.7 psia), deg F	74.9	−21.8	−114.6	−72.
Boiling point (sat 760 mm), deg C	23.8	−29.9	−81.4	−57.8
Latent heat of evap (at bp), Btu/lb	77.5	71.1	63.0	51.1
Latent heat of evap (at bp), J/kg	180 000.	165 000.	147 000.	119 000.
Freezing (melting) point, deg F (1 atm)	−168.	−252.	−294.	−270.
Freezing (melting) point, deg C (1 atm)	−111.	−157.8	−181.1	−167.8
Latent heat of fusion, Btu/lb				
Latent heat of fusion, J/kg				
Critical temperature, deg F	388.4	233.	83.9	152.
Critical temperature, deg C	198.	111.7	28.8	66.7
Critical pressure, psia	635.	582.	559.	573.
Critical pressure, MN/m^2	4.38	4.01	3.85	3.95
Critical volume, ft^3/lb	0.028 9	0.287	0.027 7	0.021 5
Critical volume, m^3/kg	0.001 80	0.018	0.001 73	0.001 34
Flammable (yes or no)	No	No	No	No
Heat of combustion, Btu/ft^3	—	—	—	—
Heat of combustion, Btu/lb	—	—	—	—
Heat of combustion, kJ/kg	—	—	—	—

[a]For $N \cdot sec/m^2$ divide by 1 000.

TABLE A.7 (continued) Chemical, Physical, and Thermal Properties of Gases: Gases and Vapors, Including Fuels and Refrigerants, English and Metric Units

Common name(s)	*Fluorocarbons*			
Chemical formula	CF_4	$CHCl_2F$	$CHClF_2$	$C_2Cl_2F_4$
Refrigerant number	*14*	*21*	*22*	*114*
CHEMICAL AND PHYSICAL PROPERTIES				
Molecular weight	88.00	102.92	86.468	170.92
Specific gravity, air = 1	3.04	3.55	2.99	5.90
Specific volume, ft^3/lb	4.34	3.7	4.35	2.6
Specific volume, m^3/kg	0.271	0.231	0.271	0.162
Density of liquid (at atm bp), lb/ft^3	102.0	87.7	88.2	94.8
Density of liquid (at atm bp), kg/m^3	1 634.	1 405.	1 413.	1 519.
Vapor pressure at 25 deg C, psia		26.4	151.4	30.9
Vapor pressure at 25 deg C, MN/m^2		0.182	1.044	0.213
Viscosity (abs), lbm/ft·sec		8.06×10^{-6}	8.74×10^{-6}	8.06×10^{-6}
Viscosity (abs), centipoises[a]		0.012	0.013	0.012
Sound velocity in gas, m/sec				
THERMAL AND THERMODYNAMIC PROPERTIES				
Specific heat, c_p, Btu/lb·deg F or cal/g·deg C		0.139	0.157	0.158
Specific heat, c_p, J/kg·K		582.	657.	661.
Specific heat ratio, c_p/c_v		1.18	1.185	1.09
Gas constant R, ft-lb/lb·deg F				
Gas constant R, J/kg·deg C				
Thermal conductivity, Btu/hr·ft·deg F			0.007	0.006
Thermal conductivity, W/m·deg C			0.012	0.010
Boiling point (sat 14.7 psia), deg F	−198.2	48.1	−41.3	38.4
Boiling point (sat 760 mm), deg C	−127.9	9.0	−40.7	3.55
Latent heat of evap (at bp), Btu/lb	58.5	104.1	100.4	58.4
Latent heat of evap (at bp), J/kg	136 000.	242 000.	234 000.	136 000.
Freezing (melting) point, deg F (1 atm)	−299.	−211.	−256.	−137.
Freezing (melting) point, deg C (1 atm)	−183.8	−135.	−160.	−93.8
Latent heat of fusion, Btu/lb	2.53			
Latent heat of fusion, J/kg	5 880.			
Critical temperature, deg F	−49.9	353.3	204.8	294.
Critical temperature, deg C	−45.5	178.5	96.5	
Critical pressure, psia	610.	750.	715.	475.
Critical pressure, MN/m^2	4.21	5.17	4.93	3.28
Critical volume, ft^3/lb	0.025	0.030 7	0.030 5	0.027 5
Critical volume, m^3/kg	0.001 6	0.001 91	0.001 90	0.001 71
Flammable (yes or no)	No	No	No	No
Heat of combustion, Btu/ft^3	—	—	—	—
Heat of combustion, Btu/lb	—	—	—	—
Heat of combustion, kJ/kg	—	—	—	—

[a]For $N \cdot sec/m^2$ divide by 1 000.

TABLE A.7 (continued) Chemical, Physical, and Thermal Properties of Gases: Gases and Vapors, Including Fuels and Refrigerants, English and Metric Units

Common name(s)	*Fluorocarbons*			*Helium*
Chemical formula	C_2ClF_5	$C_2H_3ClF_2$	$C_2H_4F_2$	*He*
Refrigerant number	*115*	*142b*	*152a*	*704*
CHEMICAL AND PHYSICAL PROPERTIES				
Molecular weight	154.47	100.50	66.05	4.002 6
Specific gravity, air = 1	5.33	3.47	2.28	0.138
Specific volume, ft^3/lb	2.44	3.7	5.9	97.86
Specific volume, m^3/kg	0.152	0.231	0.368	6.11
Density of liquid (at atm bp), lb/ft^3	96.5	74.6	62.8	7.80
Density of liquid (at atm bp), kg/m^3	1 546.	1 195.	1 006.	125.
Vapor pressure at 25 deg C, psia	132.1	49.1	86.8	
Vapor pressure at 25 deg C, MN/m^2	0.911	0.338 5	0.596	
Viscosity (abs), lbm/ft·sec				13.4×10^{-6}
Viscosity (abs), centipoises[a]				0.02
Sound velocity in gas, m/sec				1 015.
THERMAL AND THERMODYNAMIC PROPERTIES				
Specific heat, c_p, Btu/lb·deg F or cal/g·deg C	0.161			1.24
Specific heat, c_p, J/kg·K	674.			5 188.
Specific heat ratio, c_p/c_v	1.091			1.66
Gas constant R, ft-lb/lb·deg F				386.
Gas constant R, J/kg·deg C				2 077.
Thermal conductivity, Btu/hr·ft·deg F				0.086
Thermal conductivity, W/m·deg C				0.149
Boiling point (sat 14.7 psia), deg F	−38.0	14.	−13.	−452.
Boiling point (sat 760 mm), deg C	−38.9	−10.0	−25.0	4.22 K
Latent heat of evap (at bp), Btu/lb	53.4	92.5	137.1	10.0
Latent heat of evap (at bp), J/kg	124 000.	215 000.	319 000.	23 300.
Freezing (melting) point, deg F (1 atm)	−149.			[b]
Freezing (melting) point, deg C (1 atm)	−100.6			—
Latent heat of fusion, Btu/lb				—
Latent heat of fusion, J/kg				—
Critical temperature, deg F	176.		387.	−450.3
Critical temperature, deg C				5.2 K
Critical pressure, psia	457.6			33.22
Critical pressure, MN/m^2	3.155			
Critical volume, ft^3/lb	0.026 1			0.231
Critical volume, m^3/kg	0.001 63			0.014 4
Flammable (yes or no)	No	No	No	No
Heat of combustion, Btu/ft^3	—	—	—	—
Heat of combustion, Btu/lb	—	—	—	—
Heat of combustion, kJ/kg	—	—	—	—

[a]For N·sec/m^2 divide by 1 000.
[b]Helium cannot be solidified at atmospheric pressure.

TABLE A.7 (continued) Chemical, Physical, and Thermal Properties of Gases: Gases and Vapors, Including Fuels and Refrigerants, English and Metric Units

Common name(s)	*Hydrogen*	*Hydrogen chloride*	*Hydrogen sulfide*	*Krypton*
Chemical formula	H_2	HCl	H_2S	Kr
Refrigerant number	*702*	—	—	—
CHEMICAL AND PHYSICAL PROPERTIES				
Molecular weight	2.016	36.461	34.076	83.80
Specific gravity, air = 1	0.070	1.26	1.18	2.89
Specific volume, ft^3/lb	194.	10.74	11.5	4.67
Specific volume, m^3/kg	12.1	0.670	0.093 0	0.291
Density of liquid (at atm bp), lb/ft^3	4.43	74.4	62.	150.6
Density of liquid (at atm bp), kg/m^3	71.0	1 192.	993.	2 413.
Vapor pressure at 25 deg C, psia				
Vapor pressure at 25 deg C, MN/m^2				
Viscosity (abs), lbm/ft·sec	6.05×10^{-6}	10.1×10^{-6}	8.74×10^{-6}	16.8×10^{-6}
Viscosity (abs), centipoises[a]	0.009	0.015	0.013	0.025
Sound velocity in gas, m/sec	1 315.	310.	302.	223.
THERMAL AND THERMODYNAMIC PROPERTIES				
Specific heat, c_p, Btu/lb·deg F or cal/g·deg C	3.42	0.194	0.23	0.059
Specific heat, c_p, J/kg·K	14 310.	812.	962.	247.
Specific heat ratio, c_p/c_v	1.405	1.39	1.33	1.68
Gas constant R, ft-lb/lb·deg F	767.	42.4	45.3	18.4
Gas constant R, J/kg·deg C	4 126.	228.	244.	99.0
Thermal conductivity, Btu/hr·ft·deg F	0.105	0.008	0.008	0.005 4
Thermal conductivity, W/m·deg C	0.018 2	0.014	0.014	0.009 3
Boiling point (sat 14.7 psia), deg F	−423.	−121.	−76.	−244.
Boiling point (sat 760 mm), deg C	20.4 K	−85.	−60.	−153.
Latent heat of evap (at bp), Btu/lb	192.	190.5	234.	46.4
Latent heat of evap (at bp), J/kg	447 000.	443 000.	544 000.	108 000.
Freezing (melting) point, deg F (1 atm)	−434.6	−169.6	−119.2	−272.
Freezing (melting) point, deg C (1 atm)	−259.1	−112.	−84.	−169.
Latent heat of fusion, Btu/lb	25.0	23.4	30.2	4.7
Latent heat of fusion, J/kg	58 000.	54 400.	70 200.	10 900.
Critical temperature, deg F	−399.8	124.	213.	
Critical temperature, deg C	−240.0	51.2	100.4	−63.8
Critical pressure, psia	189.	1 201.	1 309.	800.
Critical pressure, MN/m^2	1.30	8.28	9.02	5.52
Critical volume, ft^3/lb	0.53	0.038	0.046	0.017 7
Critical volume, m^3/kg	0.033	0.002 4	0.002 9	0.001 1
Flammable (yes or no)	Yes	No	Yes	No
Heat of combustion, Btu/ft^3	320.	—	700.	—
Heat of combustion, Btu/lb	62 050.	—	8 000.	—
Heat of combustion, kJ/kg	144 000.	—	18 600.	—

[a]For $N \cdot sec/m^2$ divide by 1 000.

TABLE A.7 (continued) Chemical, Physical, and Thermal Properties of Gases: Gases and Vapors, Including Fuels and Refrigerants, English and Metric Units

Common name(s)	*Methane*	*Methyl chloride*	*Neon*	*Nitric oxide*
Chemical formula	CH_4	CH_3Cl	*Ne*	*NO*
Refrigerant number	*50*	*40*	*720*	—
CHEMICAL AND PHYSICAL PROPERTIES				
Molecular weight	16.044	50.488	20.179	30.006
Specific gravity, air = 1	0.554	1.74	0.697	1.04
Specific volume, ft^3/lb	24.2	7.4	19.41	13.05
Specific volume, m^3/kg	1.51	0.462	1.211	0.814
Density of liquid (at atm bp), lb/ft^3	26.3	62.7	75.35	
Density of liquid (at atm bp), kg/m^3	421.	1 004.	1 207.	
Vapor pressure at 25 deg C, psia		82.2		
Vapor pressure at 25 deg C, MN/m^2		0.567		
Viscosity (abs), lbm/ft·sec	7.39×10^{-6}	7.39×10^{-6}	21.5×10^{-6}	12.8×10^{-6}
Viscosity (abs), centipoises[a]	0.011	0.011	0.032	0.019
Sound velocity in gas, m/sec	446.	251.	454.	341.
THERMAL AND THERMODYNAMIC PROPERTIES				
Specific heat, c_p, Btu/lb·deg F or cal/g·deg C	0.54	0.20	0.246	0.235
Specific heat, c_p, J/kg·K	2 260.	837.	1 030.	983.
Specific heat ratio, c_p/c_v	1.31	1.28	1.64	1.40
Gas constant R, ft-lb/lb·deg F	96.	30.6	76.6	51.5
Gas constant R, J/kg·deg C	518.	165.	412.	277.
Thermal conductivity, Btu/hr·ft·deg F	0.02	0.006	0.028	0.015
Thermal conductivity, W/m·deg C	0.035	0.010	0.048	0.026
Boiling point (sat 14.7 psia), deg F	−259.	−10.7	−410.9	−240.
Boiling point (sat 760 mm), deg C	−434.2	−23.7	−246.	−151.5
Latent heat of evap (at bp), Btu/lb	219.2	184.1	37.	
Latent heat of evap (at bp), J/kg	510 000.	428 000.	86 100.	
Freezing (melting) point, deg F (1 atm)	−296.6	−144.	−415.6	−258.
Freezing (melting) point, deg C (1 atm)	−182.6	−97.8	−248.7	−161.
Latent heat of fusion, Btu/lb	14.	56.	6.8	32.9
Latent heat of fusion, J/kg	32 600.	130 000.	15 800.	76 500.
Critical temperature, deg F	−116.	289.4	−379.8	−136.
Critical temperature, deg C	−82.3	143.	−228.8	−93.3
Critical pressure, psia	673.	968.	396.	945.
Critical pressure, MN/m^2	4.64	6.67	2.73	6.52
Critical volume, ft^3/lb	0.099	0.043	0.033	0.033 2
Critical volume, m^3/kg	0.006 2	0.002 7	0.002 0	0.002 07
Flammable (yes or no)	Yes	Yes	No	No
Heat of combustion, Btu/ft^3	985.			—
Heat of combustion, Btu/lb	2 290.		—	—
Heat of combustion, kJ/kg			—	—

[a]For $N \cdot sec/m^2$ divide by 1 000.

TABLE A.7 (continued) Chemical, Physical, and Thermal Properties of Gases: Gases and Vapors, Including Fuels and Refrigerants, English and Metric Units

Common name(s)	*Nitrogen*	*Nitrous oxide*	*Oxygen*	*Ozone*
Chemical formula	N_2	N_2O	O_2	O_3
Refrigerant number	*728*	*744A*	*732*	—
CHEMICAL AND PHYSICAL PROPERTIES				
Molecular weight	28.013 4	44.012	31.998 8	47.998
Specific gravity, air = 1	0.967	1.52	1.105	1.66
Specific volume, ft^3/lb	13.98	8.90	12.24	8.16
Specific volume, m^3/kg	0.872	0.555	0.764	0.509
Density of liquid (at atm bp), lb/ft^3	50.46	76.6	71.27	
Density of liquid (at atm bp), kg/m^3	808.4	1 227.	1 142.	
Vapor pressure at 25 deg C, psia				
Vapor pressure at 25 deg C, MN/m^2				
Viscosity (abs), lbm/ft·sec	12.1×10^{-6}	10.1×10^{-6}	13.4×10^{-6}	8.74×10^{-6}
Viscosity (abs), centipoises[a]	0.018	0.015	0.020	0.013
Sound velocity in gas, m/sec	353.	268.	329.	
THERMAL AND THERMODYNAMIC PROPERTIES				
Specific heat, c_p, Btu/lb·deg F or cal/g·deg C	0.249	0.21	0.220	0.196
Specific heat, c_p, J/kg·K	1 040.	879.	920.	820.
Specific heat ratio, c_p/c_v	1.40	1.31	1.40	
Gas constant R, ft-lb/lb·deg F	55.2	35.1	48.3	32.2
Gas constant R, J/kg·deg C	297.	189.	260.	173.
Thermal conductivity, Btu/hr·ft·deg F	0.015	0.010	0.015	0.019
Thermal conductivity, W/m·deg C	0.026	0.017	0.026	0.033
Boiling point (sat 14.7 psia), deg F	−320.4	−127.3	−297.3	−170.
Boiling point (sat 760 mm), deg C	−195.8	−88.5	−182.97	−112.
Latent heat of evap (at bp), Btu/lb	85.5	161.8	91.7	
Latent heat of evap (at bp), J/kg	199 000.	376 000.	213 000.	
Freezing (melting) point, deg F (1 atm)	−346.	−131.5	−361.1	−315.5
Freezing (melting) point, deg C (1 atm)	−210.	−90.8	−218.4	−193.
Latent heat of fusion, Btu/lb	11.1	63.9	5.9	97.2
Latent heat of fusion, J/kg	25 800.	149 000.	13 700.	226 000.
Critical temperature, deg F	−232.6	97.7	−181.5	16.
Critical temperature, deg C	−147.	36.5	−118.6	−9.
Critical pressure, psia	493.	1 052.	726.	800.
Critical pressure, MN/m^2	3.40	7.25	5.01	5.52
Critical volume, ft^3/lb	0.051	0.036	0.040	0.029 8
Critical volume, m^3/kg	0.003 18	0.002 2	0.002 5	0.001 86
Flammable (yes or no)	No	No	No	No
Heat of combustion, Btu/ft^3	—	—	—	—
Heat of combustion, Btu/lb	—	—	—	—
Heat of combustion, kJ/kg	—	—	—	—

[a]For $N \cdot sec/m^2$ divide by 1 000.

TABLE A.7 (continued) Chemical, Physical, and Thermal Properties of Gases: Gases and Vapors, Including Fuels and Refrigerants, English and Metric Units

Common name(s)	*Propane*	*Propylene (Propene)*	*Sulfur dioxide*	*Xenon*
Chemical formula	C_3H_8	C_3H_6	SO_2	*Xe*
Refrigerant number	*290*	*1 270*	*764*	—
CHEMICAL AND PHYSICAL PROPERTIES				
Molecular weight	44.097	42.08	64.06	131.30
Specific gravity, air = 1	1.52	1.45	2.21	4.53
Specific volume, ft^3/lb	8.84	9.3	6.11	2.98
Specific volume, m^3/kg	0.552	0.58		
Density of liquid (at atm bp), lb/ft^3	36.2	37.5	42.8	190.8
Density of liquid (at atm bp), kg/m^3	580.	601.	585.	3 060.
Vapor pressure at 25 deg C, psia	135.7	166.4	56.6	
Vapor pressure at 25 deg C, MN/m^2	0.936	1.147	0.390	
Viscosity (abs), lbm/ft·sec	53.8×10^{-6}	57.1×10^{-6}	8.74×10^{-6}	15.5×10^{-6}
Viscosity (abs), centipoises[a]	0.080	0.085	0.013	0.023
Sound velocity in gas, m/sec	253.	261.	220.	177.
THERMAL AND THERMODYNAMIC PROPERTIES				
Specific heat, c_p, Btu/lb·deg F or cal/g·deg C	0.39	0.36	0.11	0.115
Specific heat, c_p, J/kg·K	1 630.	1 506.	460.	481.
Specific heat ratio, c_p/c_v	1.2	1.16	1.29	1.67
Gas constant R, ft-lb/lb·deg F	35.0	36.7	24.1	11.8
Gas constant R, J/kg·deg C	188.	197.	130.	63.5
Thermal conductivity, Btu/hr·ft·deg F	0.010	0.010	0.006	0.003
Thermal conductivity, W/m·deg C	0.017	0.017	0.010	0.005 2
Boiling point (sat 14.7 psia), deg F	−44.	−54.	14.0	−162.5
Boiling point (sat 760 mm), deg C	−42.2	−48.3	−10.	−108.
Latent heat of evap (at bp), Btu/lb	184.	188.2	155.5	41.4
Latent heat of evap (at bp), J/kg	428 000.	438 000.	362 000.	96 000.
Freezing (melting) point, deg F (1 atm)	−309.8	−301.	−104.	−220.
Freezing (melting) point, deg C (1 atm)	−189.9	−185.	−75.5	−140.
Latent heat of fusion, Btu/lb	19.1		58.0	10.
Latent heat of fusion, J/kg	44 400.		135 000.	23 300.
Critical temperature, deg F	205.	197.	315.5	61.9
Critical temperature, deg C	96.	91.7	157.6	16.6
Critical pressure, psia	618.	668.	1 141.	852.
Critical pressure, MN/m^2	4.26	4.61	7.87	5.87
Critical volume, ft^3/lb	0.073	0.069	0.03	0.014 5
Critical volume, m^3/kg	0.004 5	0.004 3	0.001 9	0.000 90
Flammable (yes or no)	Yes	Yes	No	No
Heat of combustion, Btu/ft^3	2 450.	2 310.	—	—
Heat of combustion, Btu/lb	21 660.	21 500.	—	—
Heat of combustion, kJ/kg	50 340.	50 000.	—	—

[a]For $N \cdot sec/m^2$ divide by 1 000.

TABLE A.8 Ideal Gas Properties of Air

Part a. SI Units

T(K), h and u(kJ/kg), s°(kJ/kg·K)

T	h	p_r	u	v_r	s°	T	h	p_r	u	v_r	s°
200	199.97	0.3363	142.56	1707.	1.29559	450	451.80	5.775	322.62	223.6	2.11161
210	209.97	0.3987	149.69	1512.	1.34444	460	462.02	6.245	329.97	211.4	2.13407
220	219.97	0.4690	156.82	1346.	1.39105	470	472.24	6.742	337.32	200.1	2.15604
230	230.02	0.5477	164.00	1205.	1.43557	480	482.49	7.268	344.70	189.5	2.17760
240	240.02	0.6355	171.13	1084.	1.47824	490	492.74	7.824	352.08	179.7	2.19876
250	250.05	0.7329	178.28	979.	1.51917	500	503.02	8.411	359.49	170.6	2.21952
260	260.09	0.8405	185.45	887.8	1.55848	510	513.32	9.031	366.92	162.1	2.23993
270	270.11	0.9590	192.60	808.0	1.59634	520	523.63	9.684	374.36	154.1	2.25997
280	280.13	1.0889	199.75	738.0	1.63279	530	533.98	10.37	381.84	146.7	2.27967
285	285.14	1.1584	203.33	706.1	1.65055	540	544.35	11.10	389.34	139.7	2.29906
290	290.16	1.2311	206.91	676.1	1.66802	550	554.74	11.86	396.86	133.1	2.31809
295	295.17	1.3068	210.49	647.9	1.68515	560	565.17	12.66	404.42	127.0	2.33685
300	300.19	1.3860	214.07	621.2	1.70203	570	575.59	13.50	411.97	121.2	2.35531
305	305.22	1.4686	217.67	596.0	1.71865	580	586.04	14.38	419.55	115.7	2.37348
310	310.24	1.5546	221.25	572.3	1.73498	590	596.52	15.31	427.15	110.6	2.39140
315	315.27	1.6442	224.85	549.8	1.75106	600	607.02	16.28	434.78	105.8	2.40902
320	320.29	1.7375	228.42	528.6	1.76690	610	617.53	17.30	442.42	101.2	2.42644
325	325.31	1.8345	232.02	508.4	1.78249	620	628.07	18.36	450.09	96.92	2.44356
330	330.34	1.9352	235.61	489.4	1.79783	630	638.63	19.84	457.78	92.84	2.46048
340	340.42	2.149	242.82	454.1	1.82790	640	649.22	20.64	465.50	88.99	2.47716
350	350.49	2.379	250.02	422.2	1.85708	650	659.84	21.86	473.25	85.34	2.49364
360	360.58	2.626	257.24	393.4	1.88543	660	670.47	23.13	481.01	81.89	2.50985
370	370.67	2.892	264.46	367.2	1.91313	670	681.14	24.46	488.81	78.61	2.52589
380	380.77	3.176	271.69	343.4	1.94001	680	691.82	25.85	496.62	75.50	2.54175
390	390.88	3.481	278.93	321.5	1.96633	690	702.52	27.29	504.45	72.56	2.55731
400	400.98	3.806	286.16	301.6	1.99194	700	713.27	28.80	512.33	69.76	2.57277
410	411.12	4.153	293.43	283.3	2.01699	710	724.04	30.38	520.23	67.07	2.58810
420	421.26	4.522	300.69	266.6	2.04142	720	734.82	32.02	528.14	64.53	2.60319
430	431.43	4.915	307.99	251.1	2.06533	730	745.62	33.72	536.07	62.13	2.61803
440	441.61	5.332	315.30	236.8	2.08870	740	756.44	35.50	544.02	59.82	2.63280

TABLE A.8 (continued) Ideal Gas Properties of Air

T(K), h and u(kJ/kg), s°(kJ/kg·K)											
T	h	p_r	u	v_r	s°	T	h	p_r	u	v_r	s°
750	767.29	37.35	551.99	57.63	2.64737	1300	1395.97	330.9	1022.82	11.275	3.27345
760	778.18	39.27	560.01	55.54	2.66176	1320	1419.76	352.5	1040.88	10.747	3.29160
770	789.11	41.31	568.07	53.39	2.67595	1340	1443.60	375.3	1058.94	10.247	3.30959
780	800.03	43.35	576.12	51.64	2.69013	1360	1467.49	399.1	1077.10	9.780	3.32724
790	810.99	45.55	584.21	49.86	2.70400	1380	1491.44	424.2	1095.26	9.337	3.34474
800	821.95	47.75	592.30	48.08	2.71787	1400	1515.42	450.5	1113.52	8.919	3.36200
820	843.98	52.59	608.59	44.84	2.74504	1420	1539.44	478.0	1131.77	8.526	3.37901
840	866.08	57.60	624.95	41.85	2.77170	1440	1563.51	506.9	1150.13	8.153	3.39586
860	888.27	63.09	641.40	39.12	2.79783	1460	1587.63	537.1	1168.49	7.801	3.41247
880	910.56	68.98	657.95	36.61	2.82344	1480	1611.79	568.8	1186.95	7.468	3.42892
900	932.93	75.29	674.58	34.31	2.84856	1500	1635.97	601.9	1205.41	7.152	3.44516
920	955.38	82.05	691.28	32.18	2.87324	1520	1660.23	636.5	1223.87	6.854	3.46120
940	977.92	89.28	708.08	30.22	2.89748	1540	1684.51	672.8	1242.43	6.569	3.47712
960	1000.55	97.00	725.02	28.40	2.92128	1560	1708.82	710.5	1260.99	6.301	3.49276
980	1023.25	105.2	741.98	26.73	2.94468	1580	1733.17	750.0	1279.65	6.046	3.50829
1000	1046.04	114.0	758.94	25.17	2.96770	1600	1757.57	791.2	1298.30	5.804	3.52364
1020	1068.89	123.4	776.10	23.72	2.99034	1620	1782.00	834.1	1316.96	5.574	3.53879
1040	1091.85	133.3	793.36	22.39	3.01260	1640	1806.46	878.9	1335.72	5.355	3.55381
1060	1114.86	143.9	810.62	21.14	3.03449	1660	1830.96	925.6	1354.48	5.147	3.56867
1080	1137.89	155.2	827.88	19.98	3.05608	1680	1855.50	974.2	1373.24	4.949	3.58335
1100	1161.07	167.1	845.33	18.896	3.07732	1700	1880.1	1025	1392.7	4.761	3.5979
1120	1184.28	179.7	862.79	17.886	3.09825	1750	1941.6	1161	1439.8	4.328	3.6336
1140	1207.57	193.1	880.35	16.946	3.11883	1800	2003.3	1310	1487.2	3.944	3.6684
1160	1230.92	207.2	897.91	16.064	3.13916	1850	2065.3	1475	1534.9	3.601	3.7023
1180	1254.34	222.2	915.57	15.241	3.15916	1900	2127.4	1655	1582.6	3.295	3.7354
1200	1277.79	238.0	933.33	14.470	3.17888	1950	2189.7	1852	1630.6	3.022	3.7677
1220	1301.31	254.7	951.09	13.747	3.19834	2000	2252.1	2068	1678.7	2.776	3.7994
1240	1324.93	272.3	968.95	13.069	3.21751	2050	2314.6	2303	1726.8	2.555	3.8303
1260	1348.55	290.8	986.90	12.435	3.23638	2100	2377.4	2559	1775.3	2.356	3.8605
1280	1372.24	310.4	1004.76	11.835	3.25510	2150	2440.3	2837	1823.8	2.175	3.8901
						2200	2503.2	3138	1872.4	2.012	3.9191
						2250	2566.4	3464	1921.3	1.864	3.9474

TABLE A.8 (continued) Ideal Gas Properties of Air

Part b. English Units

T(°R), h and u(Btu/lb), s^o(Btu/lb · °R)

T	h	p_r	u	v_r	s^o	T	h	p_r	u	v_r	s^o
360	85.97	0.3363	61.29	396.6	0.50369	940	226.11	9.834	161.68	35.41	0.73509
380	90.75	0.4061	64.70	346.6	0.51663	960	231.06	10.61	165.26	33.52	0.74030
400	95.53	0.4858	68.11	305.0	0.52890	980	236.02	11.43	168.83	31.76	0.74540
420	100.32	0.5760	71.52	270.1	0.54058	1000	240.98	12.30	172.43	30.12	0.75042
440	105.11	0.6776	74.93	240.6	0.55172	1040	250.95	14.18	179.66	27.17	0.76019
460	109.90	0.7913	78.36	215.33	0.56235	1080	260.97	16.28	186.93	24.58	0.76964
480	114.69	0.9182	81.77	193.65	0.57255	1120	271.03	18.60	194.25	22.30	0.77880
500	119.48	1.0590	85.20	174.90	0.58233	1160	281.14	21.18	201.63	20.29	0.78767
520	124.27	1.2147	88.62	158.58	0.59172	1200	291.30	24.01	209.05	18.51	0.79628
537	128.34	1.3593	91.53	146.34	0.59945	1240	301.52	27.13	216.53	16.93	0.80466
540	129.06	1.3860	92.04	144.32	0.60078	1280	311.79	30.55	224.05	15.52	0.81280
560	133.86	1.5742	95.47	131.78	0.60950	1320	322.11	34.31	231.63	14.25	0.82075
580	138.66	1.7800	98.90	120.70	0.61793	1360	332.48	38.41	239.25	13.12	0.82848
600	143.47	2.005	102.34	110.88	0.62607	1400	342.90	42.88	246.93	12.10	0.83604
620	148.28	2.249	105.78	102.12	0.63395	1440	353.37	47.75	254.66	11.17	0.84341
640	153.09	2.514	109.21	94.30	0.64159	1480	363.89	53.04	262.44	10.34	0.85062
660	157.92	2.801	112.67	87.27	0.64902	1520	374.47	58.78	270.26	9.578	0.85767
680	162.73	3.111	116.12	80.96	0.65621	1560	385.08	65.00	278.13	8.890	0.86456
700	167.56	3.446	119.58	75.25	0.66321	1600	395.74	71.73	286.06	8.263	0.87130
720	172.39	3.806	123.04	70.07	0.67002	1650	409.13	80.89	296.03	7.556	0.87954
740	177.23	4.193	126.51	65.38	0.67665	1700	422.59	90.95	306.06	6.924	0.88758
760	182.08	4.607	129.99	61.10	0.68312	1750	436.12	101.98	316.16	6.357	0.89542
780	186.94	5.051	133.47	57.20	0.68942	1800	449.71	114.0	326.32	5.847	0.90308
800	191.81	5.526	136.97	53.63	0.69558	1850	463.37	127.2	336.55	5.388	0.91056
820	196.69	6.033	140.47	50.35	0.70160	1900	477.09	141.5	346.85	4.974	0.91788
840	201.56	6.573	143.98	47.34	0.70747	1950	490.88	157.1	357.20	4.598	0.92504
860	206.46	7.149	147.50	44.57	0.71323	2000	504.71	174.0	367.61	4.258	0.93205
880	211.35	7.761	151.02	42.01	0.71886	2050	518.61	192.3	378.08	3.949	0.93891
900	216.26	8.411	154.57	39.64	0.72438	2100	532.55	212.1	388.60	3.667	0.94564
920	221.18	9.102	158.12	37.44	0.72979	2150	546.54	233.5	399.17	3.410	0.95222

TABLE A.8 (continued) Ideal Gas Properties of Air

T(°R), h and u(Btu/lb), s^o(Btu/lb · °R)

T	h	p_r	u	v_r	s^o	T	h	p_r	u	v_r	s^o
2200	560.59	256.6	409.78	3.176	0.95868	3700	998.11	2330	744.48	.5882	1.10991
2250	574.69	281.4	420.46	2.961	0.96501	3750	1013.1	2471	756.04	.5621	1.11393
2300	588.82	308.1	431.16	2.765	0.97123	3800	1028.1	2618	767.60	.5376	1.11791
2350	603.00	336.8	441.91	2.585	0.97732	3850	1043.1	2773	779.19	.5143	1.12183
2400	617.22	367.6	452.70	2.419	0.98331	3900	1058.1	2934	790.80	.4923	1.12571
2450	631.48	400.5	463.54	2.266	0.98919	3950	1073.2	3103	802.43	.4715	1.12955
2500	645.78	435.7	474.40	2.125	0.99497	4000	1088.3	3280	814.06	.4518	1.13334
2550	660.12	473.3	485.31	1.996	1.00064	4050	1103.4	3464	825.72	.4331	1.13709
2600	674.49	513.5	496.26	1.876	1.00623	4100	1118.5	3656	837.40	.4154	1.14079
2650	688.90	556.3	507.25	1.765	1.01172	4150	1133.6	3858	849.09	.3985	1.14446
2700	703.35	601.9	518.26	1.662	1.01712	4200	1148.7	4067	860.81	.3826	1.14809
2750	717.83	650.4	529.31	1.566	1.02244	4300	1179.0	4513	884.28	.3529	1.15522
2800	732.33	702.0	540.40	1.478	1.02767	4400	1209.4	4997	907.81	.3262	1.16221
2850	746.88	756.7	551.52	1.395	1.03282	4500	1239.9	5521	931.39	.3019	1.16905
2900	761.45	814.8	562.66	1.318	1.03788	4600	1270.4	6089	955.04	.2799	1.17575
2950	776.05	876.4	573.84	1.247	1.04288	4700	1300.9	6701	978.73	.2598	1.18232
3000	790.68	941.4	585.04	1.180	1.04779	4800	1331.5	7362	1002.5	.2415	1.18876
3050	805.34	1011	596.28	1.118	1.05264	4900	1362.2	8073	1026.3	.2248	1.19508
3100	820.03	1083	607.53	1.060	1.05741	5000	1392.9	8837	1050.1	.2096	1.20129
3150	834.75	1161	618.82	1.006	1.06212	5100	1423.6	9658	1074.0	.1956	1.20738
3200	849.48	1242	630.12	.9546	1.06676	5200	1454.4	10539	1098.0	.1828	1.21336
3250	864.24	1328	641.46	.9069	1.07134	5300	1485.3	11481	1122.0	.1710	1.21923
3300	879.02	1418	652.81	.8621	1.07585						
3350	893.83	1513	664.20	.8202	1.08031						
3400	908.66	1613	675.60	.7807	1.08470						
3450	923.52	1719	687.04	.7436	1.08904						
3500	938.40	1829	698.48	.7087	1.09332						
3550	953.30	1946	709.95	.6759	1.09755						
3600	968.21	2068	721.44	.6449	1.10172						
3650	983.15	2196	732.95	.6157	1.10584						

Source: Adapted from M.J. Moran and H.N. Shapiro, *Fundamentals of Engineering Thermodynamics*, 3rd. ed., Wiley, New York, 1995, as based on J.H. Keenan and J. Kaye, *Gas Tables*, Wiley, New York, 1945.

TABLE A.9 Equations for Gas Properties

Gas	Molar Mass M kg/kmol	Gas Constant R kJ/kg·K	Specific Heats at 25°C c_p kJ/kg·K	c_v kJ/kg·K	k	Equation Coefficients for $c_p/R = a + bT + cT^2 + dT^3 + eT^4$ Temperature Range	a	$b \times 10^3$ K^{-1}	$c \times 10^6$ K^{-2}	$d \times 10^{10}$ K^{-3}	$e \times 10^{13}$ K^{-4}	Critical State Properties p_c MPa	T_c K	Redlich–Kwong Constants a kPa·m^6·K$^{0.5}$/kmol^2	b m^3/kmol	Gas
Acetylene, C_2H_2	26.04	0.319	1.69	1.37	1.232	300–1000K	0.8021	23.51	−35.95	286.1	−87.64	6.14	308	8030	0.0362	Acetylene, C_2H_2
						1000–3000K	3.825	6.767	−3.014	6.931	−0.6469					
Air	28.97	0.287	1.01	0.718	1.400	300–1000K	3.721	−1.874	4.719	−34.45	8.531	3.77	132	1580	0.0253	Air
						1000–3000K	2.786	1.925	−0.9465	2.321	−0.2229					
Argon, Ar	39.95	0.208	0.520	0.312	1.667		2.50	0	0	0	0	4.90	151	1680	0.0222	Argon, Ar
Butane, C_4H_{10}	58.12	0.143	1.67	1.53	1.094	300–1500K	0.4756	44.65	−22.04	42.07	0	3.80	425	29000	0.0806	Butane, C_4H_{10}
Carbon Dioxide CO_2	44.01	0.189	0.844	0.655	1.289	300–1000K	2.227	9.992	−9.802	53.97	−12.81	7.38	304	6450	0.0297	Carbon Dioxide CO_2
						1000–3000K	3.247	5.847	−3.412	9.469	−1.009					
Carbon Monoxide CO	28.01	0.297	1.04	0.744	1.399	300–1000K	3.776	−2.093	4.880	−32.71	6.984	3.50	133	1720	0.0274	Carbon Monoxide, CO
						1000–3000K	2.654	2.226	−1.146	2.851	−0.2762					
Ethane, C_2H_6	30.07	0.276	1.75	1.48	1.187	300–1500K	0.8293	20.75	−7.704	8.756	0	4.88	306	9860	0.0450	Ethane, C_2H_6
Ethylene, C_2H_4	28.05	0.296	1.53	1.23	1.240	300–1000K	1.575	10.19	11.25	−199.1	81.98	5.03	282	7860	0.0404	Ethylene, C_2H_4
						1000–3000K	0.2530	18.67	−9.978	26.03	−2.668					
Helium, He	4.003	2.08	5.19	3.12	1.667		2.50	0	0	0	0	0.228	5.20	8.00	0.0165	Helium, He
Hydrogen, H_2	2.016	4.12	14.3	10.2	1.405	300–1000K	2.892	3.884	−8.850	86.94	−29.88	1.31	33.2	143	0.0182	Hydrogen, H_2
						1000–3000K	3.717	−0.9220	1.221	−4.328	0.5202					
Hydrogen, H	1.008	8.25	20.6	12.4	1.667	300–1000K	2.496	0.02977	−0.07655	0.8238	−0.3158					Hydrogen, H
						1000–3000K	2.567	−0.1509	0.1219	−0.4184	0.05182					
Hydroxyl, OH	17.01	0.489	1.76	1.27	1.384	300–1000K	3.874	−1.349	1.670	−5.670	0.6189					Hydroxyl, OH
						1000–3000K	3.229	0.2014	0.4357	−2.043	0.2696					
Methane, CH_4	16.04	0.518	2.22	1.70	1.304	300–1000K	4.503	−8.965	37.38	−364.9	122.2	4.60	191	3210	0.0298	Methane, CH_4
						1000–3000K	−0.6992	15.31	−7.695	18.96	−1.849					
Neon, Ne	20.18	0.412	1.03	0.618	1.667		2.50	0	0	0	0	2.65	44.4	146	0.0120	Neon, Ne
Nitric Oxide, NO	30.01	0.277	0.995	0.718	1.386	300–1000K	4.120	−4.225	10.77	−97.64	31.85	6.48	180	1980	0.0200	Nitric Oxide, NO
						1000–3000K	2.730	2.372	−1.338	3.604	−0.3743					
Nitrogen, N_2	28.01	0.297	1.04	0.743	1.400	300–1000K	3.725	−1.562	3.208	−15.54	1.154	3.39	126	1550	0.0267	Nitrogen, N_2
						1000–3000K	2.469	2.467	−1.312	3.401	−0.3454					
Nitrogen, N	14.01	0.594	1.48	0.890	1.667	300–1000K	2.496	0.02977	−0.07655	0.8238	−0.3158					Nitrogen, N
						1000–3000K	2.483	0.03033	−0.01517	0.001879	0.009657					
Oxygen, O_2	32.00	0.260	0.919	0.659	1.395	300–1000K	3.837	−3.420	10.99	−109.6	37.47	5.04	155	1740	0.0221	Oxygen, O_2
						1000–3000K	3.156	1.809	−1.052	3.190	−0.3629					
Oxygen, O	16.00	0.520	1.37	0.850	1.612	300–1000K	3.020	−2.176	3.793	−30.62	9.402					Oxygen, O
						1000–3000K	2.662	−0.3051	0.2250	−0.7447	0.09383					
Propane, C_3H_8	44.10	0.189	1.67	1.48	1.127	300-1500K	-0.4861	36.63	-18.91	38.14	0	4.26	370	18300	0.0626	Propane, C_3H_8
Water, H_2O	18.02	0.462	1.86	1.40	1.329	300–1000K	4.132	−1.559	5.315	−42.09	12.84	22.1	647	14300	0.0211	Water, H_2O
						1000–3000K	2.798	2.693	−0.5392	−0.01783	0.09027					

Source: Adapted from J.B. Jones and R.E. Dugan, *Engineering Thermodynamics*, Prentice-Hall, Englewood Cliffs, NJ 1996 from various sources: *JANAF Thermochemical Tables*, 3rd ed., published by the American Chemical Society and the American Institute of Physics for the National Bureau of Standards, 1986. Data for butane, ethane, and propane from K.A. Kobe and E.G. Long, "Thermochemistry for the Petrochemical Industry, Part II — Paraffinic Hydrocarbons, C_1–C_{16}" *Petroleum Refiner*, Vol. 28, No. 2, 1949, pp. 113–116.

Appendix B. Properties of Liquids

TABLE B.1 Properties of Liquid Water*

Symbols and Units:

ρ = density, lbm/ft^3. For g/cm^3 multiply by 0.016018. For kg/m^3 multiply by 16.018.
c_p = specific heat, Btu/lbm·deg R = cal/g·K. For J/kg·K multiply by 4186.8
μ = viscosity. For lbf·sec/ft^2 = slugs/sec·ft, multiply by 10^{-7}. For lbm·sec·ft multiply by 10^{-7} and by 32.174. For g/sec·cm (poises) multiply by 10^{-7} and by 478.80. For N·sec/m^2 multiply by 10^{-7} and by 478.880.
k = thermal conductivity, Btu/hr·ft·deg R. For W/m·K multiply by 1.7307.

Temp., °F	*At 1 atm or 14.7 psia*				*At 1,000 psia*				*At 10,000 psia*			
	ρ	c_p	μ	k	ρ	c_p	μ	k	ρ	c_p	μ	k†
32	62.42	1.007	366	0.3286	62.62	0.999	365	0.3319	64.5	0.937	357	0.3508
40	62.42	1.004	323	0.334	62.62	0.997	323	0.337	64.5	0.945	315	0.356
50	62.42	1.002	272	0.3392	62.62	0.995	272	0.3425	64.5	0.951	267	0.3610
60	62.38	1.000	235	0.345	62.58	0.994	235	0.348	64.1	0.956	233	0.366
70	62.31	0.999	204	0.350	62.50	0.994	204	0.353	64.1	0.960	203	0.371
80	62.23	0.998	177	0.354	62.42	0.994	177	0.358	64.1	0.962	176	0.376
90	62.11	0.998	160	0.359	62.31	0.994	160	0.362	63.7	0.964	159	0.380
100	62.00	0.998	142	0.3633	62.19	0.994	142	0.3666	63.7	0.965	142	0.3841
110	61.88	0.999	126	0.367	62.03	0.994	126	0.371	63.7	0.966	126	0.388
120	61.73	0.999	114	0.371	61.88	0.995	114	0.374	63.3	0.967	114	0.391
130	61.54	0.999	105	0.374	61.73	0.995	105	0.378	63.3	0.968	105	0.395
140	61.39	0.999	96	0.378	61.58	0.996	96	0.381	63.3	0.969	98	0.398
150	61.20	1.000	89	0.3806	61.39	0.996	89	0.3837	63.0	0.970	91	0.4003
160	61.01	1.001	83	0.383	61.20	0.997	83	0.386	62.9	0.971	85	0.403
170	60.79	1.002	77	0.386	60.98	0.998	77	0.389	62.5	0.972	79	0.405
180	60.57	1.003	72	0.388	60.75	0.999	72	0.391	62.5	0.973	74	0.407
190	60.35	1.004	68	0.390	60.53	1.001	68	0.393	62.1	0.974	70	0.409
200	60.10	1.005	62.5	0.3916	60.31	1.002	62.9	0.3944	62.1	0.975	65.4	0.4106
250	boiling point 212°F				59.03	1.001	47.8	0.3994	60.6	0.981	50.6	0.4158
300					57.54	1.024	38.4	0.3993	59.5	0.988	41.3	0.4164
350					55.83	1.044	32.1	0.3944	58.1	0.999	35.1	0.4132
400					53.91	1.072	27.6	0.3849	56.5	1.011	30.6	0.4064
500					49.11	1.181	21.6	0.3508	52.9	1.051	24.8	0.3836
600					boiling point 544.58°F				48.3	1.118	21.0	0.3493

†At 7,500 psia.
*From: "1967 ASME Steam Tables", American Society of Mechanical Engineers, Tables 9, 10, and 11 and Figures 6, 7, 8, and 9.

The ASME compilation is a 330-page book of tables and charts, including a $2\frac{1}{2} \times 3\frac{1}{2}$-ft Mollier chart. All values have been computed in accordance with the 1967 specifications of the International Formulation Committee (IFC) and are in conformity with the 1963 International Skeleton Tables. This standardization of tables began in 1921 and was extended through the International Conferences in London (1929), Berlin (1930), Washington (1934), Philadelphia (1954), London (1956), New York (1963) and Glasgow (1966). Based on these world-wide standard data, the 1967 ASME volume represents detailed computer output in both tabular and graphic form. Included are density and volume, enthalpy, entropy, specific heat, viscosity, thermal conductivity, Prandtl number, isentropic exponent, choking velocity, p-v product, etc., over the entire range (to 1500 psia 1500°F). English units are used, but all conversion factors are given.

TABLE B.2 Physical and Thermal Properties of Common Liquids

Part a. SI Units

(At 1.0 Atm Pressure (0.101 325 MN/m^2), 300 K, except as noted.)

Common name	*Density, kg/m^3*	*Specific heat, $kJ/kg \cdot K$*	*Viscosity, $N \cdot s/m^2$*	*Thermal conductivity, $W/m \cdot K$*	*Freezing point, K*	*Latent heat of fusion, kJ/kg*	*Boiling point, K*	*Latent heat of evaporation, kJ/kg*	*Coefficient of cubical expansion per K*
Acetic acid	1 049	2.18	.001 155	0.171	290	181	391	402	0.001 1
Acetone	784.6	2.15	.000 316	0.161	179.0	98.3	329	518	0.001 5
Alcohol, ethyl	785.1	2.44	.001 095	0.171	158.6	108	351.46	846	0.001 1
Alcohol, methyl	786.5	2.54	.000 56	0.202	175.5	98.8	337.8	1 100	0.001 4
Alcohol, propyl	800.0	2.37	.001 92	0.161	146	86.5	371	779	
Ammonia (aqua)	823.5	4.38		0.353					
Benzene	873.8	1.73	.000 601	0.144	278.68	126	353.3	390	0.001 3
Bromine		.473	.000 95		245.84	66.7	331.6	193	0.001 2
Carbon disulfide	1 261	.992	.000 36	0.161	161.2	57.6	319.40	351	0.001 3
Carbon tetrachloride	1 584	.866	.000 91	0.104	250.35	174	349.6	194	0.001 3
Castor oil	956.1	1.97	.650	0.180	263.2				
Chloroform	1 465	1.05	.000 53	0.118	209.6	77.0	334.4	247	0.001 3
Decane	726.3	2.21	.000 859	0.147	243.5	201	447.2	263	
Dodecane	754.6	2.21	.001 374	0.140	247.18	216	489.4	256	
Ether	713.5	2.21	.000 223	0.130	157	96.2	307.7	372	0.001 6
Ethylene glycol	1 097	2.36	.016 2	0.258	260.2	181	470	800	
Fluorine refrigerant R-11	1 476	.870[a]	.000 42	0.093[a]	162		297.0	180[b]	
Fluorine refrigerant R-12	1 311	.971[a]		0.071[a]	115	34.4	243.4	165[b]	
Fluorine refrigerant R-22	1 194	1.26[a]		0.086[a]	113	183	232.4	232[b]	
Glycerine	1 259	2.62	.950	0.287	264.8	200	563.4	974	0.000 54
Heptane	679.5	2.24	.000 376	0.128	182.54	140	371.5	318	
Hexane	654.8	2.26	.000 297	0.124	178.0	152	341.84	365	
Iodine		2.15			386.6	62.2	457.5	164	
Kerosene	820.1	2.09	.001 64	0.145				251	
Linseed oil	929.1	1.84	.033 1		253		560		
Mercury		.139	.001 53		234.3	11.6	630	295	0.000 18
Octane	698.6	2.15	.000 51	0.131	216.4	181	398	298	0.000 72
Phenol	1 072	1.43	.008 0	0.190	316.2	121	455		0.000 90
Propane	493.5	2.41[a]	.000 11		85.5	79.9	231.08	428[b]	
Propylene	514.4	2.85	.000 09		87.9	71.4	225.45	342	
Propylene glycol	965.3	2.50	.042		213		460	914	
Sea water	1 025	3.76–4.10			270.6				
Toluene	862.3	1.72	.000 550	0.133	178	71.8	383.6	363	
Turpentine	868.2	1.78	.001 375	0.121	214		433	293	0.000 99
Water	997.1	4.18	.000 89	0.609	273	333	373	2 260	0.000 20

[a] At 297 K, liquid.
[b] At .101 325 meganewtons, saturation temperature.

TABLE B.2 (continued) Physical and Thermal Properties of Common Liquids

Part b. English Units

(At 1.0 Atm Pressure 77°F (25°C), except as noted.)

For viscosity in N·s/m² (=kg m·s), multiply values in centipoises by 0.001. For surface tension in N/m, multiply values in dyne/cm by 0.001.

Common name	*Density,* $\frac{lb}{ft^3}$	*Specific gravity*	*Viscosity*		*Sound velocity,* $\frac{meters}{sec}$	*Dielectric constant*	*Refractive index*
			$lb_m/ft\ sec \times 10^4$	*cp*			
Acetic acid	65.493	1.049	7.76	1.155	1584^{50}	6.15	1.37
Acetone	48.98	.787	2.12	0.316	1174	20.7	1.36
Alcohol, ethyl	49.01	.787	7.36	1.095	1144	24.3	1.36
Alcohol, methyl	49.10	.789	3.76	0.56	1103	32.6	1.33
Alcohol, propyl	49.94	.802	12.9	1.92	1205	20.1	1.38
Ammonia (aqua)	51.411	.826	—	—	—	16.9	—
Benzene	54.55	.876	4.04	0.601	1298	2.2	1.50
Bromine	—	—	6.38	0.95	—	3.20	—
Carbon disulfide	78.72	1.265	2.42	0.36	1149	2.64	1.63
Carbon tetrachloride	98.91	1.59	6.11	0.91	924	2.23	1.46
Castor oil	59.69	0.960	—	650	1474	4.7	—
Chloroform	91.44	1.47	3.56	0.53	995	4.8	1.44
Decane	45.34	.728	5.77	0.859	—	2.0	1.41
Dodecane	47.11	—	9.23	1.374	—	—	1.41
Ether	44.54	0.715	1.50	0.223	985	4.3	1.35
Ethylene glycol	68.47	1.100	109	16.2	1644	37.7	1.43
Fluorine refrigerant R–11	92.14	1.480	2.82	0.42	—	2.0	1.37
Fluorine refrigerant R–12	81.84	1.315	—	—	—	2.0	1.29
Fluorine refrigerant R–22	74.53	1.197	—	—	—	2.0	1.26
Glycerine	78.62	1.263	6380	950	1909	40	1.47
Heptane	42.42	.681	2.53	0.376	1138	1.92	1.38
Hexane	40.88	.657	2.00	0.297	1203	—	1.37
Iodine	—	—	—	—	—	11	—
Kerosene	51.2	0.823	11.0	1.64	1320	—	—
Linseed oil	58.0	0.93	222	33.1	—	3.3	—
Mercury	—	13.633	10.3	1.53	1450	—	—
Octane	43.61	.701	3.43	0.51	1171	—	1.40
Phenol	66.94	1.071	54	8.0	1274^{100}	9.8	—
Propane	30.81	.495	0.74	0.11	—	1.27	1.34
Propylene	32.11	.516	0.60	0.09	—	—	1.36
Propylene glycol	60.26	.968	—	42	—	—	1.43
Sea water	64.0	1.03	—	—	1535	—	—
Toluene	53.83	0.865	3.70	0.550	1275^{30}	2.4	1.49
Turpentine	54.2	0.87	9.24	1.375	1240	—	1.47
Water	62.247	1.00	6.0	0.89	1498	78.54[a]	1.33

[a]The dielectric constant of water near the freezing point is 87.8; it decreases with increase in temperature to about 55.6 near the boiling point.

Appendix C. Properties of Solids

TABLE C.1 Properties of Common Solids*

Material	Specific gravity	Specific heat		Thermal conductivity	
		$\frac{Btu}{lbm \cdot deg\ R}$	$\frac{kJ}{kg \cdot K}$	$\frac{Btu}{hr \cdot ft \cdot deg\ F}$	$\frac{W}{m \cdot K}$
Asbestos cement board	1.4	0.2	.837	0.35	0.607
Asbestos millboard	1.0	0.2	.837	0.08	0.14
Asphalt	1.1	0.4	1.67		
Beeswax	0.95	0.82	3.43		
Brick, common	1.75	0.22	.920	0.42	0.71
Brick, hard	2.0	0.24	1.00	0.75	1.3
Chalk	2.0	0.215	.900	0.48	0.84
Charcoal, wood	0.4	0.24	1.00	0.05	0.088
Coal, anthracite	1.5	0.3	1.26		
Coal, bituminous	1.2	0.33	1.38		
Concrete, light	1.4	0.23	.962	0.25	0.42
Concrete, stone	2.2	0.18	.753	1.0	1.7
Corkboard	0.2	0.45	1.88	0.025	0.04
Earth, dry	1.4	0.3	1.26	0.85	1.5
Fiberboard, light	0.24	0.6	2.51	0.035	0.058
Fiber hardboard	1.1	0.5	2.09	0.12	0.2
Firebrick	2.1	0.25	1.05	0.8	1.4
Glass, window	2.5	0.2	.837	0.55	0.96
Gypsum board	0.8	0.26	1.09	0.1	0.17
Hairfelt	0.1	0.5	2.09	0.03	0.050
Ice (32°)	0.9	0.5	2.09	1.25	2.2
Leather, dry	0.9	0.36	1.51	0.09	0.2
Limestone	2.5	0.217	.908	1.1	1.9
Magnesia (85%)	0.25	0.2	.837	0.04	0.071
Marble	2.6	0.21	.879	1.5	2.6
Mica	2.7	0.12	.502	0.4	0.71
Mineral wool blanket	0.1	0.2	.837	0.025	0.04
Paper	0.9	0.33	1.38	0.07	0.1
Paraffin wax	0.9	0.69	2.89	0.15	0.2
Plaster, light	0.7	0.24	1.00	0.15	0.2
Plaster, sand	1.8	0.22	.920	0.42	0.71
Plastics, foamed	0.2	0.3	1.26	0.02	0.03
Plastics, solid	1.2	0.4	1.67	0.11	0.19
Porcelain	2.5	0.22	.920	0.9	1.5
Sandstone	2.3	0.22	.920	1.0	1.7
Sawdust	0.15	0.21	.879	0.05	0.08
Silica aerogel	0.11	0.2	.837	0.015	0.02
Vermiculite	0.13	0.2	.837	0.035	0.058
Wood, balsa	0.16	0.7	2.93	0.03	0.050
Wood, oak	0.7	0.5	2.09	0.10	0.17
Wood, white pine	0.5	0.6	2.51	0.07	0.12
Wool, felt	0.3	0.33	1.38	0.04	0.071
Wool, loose	0.1	0.3	1.26	0.02	0.3

*Compiled from several sources.

TABLE C.2 Density of Various Solids:* Approximate Density of Solids at Ordinary Atmospheric Temperature

Substance	*Grams per cu cm*	*Pounds per cu ft*
Agate	2.5–2.7	156–168
Alabaster		
Carbonate	2.69–2.78	168–173
Sulfate	2.26–2.32	141–145
Albite	2.62–2.65	163–165
Amber	1.06–1.11	66–69
Amphiboles	2.9–3.2	180–200
Anorthite	2.74–2.76	171–172
Asbestos	2.0–2.8	125–175
Asbestos slate	1.8	112
Asphalt	1.1–1.5	69–94
Basalt	2.4–3.1	150–190
Beeswax	0.96–0.97	60–61
Beryl	2.69–2.7	168–169
Biotite	2.7–3.1	170–190
Bone	1.7–2.0	106–125
Brick	1.4–2.2	87–137
Butter	0.86–0.87	53–54
Calamine	4.1–4.5	255–280
Calcspar	2.6–2.8	162–175
Camphor	0.99	62
Caoutchouc	0.92–0.99	57–62
Cardboard	0.69	43
Celluloid	1.4	87
Cement, set	2.7–3.0	170–190
Chalk	1.9–2.8	118–175
Charcoal		
Oak	0.57	35
Pine	0.28–0.44	18–28
Cinnabar	8.12	507
Clay	1.8–2.6	112–162
Coal		
Anthracite	1.4–1.8	87–112
Bituminous	1.2–1.5	75–94
Cocoa butter	0.89–0.91	56–57
Coke	1.0–1.7	62–105
Copal	1.04–1.14	65–71
Cork	0.22–0.26	14–16
Cork linoleum	0.54	34
Corundum	3.9–4.0	245–250
Diamond	3.01–3.52	188–220
Dolomite	2.84	177
Ebonite	1.15	72
Emery	4.0	250
Epidote	3.25–3.50	203–218
Feldspar	2.55–2.75	159–172
Flint	2.63	164
Fluorite	3.18	198
Galena	7.3–7.6	460–470
Gamboge	1.2	75
Garnet	3.15–4.3	197–268
Gas carbon	1.88	117
Gelatin	1.27	79
Glass		
Common	2.4–2.8	150–175
Flint	2.9–5.9	180–370
Glue	1.27	79
Granite	2.64–2.76	165–172
Graphite†	2.30–2.72	144–170
Gum arabic	1.3–1.4	81–87
Gypsum	2.31–2.33	144–145
Hematite	4.9–5.3	306–330
Hornblende	3.0	187
Ice	0.917	57.2
Ivory	1.83–1.92	114–120
Leather, dry	0.86	54
Lime, slaked	1.3–1.4	81–87
Limestone	2.68–2.76	167–171
Linoleum	1.18	74
Magnetite	4.9–5.2	306–324
Malachite	3.7–4.1	231–256
Marble	2.6–2.84	160–177
Meerschaum	0.99–1.28	62–80
Mica	2.6–3.2	165–200
Muscovite	2.76–3.00	172–187
Ochre	3.5	218
Opal	2.2	137
Paper	0.7–1.15	44–72
Paraffin	0.87–0.91	54–57
Peat blocks	0.84	52
Pitch	1.07	67
Porcelain	2.3–2.5	143–156
Porphyry	2.6–2.9	162–181
Pressed wood pulp board	0.19	12
Pyrite	4.95–5.1	309–318
Quartz	2.65	165
Resin	1.07	67
Rock salt	2.18	136
Rubber, hard	1.19	74
Rubber, soft		
Commercial	1.1	69
Pure gum	0.91–0.93	57–58
Sandstone	2.14–2.36	134–147
Serpentine	2.50–2.65	156–165
Silica		
Fused transparent	2.21	138
Translucent	2.07	129
Slag	2.0–3.9	125–240
Slate	2.6–3.3	162–205
Soapstone	2.6–2.8	162–175
Spermaceti	0.95	59
Starch	1.53	95
Sugar	1.59	99
Talc	2.7–2.8	168–174
Tallow		
Beef	0.94	59
Mutton	0.94	59
Tar	1.02	66
Topaz	3.5–3.6	219–223
Tourmaline	3.0–3.2	190–200
Wax, sealing	1.8	112
Wood (seasoned)		
Alder	0.42–0.68	26–42
Apple	0.66–0.84	41–52
Ash	0.65–0.85	40–53
Balsa	0.11–0.14	7–9
Bamboo	0.31–0.40	19–25
Basswood	0.32–0.59	20–37
Beech	0.70–0.90	32–56
Birch	0.51–0.77	32–48
Blue gum	1.00	62
Box	0.95–1.16	59–72
Butternut	0.38	24
Cedar	0.49–0.57	30–35
Cherry	0.70–0.90	43–56
Dogwood	0.76	47
Ebony	1.11–1.33	69–83
Elm	0.54–0.60	34–37
Hickory	0.60–0.93	37–58
Holly	0.76	47
Juniper	0.56	35
Larch	0.50–0.56	31–35
Lignum vitae	1.17–1.33	73–83
Locust	0.67–0.71	42–44
Logwood	0.91	57
Mahogany		
Honduras	0.66	41
Spanish	0.85	53
Maple	0.62–0.75	39–47
Oak	0.60–0.90	37–56
Pear	0.61–0.73	38–45
Pine		
Pitch	0.83–0.85	52–53
White	0.35–0.50	22–31
Yellow	0.37–0.60	23–37
Plum	0.66–0.78	41–49
Poplar	0.35–0.5	22–31
Satinwood	0.95	59
Spruce	0.48–0.70	30–44
Sycamore	0.40–0.60	24–37
Teak		
Indian	0.66–0.88	41–55
African	0.98	61
Walnut	0.64–0.70	40–43
Water gum	1.00	62
Willow	0.40–0.60	24–37

†Some values reported as low as 1.6

*Based largely on: "Smithsonian Physical Tables", 9th rev. ed., W.E. Forsythe, Ed., The Smithsonian Institution, 1956, p. 292.

Note: In the case of substances with voids, such as paper or leather, the bulk density is indicated rather than the density of the solid portion. For density in kg/m^3, multiply values in g/cm^3 by 1,000.

TABLE C.3 Specific Stiffness of Metals, Alloys, and Certain Non-Metallics*

Specific stiffness is usually expressed as the modulus of elasticity (in tension) per unit weight-density, i.e., E/ρ, in units of pounds and inches. While the stiffness of similar alloys varies considerably, there are definite ranges and groups to be recognized. Since the specific stiffness of steel is about 100 million, the values in the following table are also approximately the percentage stiffness, referred to steel.

Material	*Specific stiffness, millions*
Beryllium	650
Silicon carbide	600
Alumina ceramics	400
Mica	350
Titanium carbide cermet	250
Alumina cermet	200
Molybdenum and alloys; silica glass	130
Titanium and alloys; cobalt superalloys; soda-lime glass	110
Carbon and low-alloy steel; wrought iron	105
Stainless steel; nodular cast iron; magnesium and alloys; aluminum and alloys	100
Nickel and alloys; malleable iron	95
Iron silicon alloys (cast); iridium; vanadium	90
Monel alloys; tungsten	80
Gray cast iron; columbium alloys	70
Aluminum bronze; beryllium copper	65
Nickel silver; cupronickel; zirconium	55
Yellow brass; nickel cast iron; bronze; Muntz metal; antimony	50
Copper; red brass; tantalum	45
Silver and alloys; pewter; platinum and alloys; white gold	30
Tin; thorium	25
Gold	20
Tin-lead alloy	10
Lead	5

*Compiled from several sources.

TABLE C.4 Thermal Properties of Pure Metals—Metric Units

Metal	AT ATMOSPHERIC PRESSURE								LIQUID METAL			
				At 100°K		At 25°C (77°F)				Vapor pressure		
	Melting point, °C	Boiling point, °C	Latent heat of fusion, cal/g**	Thermal conductivity, watts/cm °C	Specific heat, cal/g °C**	Specific heat, cal/g °C**	Coeff. of linear expansion, ($\times 10^6$) $(°C)^{-1}$	Thermal conductivity, watts/cm °C	Specific heat (liquid) at 2000°K, cal/g °C**	10^{-3} atm	10^{-6} atm	10^{-9} atm
										Boiling point temperatures, °K		
Aluminum	660.	2441.	95	3.00*	.115	0.215	25	2.37	.26	1,782	1,333	1,063
Antimony	630.	1440.	38.5	—	.040	.050	9	.185	.062	1,007	741	612
Beryllium	1285.	2475.	324.	—	.049	.436	12	2.18	.78	1,793	1,347	1,085
Bismuth	271.4	1660.	12.4	—	.026	.030	13	.084	.036	1,155	851	677
Cadmium	321.	767.	13.2	1.03	.047	.055	30	.93	.063	655	486	388
Chromium	1860.	2670.	79	1.58	.046	.110	6	.91	.224	1,992	1,530	1,247
Cobalt	1495.	2925.	66	—	.057	.10	12	.69	.164	2,167	1,652	1,345
Copper	1084.	2575.	49	4.83*	.061	.092	16.6	3.98	.118	1,862	1,391	1,120
Gold	1063.	2800.	15	3.45*	.026	.031	14.2	3.15	.0355	2,023	1,510	1,211
Iridium	2450.	4390.	33	—	.022	.031	6	1.47	.0434	3,253	2,515	2,062
Iron	1536.	2870.	65	1.32*	.052	.108	12	.803	.197	2,093	1,594	1,297
Lead	327.5	1750.	5.5	0.396	.028	.031	29	.346	.033	1,230	889	698
Magnesium	650.	1090.	88.0	1.69	.016	.243	25	1.59	.32	857	638	509
Manganese	1244.	2060.	64	—	.064	.114	22	—	.20	1,495	1,131	913
Mercury	−38.86	356.55	2.7	—	.029	.033	—	.0839	—	393	287	227
Molybdenum	2620.	4651.	69	1.79	.033	.060	5	1.4	.089	3,344	2,558	2,079
Nickel	1453.	2800.	71	1.58	.055	.106	13	.899	.175	2,156	1,646	1,343
Niobium (Columbium)	2470.	4740.	68	0.552	.045	.064	7	.52	.083	3,523	2,721	2,232
Osmium	3025.	4225.	34	—	—	.031	5	.61	.039	—	—	—
Platinum	1770.	3825.	24	0.79*	.024	.032	9	.73	.043	2,817	2,155	1,757
Plutonium	640.	3230.	3	—	.019	.032	54	.08	.041	2,200	1,596	1,252
Potassium	63.3	760.	14.5	—	.150	.180	83	.99	—	606	430	335
Rhodium	1965.	3700.	50	—	—	.058	8	1.50	.092	—	—	—
Selenium	217.	700.	16	—	—	.077	37	.005	—	—	—	—
Silicon	1411.	3280.	430	—	.062	.17	3	.835	.217	2,340	1,749	1,427
Silver	961.	2212.	26.5	4.50*	.045	.057	19	4.27	.068	1,582	1,179	952
Sodium	97.83	884.	27	—	.234	.293	70	1.34	—	701	504	394
Tantalum	2980.	5365.	41	0.592	.026	.034	6.5	.54	.040	3,959	3,052	2,495
Thorium	1750.	4800.	17	—	.024	.03	12	.41	.047	3,251	2,407	1,919
Tin	232.	2600.	14.1	0.85	.039	.054	20	.64	.058	1,857	1,366	1,080
Titanium	1670.	3290.	100	0.312	.072	.125	8.5	.2	.188	2,405	1,827	1,484
Tungsten	3400.	5550.	46	2.35*	.021	.032	4.5	1.78	.040	4,139	3,228	2,656
Uranium	1132.	4140.	12	—	.022	.028	13.4	.25	.048	2,861	2,128	1,699
Vanadium	1900.	3400.	98	—	.061	.116	8	.60	.207	2,525	1,948	1,591
Zinc	419.5	910.	27	1.32	.063	.093	35	1.15	—	752	559	449

* Temperatures of maximum thermal conductivity (conductivity values in watts/cm °C): Aluminum 13°K, cond. = 71.5; copper 10°K, cond. = 196; gold 10°K, cond. = 28.2; iron 20°K, cond. = 9.97; platinum 8°K, cond. = 12.9; silver 7°K, cond. = 193; tungsten 8°K, cond. = 85.3.

** To convert to SI units note that 1 cal = 4.186 J.

TABLE C.5 Mechanical Properties of Metals and Alloys:* Typical Composition, Properties, and Uses of Common Materials

For MN/m^2 multiply strength in thousands of psi by 6.895.

No.	Material	Nominal composition	Form and condition	Typical mechanical properties: Yield strength (0.2% offset), 1000 lb/sq in.	Tensile strength, 1000 lb/sq in.	Elongation, in 2 in., %	Hardness, Brinell	Comments
			FERROUS ALLOYS					
	Ferrous alloys comprise the largest volume of metal alloys used in engineering. The actual range of mechanical properties in any particular grade of alloy steel depends on the particular history and heat treatment. The steels listed in this table are intended to give some idea of the range of properties readily obtainable. Many hundreds of steels are available. Cost is frequently an important criterion in the choice of material; in general the greater the percentage of alloying elements present in the alloy, the greater will be the cost.							
	IRON							
1	Ingot iron (Included for comparison)	Fe 99.9	Hot-rolled Annealed	29 19	45 38	26 45	90 67	
	PLAIN CARBON STEELS							
2	AISI-SAE 1020	C 0.20 Mn 0.45 Si 0.25 Fe bal.	Hot-rolled Hardened (water-quenched, 1000°F-tempered)	30 62	55 90	25 25	111 179	Bolts, crankshafts, gears, connecting rods; easily weldable
3	AISI 1025	C 0.25 Fe bal. Mn 0.45	Bar stock Hot-rolled Cold-drawn	 32 54	 58 64	 25 15	 116 126	
4	AISI-SAE 1035	C 0.35 Mn 0.75	Hot-rolled Cold-rolled	39 67	72 80	18 12	143 163	Medium-strength, engineering steel
5	AISI-SAE 1045	C 0.45 Fe bal. Mn 0.75	Bar stock Annealed Hot-rolled Cold-drawn	 73 45 77	 80 82 91	 12 16 12	 170 163 179	
6	AISI-SAE 1078	C 0.78 Fe bal. Mn 0.45	Bar stock Hot-rolled; spheroidized Annealed	 55 72	 100 94	 12 10	 207 192	
7	AISI-SAE 1095	C 0.95 Fe bal. Mn 0.40						
8	AISI-SAE 1120	C 0.2 Mn 0.8 S 0.1	Cold-drawn	58	69	—	137	Free-cutting, leaded, resulphurized steel; high-speed, automatic machining
	ALLOY STEELS							
9	ASTM A202/56	C 0.17 Mn 1.2 Cr 0.5 Si 0.75	Stress-relieved	45	75	18	—	Low alloy; boilers, pressure vessels

TABLE C.5 (continued) Mechanical Properties of Metals and Alloys:* Typical Composition, Properties, and Uses of Common Materials

No.	Material	Nominal composition	Form and condition	Typical mechanical properties				Comments
				Yield strength (0.2% offset), 1000 lb/sq in.	Tensile strength, 1000 lb/sq in.	Elongation, in 2 in., %	Hardness, Brinell	
10	AISI 4140	C 0.40 Si 0.3 Cr 1.0 Mo 0.2 Mn 0.9	Fully-tempered Optimum properties	95 132	108 150	22 18	240 —	High strength; gears, shafts
11	12% Manganese steel	12% Mn C	Tempered 600°F Rolled and heat-treated stock	200 44	220 160	10 40	— 170	Machine tool parts; wear, abrasion-resistant
12	VASCO 300	Ni 18.5 Ti 0.6 Co 9.0 C 0.03 Mo 4.8	Solution treatment 1500°F; aged 900°F	110	150	18	—	Very high strength, maraging, good machining properties in annealed state
13	T1 (AISI)	W 18.0 V 1.0 Cr 4.0 C 0.7	Quenched; tempered				R(c)	High speed tool steel, cutting tools, punches, etc.
14	M2 (AISI)	W 6.5 Mo 5.0 Cr 4.0 C 0.85 V 2.0	Quenched; tempered				65–66	M-grade, cheaper, tougher
15	Stainless steel type 304	Ni 9.0 C 0.08 Cr 19.0 max	Annealed; cold-rolled	35 to 160	85 to 185	60 8	160 to 400	General purpose, weldable; nonmagnetic austenitic steel
16	Stainless steel type 316	Cr 18.0 C 0.10 Ni 11.0 max Mo 2.5 Fe bal.	Annealed	30 to 120	90 to 150	50 8	165 275	For severe corrosive media, under stress; nonmagnetic austenitic steel
17	Stainless steel type 431	Cr 16.0 Si 1.0 Ni 2.0 C 0.20 Mn 1.0 Fe bal.	Annealed Heat-treated	85 150	120 195	25 20	250 400	Heat-treated stainless steel, with good mechanical strength; magnetic
18	Stainless steel 17–4 PH	Cr 17.0 Co 0.35 Ni 4.0 C 0.07 Cu 4.0 Fe bal.	Annealed	110	150	10	363	Precipitation hardening; heat-resisting type; retains strength up to approx. 600°F

TABLE C.5 (continued) Mechanical Properties of Metals and Alloys:* Typical Composition, Properties, and Uses of Common Materials

No.	Material	Nominal composition	Form and condition	Typical mechanical properties: Yield strength (0.2% offset), 1000 lb/sq in.	Tensile strength, 1000 lb/sq in.	Elongation, in 2 in., %	Hardness, Brinell	Comments
			CAST IRONS AND CAST STEELS					
	These alloys are used where large and/or intricate-shaped articles are required or where over-all dimensional tolerances are not critical. Thus the article can be produced with the fabrication and machining costs held to a minimum. Except for a few heat-treatable cast steels, this class of alloys does not demonstrate high-strength qualities.							
	CAST IRONS							
19	Cast gray iron ASTM A48-48, Class 25	C 3.4 Si 1.8 Mn 0.5	Cast (as cast)	—	25 min	0.5 max	180	Engine blocks, fly-wheels, gears, machine-tool bases
20	White	C 3.4 Si 0.7 Mn 0.6	Cast	—	25	0	450	
21	Malleable iron ASTM A47	C 2.5 Si 1.0 Mn 0.55 max	Cast (annealed)	33	52	12	130	Automotives, axle bearings, track wheels, crankshafts
22	Ductile or nodular iron (Mg-containing) ASTM A339 ASTM A395	C 3.4 P 0.1 max Mn 0.40 Ni 1% Mg 0.06 Si 2.5 Fe bal.	Cast Cast (as cast) Cast (quenched, tempered)	53 68 108	70 90 135	18 7 5	170 235 310	Heavy-duty machines, gears, cams, crankshafts
23	Ni-hard type 2	C 2.7 Si 0.6 Mn 0.5 Ni 4.5 Cr 2.0 Fe bal.	Sand-cast Chill-cast (tempered)	— —	55 75	— —	550 625	Strength, with heat- and corrosion-resistance
24	Ni-resist type 2	C 3.0 Si 2.0 Mn 1.0 Ni 20.0 Cr 2.5 Fe bal.	Cast (as cast)	—	27	2	140	
	CAST STEELS							
25	ASTM A27-62 (60-30)	C 0.3 Mn 0.6 Si 0.8 Ni 0.5 Cr 0.4 Mo 0.2		30	60	24	—	Low alloy, medium strength, general application
26	ASTM A148-60 (105-85)			85	105	17	—	High strength; structural application

TABLE C.5 (continued) Mechanical Properties of Metals and Alloys:* Typical Composition, Properties, and Uses of Common Materials

No.	Material	Nominal composition	Form and condition	Typical mechanical properties: Yield strength (0.2% offset), 1000 lb/sq in.	Tensile strength, 1000 lb/sq in.	Elongation, in 2 in., %	Hardness, Brinell	Comments
27	Cast 12 Cr alloy (CA-15)	C 0.15 max; Mn 1.00 max; Si 1.50 max; Cr 11.5–14; Ni 1.00 max; Fe bal.	Air-cooled from 1800°F; tempered at 600°F	150	200	7	390	Stainless, corrosion-resistant to mildly corrosive alkalis and acids
			Air-cooled from 1800°F; tempered at 1400°F	75	100	30	185	
28	Cast 29-9 alloy (CE-30) ASTM A296 63T	C 0.30 max; Mn 1.50 max; Si 2.00 max; Cr 26–30; Ni 8–11; Fe bal.	As cast	60	95	15	170	Greater corrosion resistance, especially for oxidizing condition
29	Cast 28-7 alloy (HD) ASTM A297-63T	C 0.50 max; Mn 1.50 max; Si 2.00 max; Cr 26–30; Ni 4–7; Fe bal.	As cast	48	85	16	190	Heat-resistant

SUPER ALLOYS

The advent of engineering applications requiring high temperature and high strength, as in jet engines and rocket motors, has lead to the development of a range of alloys collectively called super alloys. These alloys require excellent resistance to oxidation together with strength at high temperatures, typically 1800°F in existing engines. These alloys are continually being modified to develop better specific properties, and therefore entries in this group of alloys should be considered "fluid". Both wrought and casting-type alloys are represented. As the high temperature properties of cast materials improve, these alloys become more attractive, since great dimensional precision is now attainable in investment castings.

No.	Material	Nominal composition	Form and condition	Yield strength (0.2% offset), 1000 lb/sq in.	Tensile strength, 1000 lb/sq in.	Elongation, in 2 in., %	Hardness, Brinell	Comments
	NICKEL BASE							
30	Hastelloy X	Co 1.5 max; Fe 18.5; Cr 22.0; Mo 9.0; W 0.6; C 0.15 max (wrought); C 0.20 max (cast); Ni bal.	Wrought sheet	52	113.2	43	194	
			Mill-annealed	—	67	17	172	
			As investment cast	46.5	—	—	—	
31	Hastelloy C	Cr 16.0; Fe 6.0; W 4.0; C 0.15 max; Mo 17.0; Ni bal.	Sand-cast (annealed)	50	78	5	199	
			Rolled (annealed)	71	130	45	204	
			Investment cast	50	80	10	215	

TABLE C.5 (continued) Mechanical Properties of Metals and Alloys:* Typical Composition, Properties, and Uses of Common Materials

No.	Material	Nominal composition	Form and condition	Typical mechanical properties				Comments
				Yield strength (0.2% offset), 1000 lb/sq in.	Tensile strength, 1000 lb/sq in.	Elongation, in 2 in., %	Hardness, Brinell	
	NICKEL BASE (Cont.)							
32	Inconel 713C	Ni (+Co) bal. Cr 13.0 Cb 2.0 Mo 4.5 Ti 0.6 Al 6.0	Investment cast	102	120	6	—	
33	In 100	C 18.0 Cr 10.0 Mo 3.0 Ti 4.7 Al 55.0 Co 15.0 V 1.0	Cast					
34	Taz 8	C 125.0 Cr 6.0 Mo 4.0 Al 6.0 W 4.0 Zr 1.0 Ta 8.0 V 2.5	Cast					
35	Nimonic 90	Ni (+Co) 57.00 C 0.05 Fe 0.45 Mn 0.50 Si 0.20 S 0.007 Cr 20.55 Cu 0.05 Ti 2.60 Al 1.65 Co 16.90	Annealed; wrought	90	155	—	260	General elevated temperature applications
36	Inconel X	Ni (+Co) 72.85 C 0.04 Fe 6.80 Mn 0.65 Si 0.30 S 0.007 Cr 15.0 Cu 0.05 Ti 2.50 Al 0.75 Cb (+Ta) 0.85	Annealed Annealed; age-hardened	50 115	115 175	50 25	150 300	
37	Waspaloy	C 0.08 Cr 19.5 Mo 4.3 Ti 3.0 Co 13.5	Cold-rolled	270	275	8	Rc 51	
38	Rene 41	C 0.09 Cr 19.0 Mo 10.0 Ti 3.1 Al 1.5 Co 11.0	Wrought	100	145	—	—	

TABLE C.5 (continued) Mechanical Properties of Metals and Alloys:* Typical Composition, Properties, and Uses of Common Materials

No.	Material	Nominal composition	Form and condition	Typical mechanical properties: Yield strength (0.2% offset), 1000 lb/sq in.	Tensile strength, 1000 lb/sq in.	Elongation, in 2 in., %	Hardness, Brinell	Comments
39	Udimet 700	C 0.08 Cr 15.0 Mo 5.0 Ti 3.5 Al 4.3 Co 18.5	Cold-rolled	280	285	6	Rc 53	
40	T.D. Nickel	Ni 97.5 ThO_2 2.4	Extended and cold-worked	85	100	13	—	High temperature; jet engine parts
	COBALT BASE							
41	Haynes Stellite alloy 25 (L605)	C 0.15 max Cr 20.0 W 15.0 Ni 10.0 Co bal. Mn 1.5	Wrought sheet; mill annealed	63	140	60	244	Wrought products
42	Haynes Stellite alloy 21 AMS 5385 (cast)	C 0.25 Mo 5.5 Ni 2.5 Co bal. Cr 28.5	As investment cast	82	103	8	313 max	For castings
	ALUMINUM ALLOYS							
	Although the strength of aluminum alloys is in general less than that attainable in ferrous alloys or copper-base alloys, their major advantage lies in their high strength-to-weight ratio due to the low density of aluminum. Aluminum alloys have good corrosion resistance for most applications except in alkaline solutions.							
43	3003 ASTM B221	Cu 0.12 Al bal. Mn 1.2	Annealed-O Cold-rolled-H14 Cold-rolled-H18	6 21 27	16 22 29	40 16 10	28 40 55	Good formability, weldable, medium strength; chemical equipment
44	2017 ASTM B221	Mn 0.5 Mg 0.5 Cu 4.0 Al bal.	Annealed-O Heat-treated-T4	10 40	26 62	22 22	45 105	High strength; structural parts, aircraft, heavy forgings
45	2024 ASTM B211	Cu 4.5 Mg 1.5 Mn 0.6 Al bal.	Heat-treated-T4	47	68	19	120	
46	5052 ASTM B211	Cr 0.25 Al bal. Mg 2.5	Annealed-O Cold-rolled and stabilized-H34	13 31	28 38	30 14	47 68	Medium strength, good fatigue properties; street-light standards
47	ASTM B209		Cold-rolled and stabilized-H38	37	42	8	77	

TABLE C.5 (continued) Mechanical Properties of Metals and Alloys:* Typical Composition, Properties, and Uses of Common Materials

No.	Material	Nominal composition	Form and condition	Typical mechanical properties: Yield strength (0.2% offset), 1000 lb/sq in.	Tensile strength, 1000 lb/sq in.	Elongation, in 2 in., %	Hardness, Brinell	Comments
48	7075 ASTM B211	Cu 1.6 Mg 2.5 Cr 0.3 Al bal. Zn 5.6	Annealed-O Heat-treated and artificially aged-T6	15 73	33 83	17 11	60 150	High strength, good corrosion resistance
49	380 ASTM SC84B	Si 9.0 Al bal. Cu 3.5	Die-cast	24	48	3	—	General purpose die casting
50	195 ASTM C4A	Si 0.8 Al bal. Cu 4.5	Sand-cast; heat-treated-T4 Sand-cast; heat-treated and artificially aged-T6	16 24	32 36	8.5 5	60 75	Structural elements, aircraft, and machines
51	214 ASTM G4A	Mg 3.8 Al bal.	Sand-cast-F	12	25	9	50	Chemical equipment, marine hardware, architectural
52	220 ASTM G10A	Mg 10.0 Al bal.	Sand-cast; heat-treated-T4	26	48	16	75	Strength with shock resistance; aircraft

COPPER ALLOYS

Because of their corrosion resistance and the fact that copper alloys have been used for many thousands of years, the number of copper alloys available is second only to the ferrous alloys. In general copper alloys do not have the high-strength qualities of the ferrous alloys, while their density is comparable. The cost per strength-weight ratio is high; however, they have the advantage of ease of joining by soldering, which is not shared by other metals that have reasonable corrosion resistance.

53	Copper ASTM B152 ASTM B124, B133 ASTM B1, B2, B3	Cu 99.9 plus	Annealed Cold-drawn Cold-rolled	10 40 40	32 45 46	45 15 5	42 90 100	Bus-bars, switches, architectural, roofing, screens
54	Gilding metal ASTM B36	Cu 95.0 Zn 5.0	Cold-rolled	50	56	5	114	Coinage, ammunition
55	Cartridge 70–30 brass ASTM B14 ASTM B19 ASTM B36 ASTM B134 ASTM B135	Cu 70.0 Zn 30.0	Cold-rolled	63	76	8	155	Good cold-working properties; radiator covers, hardware, electrical
56	Phosphor bronze 10% ASTM B103 ASTM B139 ASTM B159	Cu 90.0 Sn 10.0 P 0.25	Spring temper	—	122	4	241	Good spring qualities, high-fatigue strength

TABLE C.5 (continued) Mechanical Properties of Metals and Alloys:* Typical Composition, Properties, and Uses of Common Materials

No.	Material	Nominal composition	Form and condition	Typical mechanical properties				Comments
				Yield strength (0.2% offset), 1000 lb/sq in.	Tensile strength, 1000 lb/sq in.	Elongation, in 2 in., %	Hardness, Brinell	
57	Yellow brass (high brass) ASTM B36 ASTM B134 ASTM B135	Cu 65.0 Zn 35.0	Annealed Cold-drawn Cold-rolled (HT)	18 55 60	48 70 74	60 15 10	55 115 180	Good corrosion resistance; plumbing, architectural
58	Manganese bronze ASTM B138	Cu 58.5 Zn 39.2 Fe 1.0 Sn 1.0 Mn 0.3	Annealed Cold-drawn	30 50	60 80	30 20	95 180	Forgings
59	Naval brass ASTM B21	Cu 60.0 Zn 39.25 Sn 0.75	Annealed Cold-drawn	22 40	56 65	40 35	90 150	Condensor tubing; high resistance to salt-water corrosion
60	Muntz metal ASTM B111	Cu 60.0 Zn 40.0	Annealed	20	54	45	80	Condensor tubes; valve stress
61	Aluminum bronze ASTM B169, alloy A ASTM B124 ASTM B150	Cu 92.0 Al 8.0	Annealed Hard	25 65	70 105	60 7	80 210	
62	Beryllium copper 25 ASTM B194 ASTM B197 ASTM B196	Be 1.9 Cu bal. Co or Ni 0.25	Annealed, solution-treated Cold-rolled Cold-rolled	32 104 70	70 110 190	45 5 3	B60 (Rockwell) B81 C40	Bellows, fuse clips, electrical relay parts, valves, pumps
63	Free-cutting brass	Cu 62.0 Zn 35.5 Pb 2.5	Cold-drawn	44	70	18	B80 (Rockwell)	Screws, nuts, gears, keys
64	Nickel silver 18% Alloy A (wrought) ASTM B122, No. 2	Cu 65.0 Zn 17.0 Ni 18.0	Annealed Cold-rolled Cold-drawn wire	25 70 —	58 85 105	40 4 —	70 170 —	Hardware, optical goods, camera parts
65	Nickel silver 13% (cast) 10A ASTM B149, No. 10A	Ni 12.5 Pb 9.0 Sn 2.0 Cu bal. Zn 20.0	Cast	18	35	15	55	Ornamental castings, plumbing; good machining qualities
66	Cupronickel 10% ASTM B111 ASTM B171	Cu 88.35 Ni 10.0 Fe 1.25 Mn 0.4	Annealed Cold-drawn tube	22 57	44 60	45 15	— —	Condensor, salt-water piping

TABLE C.5 (continued) Mechanical Properties of Metals and Alloys:* Typical Composition, Properties, and Uses of Common Materials

No.	Material	Nominal composition	Form and condition	Typical mechanical properties				Comments
				Yield strength (0.2% offset), 1000 lb/sq in.	Tensile strength, 1000 lb/sq in.	Elongation, in 2 in., %	Hardness, Brinell	
67	Cupronickel	Cu 70.0 Ni 30.0	Wrought					Heat-exchanger process equipment, valves
68	Red brass (cast) ASTM B30, No. 4A	Cu 85.0 Zn 5.0 Pb 5.0 Sn 5.0	As-cast	17	35	25	60	
69	Silicon bronze ASTM B30, alloy 12A	Si 4.0 Fe 2.0 Zn 4.0 Al 1.0 Mn 1.0	Castings					Cheaper substitute for tin bronze
70	Tin bronze ASTM B30, alloy 1B	Sn 8% Zn 4.0	Castings					Bearings, high-pressure bushings, pump impellers
71	Navy bronze		Cast					
	TIN AND LEAD-BASE ALLOYS							
	Major uses for these alloys are as "white"-metal bearing alloys, extruded cable sheathing, and solders. Tin forms the basis of pewter used for culinary applications.							
72	Lead-base Babbitt ASTM B23, alloy 19	Pb 85.0 Sn 5.0 Sb 10.0 As 0.6 Cu 0.5	Chill cast	—	10	5	19	Bearings, light loads and low speeds
73	Arsenical-lead Babbitt ASTM B23, alloy 15	Pb 83.0 Sn 1.0 Sb 16.0 As 1.1 Cu 0.6	Chill cast	—	10.3	2	20	Bearings, high loads and speeds, diesel engines, steel mills
74	Chemical lead	Pb 99.9 Cu 0.06 Bi 0.005 max	Rolled 95%	1.9	2.5	50	5	
75	Antimonial lead (hard lead)	Pb 94.0 Sb 6.0	Chill cast Rolled 95%	— —	6.8 4.1	22 47	(500 kg) 9	Good corrosion resistance and strength
76	Calcium lead	Pb 99.9 Ca 0.025 Cu 0.10	Extruded and aged	—	4.5	25	—	Cable sheathing, creep-resistant pipe
77	Tin Babbitt alloy ASTM B23-61, grade 1	Sb 4.5 Sn bal. Cu 4.5	Chill cast	—	9.3	2	17	General bearings and die casting
78	Tin die-casting alloy ASTM B102-52	Sb 13.0 Sn bal. Cu 5.0	Die-cast	—	10	1	29	Die-casting alloy

TABLE C.5 (continued) Mechanical Properties of Metals and Alloys:* Typical Composition, Properties, and Uses of Common Materials

No.	Material	Nominal composition	Form and condition	Typical mechanical properties: Yield strength (0.2% offset), 1000 lb/sq in.	Tensile strength, 1000 lb/sq in.	Elongation, in 2 in., %	Hardness, Brinell	Comments
79	Pewter	Sn 91.0 Sb 7.0 Cu 2.0	Rolled sheet, annealed	—	8.6	40	9.5	Ornamental and household items
80	Solder 50–50	Sn 50.0 Pb 50.0	Cast	4.8	6.1	60	14	General-purpose solder
81	Solder	Sn 20.0 Pb 80.0	Cast	3.6	5.8	16	11	Coating and joining, filling seams on automobile bodies
	MAGNESIUM ALLOYS							
	Because of their low density these alloys are attractive for use where weight is at a premium. The major drawback to the use of these alloys is their ability to ignite in air (this can be a problem in machining); they are also costly. Magnesium alloys are used in both the wrought and die-cast forms, the latter being the most frequently used form.							
82	Magnesium alloy AZ31B	Zn 1.0 Mn 0.20 min Al 3.0 Mg bal.	Rolled-plate (strain-hardened, then partially annealed)	24	37	18	—	Structural applications of medium strength
			Rolled-sheet (strain-hardened, then partially annealed)	32	42	15	73	
			Annealed	22	37	21	56	
			Extruded	28	38	14	—	
83	Magnesium alloy AZ80A	Zn 0.5 Mn 0.15 min Al 8.5 Mg bal.	Extruded	36	49	11	60	General extruded and forged products
			Extruded (age-hardened)	39	53	6	82	
			Forged (age-hardened)	34	50	6	72	
84	Magnesium alloy AZ92A	Zn 2.0 Mn 0.10 min Al 9.0 Mg bal.	Sand-cast (as cast)	14	24	6	50	
			Sand-cast (solution heat-treated)	14	40	12	55	
			Sand-cast (solution heat-treated and aged)	19	40	5	83	
			Sand-cast (age-hardened)	16	30	18	—	
			Sand-cast and tempered	22	40	3	81	Pressure-tight sand and permanent mold castings; high UTS and good yield strength
85	Magnesium alloy ZK60A	Zn 5.7 Mg bal. Zr 0.55	Extruded	43	52	12	82	

TABLE C.5 (continued) Mechanical Properties of Metals and Alloys:* Typical Composition, Properties, and Uses of Common Materials

No.	Material	Nominal composition	Form and condition	Typical mechanical properties: Yield strength (0.2% offset), 1000 lb/sq in.	Tensile strength, 1000 lb/sq in.	Elongation, in 2 in., %	Hardness, Brinell	Comments
86	Magnesium alloy AZ91A and AZ91B	Zn 0.6 Mn 0.13 min Al 9.0 Mg bal.	Die-cast (as cast)	22	33	3	67	General die-casting applications
			BERYLLIUM					
87	Beryllium			27	33	1–3	—	Windows, X-ray tubes
			Hot-pressed	38	51			
			Cross-rolled	40	60	10–40	—	Moderator- and reflector-cladding nuclear reactors; heat-shield and structural-member missiles
				60	90			
			NICKEL ALLOYS					
	Nickel and its alloys are expensive and used mainly either for their high-corrosion resistance in many environments or for high-temperature and strength applications. (See Super Alloys, above.)							
88	Nickel (cast)	Ni 95.6 Cu 0.5 Fe 0.5 Mn 0.8 Si 1.5 C 0.8	As cast	25	57	22	110	Good corrosion-resistance applications
89	K Monel	Ni (+Co) 65.25 C 0.15 Mn 0.60 Fe 1.00 S 0.005 Si 0.15 Cu 29.60 Al 2.75 Ti 0.45	Annealed	45	100	40	155	High strength and corrosion resistance; aircraft parts, valve stems, pumps
			Annealed, age-hardened	100	155	25	270	
			Spring	140	150	5	300	
			Spring, age-hardened	160	185	10	335	
90	A nickel ASTM B160 ASTM B161 ASTM B162	Ni (+Co) 99.40 C 0.06 Fe 0.15 Mn 0.25 Si 0.05 S 0.005 Cu 0.05	Annealed	20	70	40	100	Chemical industry for resistance to strong alkalis, plating nickel
			Hot-rolled	25	75	40	110	
			Cold-drawn	70	95	25	170	
			Cold-rolled	95	105	5	210	
91	Duranickel	Ni (+Co) 93.90 C 0.15 Fe 0.15 Mn 0.25 Si 0.55 S 0.005 Al 4.50 Cu 0.05 Ti 0.45	Annealed	45	100	40	160	High strength and corrosion resistance; pump rods, shafts, springs
			Annealed, age-hardened	125	170	25	330	
			Spring	—	175	5	320	
			Spring, age-hardened	—	205	10	370	

TABLE C.5 (continued) Mechanical Properties of Metals and Alloys:* Typical Composition, Properties, and Uses of Common Materials

No.	Material	Nominal composition	Form and condition	Typical mechanical properties: Yield strength (0.2% offset), 1000 lb/sq in.	Tensile strength, 1000 lb/sq in.	Elongation, in 2 in., %	Hardness, Brinell	Comments
92	Cupronickel 55–45 (Constantan)	Cu 55.0 Ni 45.0	Annealed Cold-drawn Cold-rolled	30 50 65	60 65 85	45 30 20	— — —	Electrical-resistance wire; low temperature coefficient, high resistivity
93	Nichrome	Ni 80.0 Cr 20.0						Heating elements for furnaces
94	"S" Monel	Ni 60.0 Cu 29.0 Fe 2.50 max Mn 1.5 max Si 4.0 Al 0.5 max	Sand-casting	80–115	110–145	2	270–350	High-strength casting alloy; good bearing properties for valve seats
	TITANIUM ALLOYS							
	The main application for these alloys is in the aerospace industry. Because of the low density and high strength of titanium alloys, they present excellent strength-to-weight ratios.							
95	Commercial titanium ASTM B265–58T	Ti 99.4	Annealed at 1100 to 1350°F (593 to 732°C)	70	80	20	—	Moderate strength, excellent fabricability; chemical industry pipes
96	Titanium alloy ASTM B265–58T–5 Ti–6 Al–4V		Water-quenched from 1750°F (954°C); aged at 1000°F (538°C) for 2 hr	160	170	13	—	High-temperature strength needed in gas-turbine compressor blades
97	Titanium alloy Ti–4 Al–4Mn		Water-quenched from 1450°F (788°C); aged at 900°F (482°C) for 8 hr	170	185	13	—	Aircraft forgings and compressor parts
98	Ti–Mn alloy ASTM B265–58T–7	Fe 0.5 Ti bal. Mn 7.0–8.0	Sheet	140	150	18	—	Good formability, moderate high-temperature strength; aircraft skin
	ZINC ALLOYS							
	A major use for these alloys is for low-cost die-cast products, such as household fixtures, automotive parts, and trim.							
99	Zinc ASTM B69	Cd 0.35 Zn bal. Pb 0.08	Hot-rolled	—	19.5	65	38	Battery cans, grommets, lithographer's sheet

TABLE C.5 (continued) Mechanical Properties of Metals and Alloys:* Typical Composition, Properties, and Uses of Common Materials

No.	*Material*	*Nominal composition*	*Form and condition*	*Typical mechanical properties*				*Comments*
				Yield strength (0.2% offset), 1000 lb/sq in.	*Tensile strength, 1000 lb/sq in.*	*Elongation, in 2 in., %*	*Hardness, Brinell*	
100	Zilloy-15	Cu 1.00 Zn bal. Mg 0.010	Hot-rolled Cold-rolled	— —	29 36	20 25	61 80	Corrugated roofs, articles with maximum stiffness
101	Zilloy-40	Cu 1.00 Zn bal.	Hot-rolled Cold-rolled	— —	24 31	50 40	52 60	Weatherstrip, spun articles
102	Zamac-5 ASTM 25	Zn (99.99% pure remainder) Al 3.5–4.3 Cu 0.75–1.25 Mg 0.03–0.08	Die-cast	—	47.6	7	91	Die casting for automobile parts, padlocks; used also for die material
		ZIRCONIUM ALLOYS						
	These alloys have good corrosion resistance but are easily oxidized at elevated temperatures in air. The major application is for use in nuclear reactors.							
103	Zirconium, commercial	O_2 0.07 C 0.15 Hf 1.90 Zr bal.	Annealed	40	65	27	B80 (Rockwell)	
104	Zircaloy 2	Hf 0.02 Ni 0.05 Fe 0.15 Other 0.25 Sn 1.46 Zr bal.	Annealed	50	75	22	B90 (Rockwell)	Nuclear power-reactor cores at elevated temperatures

*Compiled from various sources.

TABLE C.6 Miscellaneous Properties of Metals and Alloys

Part a. Pure Metals

At Room Temperature

Common name	*PROPERTIES (TYPICAL ONLY)*						
	Thermal conductivity, Btu/hr ft °F	*Specific gravity*	*Coeff. of linear expansion, μ in./ in. °F*	*Electrical resistivity, microhm-cm*	*Poisson's ratio*	*Modulus of elasticity, millions of psi*	*Approximate melting point, °F*
Aluminum	137	2.70	14	2.655	0.33	10.0	1220
Antimony	10.7	6.69	5	41.8		11.3	1170
Beryllium	126	1.85	6.7	4.0	0.024–.030	42	2345
Bismuth	4.9	9.75	7.2	115		4.6	521
Cadmium	54	8.65	17	7.4		8	610
Chromium	52	7.2	3.3	13		36	3380
Cobalt	40	8.9	6.7	9		30	2723
Copper	230	8.96	9.2	1.673	0.36	17	1983
Gold	182	19.32	7.9	2.35	0.42	10.8	1945
Iridium	85.0	22.42	3.3	5.3		75	4440
Iron	46.4	7.87	6.7	9.7		28.5	2797
Lead	20.0	11.35	16	20.6	0.40–.45	2.0	621
Magnesium	91.9	1.74	14	4.45	0.35	6.4	1200
Manganese		7.21–7.44	12	185		23	2271
Mercury	4.85	13.546		98.4			−38
Molybdenum	81	10.22	3.0	5.2	0.32	40	4750
Nickel	52.0	8.90	7.4	6.85	0.31	31	2647
Niobium (Columbium)	30	8.57	3.9	13		15	4473
Osmium	35	22.57	2.8	9		80	5477
Platinum	42	21.45	5	10.5	0.39	21.3	3220
Plutonium	4.6	19.84	30	141.4	0.15–.21	14	1180
Potassium	57.8	0.86	46	7.01			146
Rhodium	86.7	12.41	4.4	4.6		42	3569
Selenium	0.3	4.8	21	12.0		8.4	423
Silicon	48.3	2.33	2.8	1×10^5		16	2572
Silver	247	10.50	11	1.59	0.37	10.5	1760
Sodium	77.5	0.97	39	4.2			208
Tantalum	31	16.6	3.6	12.4	0.35	27	5400
Thorium	24	11.7	6.7	18	0.27	8.5	3180
Tin	37	7.31	11	11.0	0.33	6	450
Titanium	12	4.54	4.7	43	0.3	16	3040
Tungsten	103	19.3	2.5	5.65	0.28	50	6150
Uranium	14	18.8	7.4	30	0.21	24	2070
Vanadium	35	6.1	4.4	25		19	3450
Zinc	66.5	7	19	5.92	0.25	12	787

TABLE C.6 Miscellaneous Properties of Metals and Alloys

Part b. Commercial Metals and Alloys

CLASSIFICATION AND DESIGNATION		*PROPERTIES (TYPICAL ONLY)*					
Material No. (from Table 1-57)	*Common name and classification*	*Thermal conductivity, Btu/hr ft °F*	*Specific gravity*	*Coeff. of linear expansion, μ in./ in. °F*	*Electrical resistivity, microhm-cm*	*Modulus of elasticity, millions of psi*	*Approximate melting point, °F*
1	Ingot iron (included for comparison)	42.	7.86	6.8	9.	30	2800
2	Plain carbon steel AISI–SAE 1020	30.	7.86	6.7	10.	30	2760
15	Stainless steel type 304	10.	8.02	9.6	72.	28	2600
19	Cast gray iron ASTM A48–48, Class 25	26.	7.2	6.7	67.	13	2150
21	Malleable iron ASTM A47	—	7.32	6.6	30.	25	2250
22	Ductile cast iron ASTM A339, A395	19	7.2	7.5	60.	25	2100
24	Ni-resist cast iron, type 2	23	7.3	9.6	170.	15.6	2250
29	Cast 28–7 alloy (HD) ASTM A297–63T	1.5	7.6	9.2	41.	27	2700
31	Hastelloy C	5	3.94	6.3	139.	30	2350
36	Inconel X, annealed	9	8.25	6.7	122.	31	2550
41	Haynes Stellite alloy 25 (L605)	5.5	9.15	7.61	88.	34	2500
43	Aluminum alloy 3003, rolled ASTM B221	90	2.73	12.9	4.	10	1200
44	Aluminum alloy 2017, annealed ASTM B221	95	2.8	12.7	4.	10.5	1185
49	Aluminum alloy 380 ASTM SC84B	56	2.7	11.6	7.5	10.3	1050
53	Copper ASTM B152, B124, B133, B1, B2, B3	225	8.91	9.3	1.7	17	1980
57	Yellow brass (high brass) ASTM B36, B134, B135	69	8.47	10.5	7.	15	1710
61	Aluminum bronze ASTM B169, alloy A; ASTM B124, B150	41	7.8	9.2	12.	17	1900
62	Beryllium copper 25 ASTM B194	7	8.25	9.3	—	19	1700
64	Nickel silver 18% alloy A (wrought) ASTM B122, No. 2	19	8.8	9.0	29.	18	2030
67	Cupronickel 30%	17	8.95	8.5	35.	22	2240
68	Red brass (cast) ASTM B30, No. 4A	42	8.7	10.	11.	13	1825
74	Chemical lead	20	11.35	16.4	21.	2	621
75	Antimonial lead (hard lead)	17	10.9	15.1	23.	3	554
80	Solder 50–50	26	8.89	13.1	15.	—	420
82	Magnesium alloy AZ31B	45	1.77	14.5	9.	6.5	1160
89	K Monel	11	8.47	7.4	58.	26	2430
90	Nickel ASTM B160, B161, B162	35	8.89	6.6	10.	30	2625
92	Cupronickel 55–45 (Constantan)	13	8.9	8.1	49.	24	2300
95	Commercial titanium	10	5.	4.9	80.	16.5	3300
99	Zinc ASTM B69	62	7.14	18	6.	—	785
103	Zirconium, commercial	10	6.5	2.9	41.	12	3350

*Compiled from several sources.

TABLE C.7 Composition and Melting Points of Binary Eutectic Alloys:* Binary Alloys and Solid Solutions of Metallic Components

This table represents most of the common binary combinations of metals. For many pairs no eutectic exists; for many others the information is uncertain or unavailable. In a fair number of cases, there is complete mutual solubility in all proportions; hence, there is a smooth temperature vs. composition curve, with no point of inflection from the melting point of one constituent to that of the other. For purposes of comparison, all values must be considered approximate in view of the experimental difficulties and the many sources of data.

Those pairs for which the liquidus curve exhibits more than one cusp are designated by a superscript *a*. In a few cases the cusp selected for this table does not represent the lowest possible melting point for the binary mixture.

Constituents		*Composition*		*Melting point*		*Constituents*		*Composition*		*Melting point*	
A	*B*	*Mol % B*	*Weight % B*	*K*	*deg F*	*A*	*B*	*Mol % B*	*Weight % B*	*K*	*deg F*
Ag	Al	57	25	835	1 044	Au	Bi	86.8	85	514	466
Ag	As	24	18	813	1 004	Au	Cd	70	57.1	773	932
Ag	Ca[a]	37	18	820	1 017	Au	Ce[a]	86	81	793	968
Ag	Ce[a]	80	84	798	977	Au	Ge	27	12	629	673
Ag	Cu	40	28	1 050	1 431	Au	La[a]	83	78	834	1 042
Ag	Ge	25	18	924	1 204	Au	Mg	93	62	848	1 067
Ag	La[a]	72	77	791	964	Au	Mn[a]	32	12	1 233	1 760
Ag	Li	99	89	418	293	Au	Na	17	2.3	1 149	1 609
Ag	Mg[a]	83	52	745	882	Au	Pb	84	85	488	419
Ag	Pb	95.3	97.5	577	579	Au	Sb	34.8	24.8	633	680
Ag	Pd	25.9	25.6	924	1 204	Au	Si	18.6	3.15	636	685
Ag	Sb	41	44	758	905	Au	Sn[a]	29.3	19.9	551	532
Ag	Si	10.5	2.96	1 110	1 539	Au	Te	88	83	689	781
Ag	Sr[a]	77	73	709	817	Au	Tl	72	73	404	268
Ag	Te	65	69	623	662	Au	U	14	16	1 128	1 571
Ag	Th	7.6	15	1 167	1 641	B	Hf	13	71	2 130	3 375
Ag	Zr	97	93	1 100	1 521	B	Ni	57	88	1 263	1 814
Al	Au[a]	59,5	90.0	842	1 056	B	Ti	7	25	1 700	2 601
Al	Ca[a]	65	73	818	1 013	B	Zr	88	98	1 920	2 997
Al	Cd	81	90	1 650	2 511	Ba	Mg	97	87	891	1 144
Al	Ce	69	92	928	1 211	Be	Ni	33	76	1 468	2 183
Al	Cu[a]	17.3	33.0	821	1 018	Be	Pu	97	99	910	1 179
Al	Fe	32	49.34	1 426	2 107	Be	Si	33	61	1 363	1 994
Al	Ge	29	55	700	801	Be	Ti	75	94	1 300	2 061
Al	In	5	18	910	1 179	Be	Y	61	94	1 390	2 043
Al	Mg	70	67.0	710	819	Be	Zr	65	95	1 250	1 791
Al	Ni[a]	76	87	1 658	2 525	Bi	Ca	88	58.5	1 059	1 447
Al	Pt[a]	57	90	1 533	2 300	Bi	Cd	56	40	420	297
Al	Si	13	13	850	1 071	Bi	In[a]	78	66	340	153
Al	Th	80	97	1 510	2 259	Bi	K	50	16	615	648
Al	Zn	88.7	95.0	655	720	Bi	Mg	85	40	820	1 017
As	Co	75	70	1 189	1 681	Bi	Na	22	3.0	500	441
As	Cu[a]	81.6	78.0	958	1 265	Bi	Pb	44	44	397	255
As	Fe	75	69	1 103	2 017	Bi	Sn	57	43	415	288
As	In	13	18	1 004	1 348	Bi	Te	90	84	686	775
As	Mn	57	49	1 143	1 598	Bi	Tl[a]	53	52	465	378
As	Ni[a]	63	57	1 077	1 479	C	Cr	87	96	1 775	2 736
As	Sb	80	87	878	1 121	C	Hf	35	88	3 450	5 751
As	Sn[a]	40	51	852	1 074	C	Mo	17	45	2 480	4 005
As	Zn[a]	20	18	996	1 333	C	Nb	40	84	3 580	5 985

*Compiled from several sources.

TABLE C.7 (continued) Composition and Melting Points of Binary Eutectic Alloys:* Binary Alloys and Solid Solutions of Metallic Components

Constituents		*Composition*		*Melting point*		*Constituents*		*Composition*		*Melting point*	
A	*B*	*Mol % B*	*Weight % B*	*K*	*deg F*	*A*	*B*	*Mol % B*	*Weight % B*	*K*	*deg F*
C	Ti	36	69	3 050	5 031	Gd	Ni[a]	32	15	943	1 238
C	V	84	96	1 900	2 961	Ge	Mg	38	17	953	1 256
C	W	59	96	2 980	4 905	Ge	Mn[a]	48	41	970	1 287
Ca	Cu	51	62	833	1 040	Hf	Ta	24	24	1 300	1 881
Ca	Mg[a]	32	22	718	833	In	Ni	30	17.97	1 143	1 598
Ca	Na	22	14	983	1 310	In	Sb	68	69	780	945
Ca	Ni	16	22	878	1 121	In	Sn	47	48	390	243
Ca	Sn	19	41	1 010	1 359	Ir	Mo	68	52	2 350	3 771
Cd	Cu	52	38	810	999	Ir	Nb	55	23	2 110	3 339
Cd	In	74	74	400	261	Ir	W	22	12	2 590	4 203
Cd	Pb	71	82	540	513	K	Na	32	21.67	260	−8.6
Cd	Pu	40	59	1 170	1 647	K	Rb	70	84	307	93
Cd	Sb	7.4	8	563	554	K	Sb[a]	68	84	680	765
Cd	Sn	68	69	450	351	K	Tl	84	96	440	333
Cd	Tl	73	83	475	396	La	Mg[a]	38	9.7	970	1 287
Cd	Zn	27	18	540	513	La	Pb[a]	11	15	1 049	1 429
Ce	Cu[a]	28	15	688	779	La	Sn[a]	10	9	993	1 328
Ce	Ru	33	26	923	1 202	La	Tl	16	22	913	1 184
Co	Gd	65	83	913	1 184	Mg	Ni	11	22.98	780	945
Co	Mo	27	38	1 610	2 439	Mg	Pr	4.9	23	858	1 085
Co	Nb	15	22	1 500	2 241	Mg	Pu	15	63	815	1 008
Co	Si[a]	71	54	1 486	2 215	Mg	Sb[a]	86	97	855	1 080
Co	Sn	21	35	1 380	2 025	Mg	Si	53	57	1 223	1 742
Co	Ti[a]	22	19	1 430	2 115	Mg	Sr[a]	70	89	699	799
Co	V	41	38	1 521	2 278	Mg	Th	7	42	855	1 080
Cr	Mo	14	23	1 973	3 092	Mg	Zn	30	53	615	648
Cr	Ni	46	47	1 610	2 439	Mn	Ni	40	42	1 300	1 881
Cr	Ta	13	34	1 950	3 051	Mn	Pd	26	41	1 398	2 057
Cr	Ti	86	85	950	1 251	Mn	Sb	82	91	843	1 058
Cr	V	33	32	2 050	3 231	Mn	Ti[a]	9	7.9	1 460	2 169
Cs	K	50	23	235	−36	Mn	U[a]	75	93	988	1 319
Cs	Na	20.9	4.37	241	−26	Mn	Y[a]	65	75	1 163	1 634
Cs	Rb	50	39	282	48	Mo	Nb	66	65	2 570	4 167
Cu	Ge	34	37	913	1 184	Mo	Ni	64	52	1 590	2 403
Cu	Mg[a]	85.5	69.3	758	905	Mo	Os	21	34	2 650	4 311
Cu	Mn	37	34	1 143	1 598	Mo	Pd	54	57	2 020	3 177
Cu	Pb	15	36	1 230	1 755	Mo	Re	48	64	2 780	4 545
Cu	Pr[a]	69	83	745	882	Mo	Ru	41	42	2 200	3 501
Cu	Sb[a]	63	76	800	981	Mo	Si[a]	17	5.7	2 350	3 771
Cu	Si	30	16	1 075	1 476	Na	Rb	82.1	94.5	269	25
Cu	Te	69	82	617	207	Na	Sb	60	89	678	761
Cu	Ti[a]	27	22	1 133	1 580	Na	Sn	37	75	718	833
Cu	Tl	14.5	35.3	1 357	1 983	Na	Te	55	87	592	606
Cu	U	8.2	25	1 213	1 724	Nb	Ni	58	47	1 450	2 151
Cu	Zr	9.4	13	1 253	1 796	Nb	Pt	54	71	1 970	3 087
Fe	Gd	69	86	1 123	1 562	Nb	Rh	45	31	1 770	2 727
Fe	Mo	21	31	1 725	2 646	Nb	Ru[a]	64	49	2 050	3 231
Fe	Nb	12	18.49	1 643	2 498	Nb	Zr	77	77	2 010	3 159
Fe	Sb	88	94.10	1 021	1 378	Ni	Sb	22	36.90	1 375	2 016
Fe	Si[a]	35	21	1 475	2 196	Ni	Sn	19	32.16	1 403	2 066
Fe	Sn	31	49	1 400	2 061	Ni	Th[a]	35	68	1 303	1 886
Fe	Y	65	75	1 173	1 652	Ni	Ti[a]	39	34	1 390	2 043
Fe	Zr[a]	11	17	1 600	2 421	Ni	V	52	48	1 473	2 192
Ga	Mg[a]	80	58	698	797	Ni	W	20.7	45	1 773	2 732
Ga	Ni	70	66	1 477	2 199	Ni	Zn	69	71	1 148	1 607

TABLE C.7 (continued) Composition and Melting Points of Binary Eutectic Alloys:* Binary Alloys and Solid Solutions of Metallic Components

Constituents		*Composition*		*Melting point*		*Constituents*		*Composition*		*Melting point*	
A	*B*	*Mol % B*	*Weight % B*	*K*	*deg F*	*A*	*B*	*Mol % B*	*Weight % B*	*K*	*deg F*
Pb	Pr	40	31	1 315	1 908	Si	Th[a]	88	98	1 710	2 619
Pb	Pt	5.3	5.0	563	554	Si	Ti[a]	86	91	1 600	2 421
Pb	Sb	18	11	520	477	Si	Zr[a]	9	24	1 570	2 367
Pb	Sn	73	61	460	369	Sn	Te	84	85	678	761
Pb	Te	85	78	680	765	Sn	Tl	31	44	440	333
Pb	Ti	92	74	998	1 337	Sn	Zn	16	9.5	465	378
Pd	Sb	89	90	868	1 103	Te	Tl	30	41	483	410
Pt	Sn	40	29	1 345	1 962	Th	Ti	40	12	1 463	2 174
Pu	Zn	73	42	1 100	1 521	Th	Zn[a]	49	21	1 220	1 737
Re	W	26	26	3 100	5 121	Ti	U	17	51	933	1 220
Sb	Tl	70	80	468	383	Ti	Y	6.8	12	1 593	2 408
Sb	Zn	33	21	780	945	Ti	Zr	50	66	790	963
Sb	Zr	82	77	1 700	2 601	U	Zr	70	47	879	1 123
Se	Sn	39	49	913	1 184						
Se	Tl	26	48	424	304						

REFERENCES

"Selected Values of Thermodynamic Properties of Metals and Alloys", R. Hultgren, R.L. Orr, P.D. Anderson, K.K. Kelley, John Wiley & Sons, Inc., 1963; a supplement to this publication has been issued periodically by the University of California, 1964–1971.

"Constitution of Binary Alloys", 2nd ed., M. Hansen, McGraw-Hill Book Company, 1958.

"Metals Reference Book", 4th ed., C.J. Smithells, Vol. 2, Butterworth & Co., London, 1967.

"Handbook of Binary Metallic Systems", 2 volumes; translated from Russian, Israel Program for Scientific Translations, Jerusalem. Available from Clearinghouse for Federal Scientific and Technical Information, Springfield, Virginia 22151.

See also *Trans. AIME*, *J. Inst. Metals*, and *Z. Metallkunde*, by indexes.

TABLE C.8 Melting Points of Mixtures of Metals**

Melting Points, °C

Metals	*Percentage of metal in second column*										
	0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
Pb. Sn.*	326	295	276	262	240	220	190	185	200	216	232
Bi.	322	290	...	...	179	145	126	168	205	...	268
Te.	322	710	790	880	917	760	600	480	410	425	446
Ag.	328	460	545	590	620	650	705	775	840	905	959
Na.	...	360	420	400	370	330	290	250	200	130	96
Cu.	326	870	920	925	945	950	955	985	1005	1020	1084
Sb.	326	250	275	330	395	440	490	525	560	600	632
Al. Sb.	650	750	840	925	945	950	970	1000	1040	1010	632
Cu.	650	630	600	560	540	580	610	755	930	1055	1084
Au.	655	675	740	800	855	915	970	1025	1055	675	1062
Ag.	650	625	615	600	590	580	575	570	650	750	954
Zn.	654	640	620	600	580	560	530	510	475	425	419
Fe.	653	860	1015	1110	1145	1145	1220	1315	1425	1500	1515
Sn.	650	645	635	625	620	605	590	570	560	540	232
Sb. Bi.	632	610	590	575	555	540	520	470	405	330	268
Ag.	630	595	570	545	520	500	505	545	680	850	959
Sn.	622	600	570	525	480	430	395	350	310	255	232
Zn.	632	555	510	540	570	565	540	525	510	470	419

Metals	*Percentage of metal in second column*										
	0%	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
Ni. Sn.*	1455	1380	1290	1200	1235	1290	1305	1230	1060	800	232
Na. Bi.	96	425	520	590	645	690	720	730	715	570	268
Cd.	96	125	185	245	285	325	330	340	360	390	322
Cd. Ag.	322	420	520	610	700	760	805	850	895	940	954
Tl.	321	300	285	270	262	258	245	230	210	235	302
Zn.	322	280	270	295	313	327	340	355	370	390	419
Au. Cu.	1063	910	890	895	905	925	975	1000	1025	1060	1084
Ag.	1064	1062	1061	1058	1054	1049	1039	1025	1006	982	963
Pt.	1075	1125	1190	1250	1320	1380	1455	1530	1610	1685	1775
K. Na.	62	17.5	−10	−3.5	5	11	26	41	58	77	97.5
Hg.	...	...	...	...	...	90	110	135	162	265	...
Tl.	62.5	133	165	188	205	215	220	240	280	305	301
Cu. Ni.	1080	1180	1240	1290	1320	1355	1380	1410	1430	1440	1455
Ag.	1082	1035	990	945	910	870	830	788	814	875	960
Sn.	1084	1005	890	755	725	680	630	580	530	440	232
Zn.	1084	1040	995	930	900	880	820	780	700	580	419
Ag. Zn.	959	850	755	705	690	660	630	610	570	505	419
Sn.	959	870	750	630	550	495	450	420	375	300	232
Na. Hg.	96.5	90	80	70	60	45	22	55	95	215	...

*The data in this table are compiled from various sources—hence the variations in the melting point of the metals as shown in this column.

**Based largely on: "Smithsonian Physical Tables", 9th rev. ed., W.E. Forsythe, Ed., The Smithsonian Institution, 1956.

TABLE C.9 Trade Names, Composition, and Manufacturers of Various Plastics

Trade name	*Composition*	*Manufacturer*
Abson	Acrylonitrile-butadiene, ABS polymers	B. F. Goodrich Chemical Co.
Alathon	Polyethylene resins	E. I. du Pont de Nemours & Co., Inc.
Alkor	Furane resin cement	Atlas Minerals & Chemicals Div., The Electric Storage Battery Co.
Amres	Phenolics, urea, and melamine resins	American Marietta Co., Pacific Resins & Chemicals, Inc.
Araldite	Epoxy resins	CIBA Products Co., Div. CIBA Corp.
Atlac	Polyester resins	Atlas Chemical Industries, Inc.
Bakelite	Acrylics, epoxies, phenolics, polyethylenes, copolymers	Union Carbide Corp., Chemicals and Plastics Div.
Bavick-11	Methyl methacrylate and methylstyrene copolymer	J. T. Baker Chemical Co.
Boltaflex	Supported and unsupported flexible vinyl sheeting	The General Tire & Rubber Co.
Boltaron	Rigid polyvinyl chloride sheet	The General Tire & Rubber Co., Chemical & Plastics Div.
Butacite	Polyvinyl butyral resins	E. I. du Pont de Nemours & Co., Inc.
Conolite	Polyester resins and laminates	Shellmar-Betner, Div. Continental Can Co. Woodall Industries, Inc., Conolite Div.
Corvel	Epoxies, vinyls	The Polymer Corp., Export-Polypenco Div.
	Fusion-bond finishes	
Cumar	Paracoumarone-indene resins	Allied Chemical Corp., Plastics Div.
Cycolac	ABS polymers, acrylonitrile-butadiene-styrene copolymers	Marbon Chemical Div., Borg-Warner Corporation
Dacovin	Polyvinyl chlorides	Diamond Shamrock Corp.
Dapon	Diallyl phthalate resins	FMC Corp., Organic Div.
Delrin	Acetal resin and pipe	E. I. du Pont de Nemours & Co., Inc.
Dylan	Polyethylene	Sinclair-Koppers Co.
Dylene	Polystyrene	Sinclair-Koppers Co.
Dylite	Expandable polystyrene	Sinclair-Koppers Co.
Epi-Rez	Epoxy resins	Celanese Coatings Co., Celanese Resin Div.
Epolene	Low molecular-weight polyethylene resins	Eastman Chemical Products, Inc., Sub. Eastman Kodak Company
Epoxical	Epoxy resins	United States Gypsum Co.
Epon	Epoxy resins and curing agents	The Shell Chemical Company, Plastics and Resins Div.
Escon	Polypropylene resins	Enjay Chemical Co., Div. Humble Oil & Refining Company
Estane	Polyurethane materials	B. F. Goodrich Chemical Company
Fluorogreen	Teflon with glass and ceramic fibers, fluorocarbons	John L. Dore Co.
Fluororay	Ceramic-filled fluorocarbons	Raybestos-Manhattan, Inc., Plastic Products Div.
Formica	Melamines	Formica Corp. of American Cyanamid
Forticel	Cellulose propionate sheet films, molding powders	Celanese Plastics Co.
Fortiflex	Polyethylene resins	Celanese Plastics Co.
Fosta-Tuf-Flex	Polystyrene, high-impact	Foster-Grant, Inc.
Furnane	Furanes	Atlas Minerals & Chemicals Div., The Electric Storage Battery Co.
GenEpoxy	Epoxy resins for adhesives, coatings, etc.	General Mills, Inc., Chemical Div.
Genetron	Fluorinated hydrocarbons, monomers, and polymers	Allied Chemical Corp., General Chemical Div.
Geon	Polyvinyl chloride materials	B. F. Goodrich Chemical Co.
Grex	High-density polyethylenes	Allied Chemical Corp., Plastics Div.
Halon	Fluorohalocarbon resins	Allied Chemical Corp.
Hetron	Fire-retardant polyester resin	Hooker Chemical Corp., Durez Plastics Div.
Isothane	Polyurethane foam, ester, and ether	Bernel Foam Products Co., Inc.
Kel-F	Chlorotrifluoroethylene, molding resins, and dispersions	3M Company
Kralac	High-styrene resins, styrene-butadiene copolymers	Uniroyal Chemical, Div. of Uniroyal Inc.
Kralastic	ABS polymers, copolymers	Uniroyal Chemical, Div. of Uniroyal Inc.
Kynar	Polyvinylidene fluoride	Pennsalt Chemical Corp.
Lexan	Polycarbonate resin, film, and sheet	General Electric Company, Plastics Dept.
Lucite	Acrylic resin and syrup	E. I. du Pont de Nemours & Co., Inc.
Lustran	ABS polymers	Monsanto Co.
Lustrex	Styrene molding and extrusion resins	Monsanto Co.
Lytron	Styrene molding and extrusion resins	Monsanto Co.
Madurit	Melamine resins and compounds	Cassella Farbwerke Mainkur, A.G.
Maraglas	Epoxy-casting resin	The Marblette Corporation, Div. of Allied Products
Marlex	Polyethylenes, polypropylenes, copolymers	Phillips Petroleum Co.
Marvinol	Vinyl chloride resins and compounds	Uniroyal Chemical, Div. of Uniroyal Inc.
Merlon	Polycarbonate resins	Mobay Chemical Co.
Micarta	Melamines, phenolics, polyesters	Westinghouse Electric Co., Industrial Micarta Div.
Microthene	Polyethylenes, polyolefins	U.S. Industrial Chemicals Co.
Multrathane	Urethane elastomers	Mobay Chemical Company
Nopcofoam	Polyurethane plastics	Nopco Chemical Co., Plastics Div.
Novodur	Polyacrylonitrile-butadiene-styrene	Farbenfabriken Bayer, A. G.
Opalon	Vinyl chloride resins and compounds	Monsanto Co.
Paraplex	Polyester resins, acrylic-modified polyester resins	Rohm & Haas Company

TABLE C.9 (continued) Trade Names, Composition, and Manufacturers of Various Plastics

Trade name	*Composition*	*Manufacturer*
Permelite	Melamines	Melamine Plastics, Inc., Div. of Fiberlite Corp.
Petrothene	Polyethylene resins, polypropylene resins	U.S. Industrial Chemicals Co.
Piccoflex	Styrene-copolymer resins	Pennsylvania Industrial Chemical Corp.
Piccolastic	Styrene-polymer resins	Pennsylvania Industrial Chemical Corp.
Plaskon	Nylons, melamines, phenolics, polyesters	Allied Chemical Corp.
Pleogen	Alkyds, polyesters, copolymers	Mol-Rez Div., American Petrochemical Corp.
Plexiglas	Acrylics	Rohm & Haas Company
Pliovic	Polyvinyl chlorides	The Goodyear Tire & Rubber Co., Chemical Div.
Plyophen	Phenolic resins	Reichhold Chemicals, Inc.
Poly-Eth	Polyethylene resins	Gulf Oil Corp., U.S. Div. of Gulf Oil Corp.
Polylite	Polyester resins	Reichhold Chemicals, Inc.
Polypenco	Acrylics, chlorinated polyethers, fluoro-carbons, nylons, polycarbonates	Polymer Corp.
Resimene	Urea and melamine resins	Monsanto Co.
Resinox	Phenolic resins and compounds	Monsanto Co.
Rhonite	Urea resins	Rohm & Haas Company
Roylar	Polyurethanes	Uniroyal Chemical, Div. of Uniroyal Inc.
Ryertex	Laminated phenolics and rigid polyvinyl chloride extrusions	Joseph T. Ryerson & Son, Inc., Industrial Plastics and Bearings Sales Div.
Super Dylan	Polyethylene	Sinclair-Koppers Co.
Supreme	Polyethylenes	Johns-Manville Company
Sylplast	Urea-formaldehyde compounds	FMC Corp., Organic Chemicals Div.
Teflon	Fluorocarbon resins	E. I. du Pont de Nemours & Co., Inc.
Tenite	Cellulose acetate, cellulose-acetate-polyethylene, poly-propylenes, urethane elastomers, copolymers	Eastman Chemical Products, Inc., Sub. Eastman Kodak Co.
Tetran	Fluorocarbons	Pennsalt Chemicals Corp.
Texin	Urethane elastomers	Mobay Chemical Company
Thioment	Polyisoprenes	Atlas Minerals & Chemicals Div., The Electric Storage Battery Co.
Ultrapas	Melamine resins	Dynamit Nobel, A. G.
Ultrathene	Ethylene-vinyl acetates	U.S. Industrial Chemicals Co.
Ultron	Polyvinyl chlorides	Monsanto Co.
Vibrathane	Urethane elastomers	Uniroyal Chemical, Div. of Uniroyal Inc.
Vibrin	Polyester resins	Uniroyal Chemical, Div. of Uniroyal Inc.
Vitel	Polyesters	The Goodyear Tire & Rubber Co., Chemical Div.
Viton	Synthetic rubbers	E. I. du Pont de Nemours & Co., Inc.
Vitroplast	Polyester cements	Atlas Minerals & Chemicals Div., The Electric Storage Battery Co.
Vyron	Polyvinyl chlorides	Industrial Vinyls, Inc.

TABLE C.10 Properties of Commercial Nylon Resins*

Property	*Type 6/6*	*Type 6*	*Type 6/10*	*Type 11*	*Glass-reinforced Type 6/6, 40%*	*MoS_2-filled, 2½%*	*Direct polymerized, castable*
Mechanical							
Tensile strength, psi	11,800	11,800	8209	8500	30,000	10,000 to 14,000	11,000 to 14,000
Elongation, %	60	200	240	120	1.9	5 to 150	10 to 50
Tensile yield stress, psi	11,800	11,800	8500		30,000		
Flexural modulus, psi	410,000	395,000	280,000	151,000	1,800,000	450,000	
Tensile modulus, psi	420,000	380,000	280,000	178,000		450,000 to 600,000	350,000 to 450,000
Hardness, Rockwell	118R	119R	111R	55A	75E–80E	110R–125R	112R–120R
Impact strength, tensile, ft-lb/sq in.	76		160			50–180	80–100
Impact strength, Izod, ft-lb/in. of notch	0.9	1.0	1.2	3.3	3.7**	0.6	0.9
Deformation under load, 2000 psi, 122°F, %	1.4	1.8	4.2	2.02†	0.4§	0.5 to 2.5	0.5 to 1
Thermal							
Heat-deflection temp, °F							
At 66 psi	360	365	300	154	509	400 to 490	400 to 425
At 264 psi	150	152	135	118	502	200 to 470	300 to 425
Coefficient of thermal expansion, per °F	4.5×10^{-5}	4.6×10^{-5}	5×10^{-5}	10×10^{-5}	0.9×10^{-5}	3.5×10^{-5}	5.0×10^{-5}
Coefficient of thermal conductivity, Btu in./hr ft² °F	1.7	1.7	1.5				
Specific heat	0.3–0.5	0.4	0.3–0.5	0.58			
Brittleness temp, °F	−112		−166				
Electrical							
Dielectric strength, short time, v/mil	385	420	470	425	480	300 to 400	500 to 600‡
Dielectric constant							
At 60 hz	4.0	3.8	3.9		4.45		3.7
At 10^3 hz	3.9	3.7	3.6	3.3	4.40		3.7
At 10^6 hz	3.6	3.4	3.5		4.10		3.7
Power factor							
At 60 hz	0.014	0.010	0.04	0.03	0.009		0.02
At 10^3 hz	0.02	0.016	0.04	0.03	0.011		0.02
At 10^6 hz	0.04	0.020	0.03	0.02	0.018		0.02
Volume resistivity, ohm-cm	10^{14} to 10^{15}	3×10^{15}	10^{14} to 10^{15}	2×10^{13}	2.6×10^{15}	2.5×10^{13}	
General							
Water absorption, 24 hr, %	1.5	1.6	0.4	0.4	0.6	0.5 to 1.4	0.9
Specific gravity	1.13 to 1.15	1.13	1.07 to 1.09	1.04	1.52	1.14 to 1.18	1.15 to 1.17
Melting point, °F	482 to 500	420 to 435	405 to 430	367	480 to 490	496 ± 9	430 ± 10
Flammability	self ext	self ext	self ext	self ext	self ext	self ext	self ext
Chemical resistance to							
Strong acids	Poor	Poor	Poor	Poor	Poor	Poor	Poor
Strong bases	Good	Good	Good	Fair	Good	Good	Good
Hydrocarbons	Excellent	Excellent	Excellent	Good	Excellent	Excellent	Excellent
Chlorinated hydrocarbons	Good	Good	Good	Fair	Good	Good	Good
Aromatic alcohols	Good	Good	Good	Good	Good	Good	Good
Aliphatic alcohols	Good	Good	Good	Fair	Good	Good	Good

Notes:

Most nylon resins listed in this table are used for injection molding, and test values are determined from standard injection-molded specimens. In these cases a single typical value is listed. Exceptions are MoS_2-filled nylon and direct-polymerized (castable) nylon, which are sold principally in semifinished stock shapes. Ranges of values listed are based on tests on various forms and sizes produced under varying processing conditions.

Because single values apply only to standard molded specimens, and properties vary in finished parts of different sizes and forms produced by various processes, these values should be used for comparison and preliminary design considerations only. For final design purposes the manufacturer should be consulted for test experience with the form being considered. Listed values should not be used for specification purposes.

†2000 psi, 73°F.

‡0.040-in. thick.

**$\frac{1}{2} \times \frac{1}{2}$-in. bar.

§4000 psi, 122°F.

*From: "Nylons", D.D. Carswell, *Machine Design*, 40(29):62, Dec. 12, 1968.

For Conversion factors see Table C.10.

TABLE C.11 Properties of Silicate Glasses*

Most of the commercially produced glass is for windows, bottles, and inexpensive containers; it is a soda-lime-silica glass of fairly uniform composition, similar to glass No. 0080 in the table below and in Table 1-93. The following tables on glasses (Tables 1-92 through 1-103) deal largely with that one-tenth of the glass output for which special properties are required. All data are subject to normal manufacturing variations.

Silica glass is inherently high in viscosity and melting point. These are reduced by fluxes such as Na_2O, K_2O, and B_2O_3. Soda and potash glasses have a high expansion coefficient (column 7), while that of fused silica is very low. Because the borosilicate glasses are intermediate, and their thermal shock resistance is high (e.g., Corning Code 7740 glass), they are widely used for laboratory and kitchen glassware. Aluminosilicate glasses are hard, heat-resisting, and of high chemical durability. Glass hardness (indentation) correlates closely with the elastic modulus (column 14). Lead oxide is also used as flux, with a result of reduced softening point and high refractive index: hence its uses for optical glass and art glass.

Sealing of glass with metal calls for close control of the coefficient of expansion (column 7 and Figure 1-99).

EXPLANATION OF COLUMNS:

Column 5:

B—blown ware	P—pressed ware	S—plate glass
M—multiform	R—rolled sheet	T—tubing and rod
U—panels	LC—large castings	

Column 6:

[2]Since weathering is determined primarily by clouding, which changes transmission, a rating for the opal glasses is omitted.

[3]These borosilicate glasses may rate differently if subjected to excessive heat treatment.

Column 8:

Normal service: No breakage from excessive thermal shock is assumed.

Extreme limits: Glass will be very vulnerable to thermal shock. Recommendations in this range are based on mechanical stability considerations only. Tests should be made before adopting final designs. These data are approximate only.

Column 9:

Based on plunging sample into cold water after oven heating. Resistance of 100°C means no breakage if heated to 110°C and plunged into water at 10°C. Tempered samples have over twice the resistance of annealed glass. These data are approximate only.

Column 10:

[4]These data are estimated.

Resistance in °C is the temperature differential between the two surfaces of a tube or a constrained plate that will cause a tensile stress of 1000 psi on the cooler surface.

Column 11:

Viscosity is given in poises. At the strain point the stresses are significantly reduced in a matter of hours, while at the annealing point there is adequate stress reduction in minutes.

Column 12:

Data show relative resistance to sandblasting.

Column 15:

Data at 25°C are extrapolated from high temperature readings and are approximate only.

*From: "Properties of Selected Commercial Glasses", Publication B-83, Corning Glass Works.

TABLE C.11 (continued) Properties of Silicate Glasses*

1	2	3	4	5	6			7		8				9		
					Corrosion Resistance			Thermal Expansion, 10^{-7} in./in./°C.		Upper Working Temperatures (Mechanical Considerations Only)				Thermal Shock Res. Plates 6" × 6"		
										Annealed		Tempered		Annealed		
Glass Code†	Type	Color	Principal Use	Forms Usually Available	Weathering	Water	Acid	0–300°C 32–572°F	Room Temp.-Setting Point	Normal Service °C.	Extreme Limit °C.	Normal Service °C.	Extreme Limit °C.	⅛" Thk. °C.	¼" Thk. °C.	½" Thk. °C.
0010	Potash Soda Lead	Clear	Lamp Tubing	T	2	2	2	93	100	110	380			65	50	35
0080	Soda Lime	Clear	Lamp Bulbs	B M T	3	2	2	92	103	110	460	220	250	65	50	35
0120	Potash Soda Lead	Clear	Lamp Tubing	T M	2	2	2	89	98	110	380	—		65	50	35
1720	Aluminosilicate	Clear	Ignition Tube	B T	1	1	3	42	52	200	650	400	450	135	115	75
1723	Aluminosilicate	Clear	Electron Tube	B T	1	1	3	46	54	200	650	400	450	125	100	70
1990	Potash Soda Lead	Clear	Iron Sealing	—	3	3	4	124	136	100	310	—	—	45	35	25
2405	Borosilicate	Red	General	B P U			—	43	51	200	480	—	—	135	115	75
2475	Soda Zinc	Red	Neon Signs	T	3	2	2	93	—	110	440	—	—	65	50	35
3320	Borosilicate	Canary	Tungsten Sealing		[3]1	[3]1	[3]2	40	43	200	480	—	—	145	110	80
6720	Soda Zinc	Opal	General	P	[2]—	1	2	80	92	110	480	220	275	70	60	40
6750	Soda Barium	Opal	Lighting Ware	B P R	[2]	2	2	88	—	110	420	220	220	65	50	35
6810	Soda Zinc	Opal	Lighting Ware	B P R	[2]—	1	2	69	—	120	470	240	270	85	70	45
7040	Borosilicate	Clear	Kovar Sealing	B T	[3]3	[3]3	[3]4	48	54	200	430	—	—	—	—	—
7050	Borosilicate	Clear	Series Sealing	T	[3]3	[3]3	[3]4	46	51	200	440	235	235	125	100	70
7052	Borosilicate	Clear	Kovar Sealing	B M P T	[3]2	[3]2	[3]4	46	53	200	420	210	210	125	100	70
7056	Borosilicate	Clear	Kovar Sealing	B T P	2	2	4	51	57	200	460	—	—	—	—	—
7070	Borosilicate	Clear	Low Loss Electrical	B M P T	[3]2	[3]2	[3]2	32	39	230	430	230	230	180	150	100
7250	Borosilicate	Clear	Seal Beam Lamps	P	[3]1	[3]2	[3]2	36	38	230	460	260	260	160	130	90
7570	High Lead	Clear	Solder Sealing	—	1	1	4	84	92	100	300	—	—	—	—	—
7720	Borosilicate	Clear	Tungsten Sealing	B P T	[3]2	[3]2	[3]2	36	43	230	460	260	260	160	130	90
7740	Borosilicate	Clear	General	B P S T U	[3]1	[3]1	[3]1	33	35	230	490	260	290	180	150	100
7760	Borosilicate	Clear	General	B P	2	2	2	34	37	230	450	250	250	160	130	90
7900[1]	96% Silica	Clear	High Temp.	B P T U M	1	1	1	8	7	800	1100	—	—	1250	1000	750
7913	96% Silica	Clear	High Temp.	B P R S T	1	1	1	8	7	900	1200	—	—	—	—	—
7940	Fused Silica	Clear	Ultrasonic	U	1	1	1	5.5	7	900	1100	—	—	1250	1000	750
8160	Potash Soda Lead	Clear	Electron Tubes	P T	2	2	3	91	100	110	380		—	65	50	35
8161	Potash Lead	Clear	Electron Tubes	P T	2	1	4	90	97	110	390	—	—	—	—	—
8363	High Lead	Clear	Radiation Shielding	L C	3	1	4	104	112	100	200	—	—	—	—	—
8871	Potash Lead	Clear	Capacitors		2	1	4	102	113	125	300	—	—	55	45	35
9010	Potash Soda Barium	Grey	TV Bulbs	P	2	2	2	89	102	110	380		—			
9700	Borosilicate	Clear	u v Transmission	T U	[3]1	[3]1	[3]2	39	39	220	500	—	—	150	120	80
9741	Borosilicate	Clear	u v Transmission	B U T	[3]3	[3]3	[3]4	39	49	200	390	—		150	120	80

† Corning Glass Works code numbers are used in this table.

TABLE C.11 (continued) Properties of Silicate Glasses*

10	11				12	13	14			15			16			17	18
Thermal Stress Resistance °C.	Viscosity Data†				Impact Abrasion Resistance	Density grams per C.C.	Young's Modulus		Poisson's Ratio	Log_{10} of Volume Resistivity			Dielectric Properties at 1 Mc and 20°C			Refractive Index Sod. D Line (.5893 Microns)	Glass Code
	Strain Point °C.	Annealing Point °C.	Softening Point °C.	Working Point °C.			(10^6lb./sq. in)	(10^6kg/cm²)		25°C. 77°F	250°C. 482°F	350°C. 662°F	Power Factor	Dielectric Const.	Loss Factor		
19	395	435	625	985	0.8	2.86	8.9	0.63	.21	17.+	8.9	7.0	.16%	6.7	1.%	1.539	0010
17	470	510	695	1005	1.2	2.47	10.0	0.70	.24	12.4	6.4	5.1	.9	7.2	6.5	1.512	0080
20	395	435	630	980	0.8	3.05	8.6	0.60	.22	17.+	10.1	8.0	.12	6.7	.8	1.560	0120
28	670	715	915	1190	2.0	2.52	12.7	0.89	0.25	—	11.4	9.5	.38	7.2	2.7	1.530	1720
25	670	710	910	1175	2.0	2.64	12.5	0.88	0.25	—	13.5	11.3	.16%	6.3	1.0%	1.547	1723
14	330	360	500	755	—	3.47	8.4	0.59	.25	—	10.1	7.7	.04	8.3	.33	—	1990
‘37	500	530	770	1085	—	2.50	9.9	0.70	0.21	—	—	—	—	—	—	1.507	2405
‘17	440	480	690	1040	—	2.59	10.0	0.70	—	—	7.8	6.2	—	—	—	1.511	2475
‘40	500	540	780	1155	—	2.27	9.4	0.66	0.19	—	8.6	7.1	.30	4.9	1.5	1.481	3320
19	510	550	775	1010	—	2.58	10.2	0.72	.21	—	—	—	—	—	—	1.507	6720
‘18	445	485	670	1040	—	2.59	—	—	—	—	—	—	—	—	—	1.513	6750
‘23	490	530	770	1010	—	2.65	—	—	—	—	—	—	—	—	—	1.508	6810
37	450	490	700	1080	—	2.24	8.6	0.60	.23	—	9.6	7.8	.20	4.8	1.0	1.480	7040
39	460	500	705	1025	—	2.24	8.7	0.61	.22	16.	8.8	7.2	.33	4.9	1.6	1.479	7050
41	435	480	710	1115	—	2.28	8.2	0.58	.22	17.	9.2	7.4	.26	4.9	1.3	1.484	7052
34	470	510	720	1045	—	2.29	9.2	0.65	.21	—	10.2	8.3	.27	5.7	1.5	1.487	7056
66	455	495	—	1070	4.1	2.13	7.4	0.52	.22	17.+	11.2	9.1	.06	4.1	.25	1.469	7070
48	490	540	780	1190	3.2	2.24	9.2	0.65	.20	15.	8.2	6.7	.27	4.7	1.3	1.475	7250
21	340	365	440	560	—	5.42	8.0	0.56	.28	—	10.6	8.7	.22	15.	3.3	—	7570
49	485	525	755	1140	3.2	2.35	9.1	0.64	.20	16.	8.8	7.2	.27	4.7	1.3	1.487	7720
53	515	565	820	1245	3.1	2.23	9.1	0.64	.20	15.	8.1	6.6	.50	4.6	2.6	1.474	7740
52	480	525	780	1210	—	2.23	9.1	0.64	—	17.	9.4	7.7	.18	4.5	.79	1.473	7760
202	820	910	1500	—	3.5	2.18	10.0	0.70	.19	17.	9.7	8.1	.05	3.8	.19	1.458	7900‘
211	820	910	1500	—	3.5	2.18	9.6	0.67	.19	—	9.7	8.1	.04	3.8	0.15	1.458	7913
290	990	1050	1580	—	3.6	2.20	10.5	0.74	.16	—	11.8	10.2	.001	3.8	.0038	1.459	7940
‘18	395	435	630	975	—	2.98	—	—	—	—	10.6	8.4	.09	7.0	.63	1.553	8160
22	400	435	600	860	—	4.00	7.8	0.55	.24	—	12.0	9.9	.06	8.3	0.50	1.659	8161
19	300	315	380	460	—	6.22	7.4	0.52	.27	—	9.2	7.5	.19	17.0	3.2	1.97	8363
17	350	385	525	785	—	3.84	8.4	0.59	.26	—	11.1	8.8	.05	8.4	.42	—	8871
18	405	445	650	1010	—	2.64	9.8	0.69	.21	—	8.9	7.0	.17	6.3	1.1	1.507	9010
45	520	565	805	1200	—	2.26	9.6	0.67	.20	15.	8.0	6.5	—	—	—	1.478	9700
55	410	450	705	—	—	2.16	7.2	0.51	.23	17.+	9.4	7.6	—	—	—	1.468	9741

†Viscosities at these four temperatures are approximately as follows: $10^{14.5}$ poises at the strain point, 10^{13} poises at the annealing point, $10^{7.8}$ poises at the softening point, at 10^4 poises at the working point.

TABLE C.12 Properties of Window Glass*: Transmittance of Sheet and Plate Glass

Type or tint	*Nominal thickness, in.*	*Weight, lb/ft²*	*Transmittance*	
			Total visible daylight, %	*Direct 90° solar energy, %*
Sheet	$\frac{1}{16}$	0.81	91	89
Sheet	$\frac{5}{64}$	1.00	91	88
Sheet	$\frac{3}{32}$	1.22	90	87
Sheet	$\frac{1}{8}$	1.64	90	86
Sheet	$\frac{3}{16}$	2.47	89	84
Sheet	$\frac{7}{32}$	2.85	89	82
Plate or float	$\frac{1}{8}$	1.64	90	86
Plate or float	$\frac{1}{4}$	3.28	88	79
Plate or float	$\frac{5}{16}$	4.09	88	77
Plate or float	$\frac{3}{8}$	4.91	87	74
Plate or float	$\frac{1}{2}$	6.55	86	70
Plate or float	$\frac{5}{8}$	8.18	85	65
Plate or float	$\frac{3}{4}$	9.83	83	60
Plate or float	$\frac{7}{8}$	11.45	81	55
Plate or float	1	13.13	79	49
Gray[a]	$\frac{1}{4}$	3.28	43	46
Bronze[a]	$\frac{1}{4}$	3.28	49	45
Green[a]	$\frac{1}{4}$	3.28	75	46
Double[b]	$\frac{1}{4}$ each	6.56	78	—

Note: Many types of glass are available, including tempered heat-strengthened glass, laminated shatterproof glass, conductive-coated glass, reflective-coated glass. Several double-pane combinations are offered.

Direct 90° transmittance of solar ultraviolet radiation through non-tinted window glass is about 85 percent as high as the values for total solar energy transmittance. Ultraviolet transmittance of gray or bronze glass is lower.

Infrared transmittance is considerably lower than visual transmittance. This is significant in view of the large percentage of infrared radiation from most sources.

Visible reflectance of untinted glass is about 8 percent.

Approximate shading coefficients, ASHRAE, $\frac{1}{4}$-in. glass only: clear, 0.93; gray, 0.67; bronze, 0.65; green, 0.67.

Overall heat transfer coefficient of window area (air to air) is usually assumed to be 1.0 Btu/ft² hr, but it is lower if there is no wind.

For other data on shading coefficients, spectral transmittance, coated glass, special glasses, etc., see Tables 1-91, 2-19, and 7-19.

[a]Transmittance of tinted glass depends on depth of tint.
[b]Two $\frac{1}{4}$-in. panes with $\frac{1}{2}$-in. air space, sealed.

*Tables compiled from several sources.

TABLE C.13 Properties and Uses of American Woods*

Species	*Specific gravity*		*Characteristics*	*Uses*	*Weight*		
	Green	*Dry*			*lb/cu ft, green*	*lb/cu ft, air-dry 12%*	*lb/1000 board ft, air-dry 12%*
Alder, red	0.37	0.41	Low shrinkage; moderate in strength, shock resistance, hardness, and weight†	Furniture; sash; doors; millwork	46	28	2330
Ash, black	0.45	0.49	Light in weight†	Cabinets; veneer; cooperage, containers	52	34	2830
Ash, Oregon	0.50	0.55	Similar to but lighter than white ash†	Similar to white ash	46	38	3160
Ash, white	0.54	0.58	Heavy; hard; stiff; strong; high shock resistance†	Handles; ladder rungs; baseball bats; farm implements; car parts	48	41	3420
Bald cypress (Southern cypress)			Moderate in strength, weight, hardness, and shrinkage**	Building construction; beams; posts; ties; tanks; ships; paneling	51	32	2670
Beech, American	0.56	0.64	Heavy; high strength, shock resistance, and shrinkage; uniform texture†	Flooring; furniture; handles; kitchenwear; ties (treated)	54	45	3750
Birch	0.57	0.63	Heavy; high strength, shock resistance, and shrinkage; uniform texture†	Interior finish; dowels; ties (treated); veneer; musical instruments	57	44	3670
Cottonwood	0.37	0.40	Uniform texture; does not split readily; moderate in weight, strength, hardness, and shrinkage	Crates; trunks; car parts; farm implements	49	28	2330
Douglas fir	0.41	0.44	Moderate in strength, weight, shock resistance, and shrinkage‡	Building and construction; poles; veneer; plywood; ships; furniture; boxes	38	34	2830
Elm	0.57	0.63	Moderate in strength, weight, and hardness; high in shock resistance and shrinkage; good in bending†	Cooperage; baskets; crates; veneer; vehicle parts	54	34	2920
Hemlock, Eastern	0.38	0.40	Moderate in weight, strength, and hardness†	Building and construction; boxes	50	28	2330
Hemlock, Western	0.38	0.42	Moderate in weight, strength, and hardness†	Sash; doors; posts; piles; building and construction	41	29	2420
Hickory, true	0.65	0.73	High toughness, hardness, shock resistance, strength, and shrinkage†	Dowels; spokes; poles; shafts; gymnasium equipment	63	51	4250
Incense cedar	0.35		Uniform texture; easy to season; low shrinkage; shock resistance, weight, and stiffness**	Lumber; fence posts; ties; poles; shingles	45		
Larch, Western	0.48	0.52	Moderate in strength, weight, shock resistance, hardness, and shrinkage‡	Doors; sash; posts; pilings; building and construction	48	36	3000

TABLE C.13 (continued) Properties and Uses of American Woods*

Species	*Specific gravity*		*Characteristics*	*Uses*	*Weight*		
	Green	*Dry*			*lb/cu ft, green*	*lb/cu ft, air-dry 12%*	*lb/1000 board ft, air-dry 12%*
Locust, black	0.66	0.69	High in shock resistance, weight, and hardness; very high strength; moderate shrinkage**	Mine timbers; posts; poles; ties	58	48	4000
Maple	0.44	0.48	High in hardness, weight, strength, shock resistance, and shrinkage; uniform texture†	Flooring; furniture; trim; spools; farm implements	54	40	3330
Oak, red and white	0.57	0.63	High in hardness, weight, strength, shock resistance, and shrinkage; red†, white‡	Trim; ships; flooring; ties; furniture; cooperage; piles	64	44	3670
Pine, jack			Coarse texture; low strength, stiffness, shock resistance, and shrinkage	Box lumber; fuel; mine timber; ties; poles; posts			
Pine, lodgepole	0.38	0.41	Moderate in weight, hardness, strength, shock resistance, and shrinkage; easy to work‡	Poles; mine timber; ties; construction	39	29	2420
Pine			High shrinkage; moderate strength, stiffness, hardness, and shock resistance	General construction; ties; poles; posts			
Pine, Ponderosa	0.38	0.40	Moderate in weight, shock resistance, shrinkage, and hardness; easy to work†	Building; paneling; sash; frames	45	28	2330
Pine, S. yellow	0.47	0.51	Moderate in shock resistance, shrinkage, and hardness; high in strength‡	Building and construction; poles; pilings; boxes	55	41	3420
Pine, sugar	0.35	0.36	Low shock resistance; easy to work; moderate strength†	Sash; counters; blinds; patterns	52	25	2080
Pine, Western white	0.36	0.38	Moderate in strength, shock resistance, shrinkage, and hardness; easy to work†	Building and construction; patterns; boxes	35	27	2250
Red cedar, Eastern and Western	0.44	0.47	High shock resistance; low stiffness and shrinkage; moderate in strength and hardness**	Fence posts; closet liners; chests; flooring	37	37	2750
Redwood	0.38	0.40	Low shrinkage; medium in weight, strength, hardness, and shock resistance**	Posts; doors; interiors; cooling towers	50	28	2330
Spruce, Eastern	0.38	0.40	Moderate in hardness, shock resistance, weight, shrinkage, and strength†	Building; millwork; boxes; ladders	34	28	2330

TABLE C.13 (continued) Properties and Uses of American Woods*

Species	*Specific gravity*		*Characteristics*	*Uses*	*Weight*		
	Green	*Dry*			*lb/cu ft, green*	*lb/cu ft, air-dry 12%*	*lb/1000 board ft, air-dry 12%*
Spruce, Engelmann	0.31	0.33	Generally straight grained; light in weight; low strength as a beam or post; low shock resistance; moderate shrinkage	Mine timber; ties; poles; flooring; studding; paper	39	23	1920
Spruce, Sitka	0.37	0.40	Moderate in weight, hardness, strength, shock resistance, and shrinkage†	Important in boat and plane construction; sash; doors; boxes; siding	33	28	2330
Sycamore	0.46	0.49	High shrinkage; moderate in weight, strength, hardness, and shock resistance†	Boxes; ties; posts; veneer; flooring; butcher blocks	52	34	2830
Tamarack	0.49	0.53	Coarse texture; moderate in strength, hardness, shrinkage, and shock resistance	Ties; mine timber; posts; poles; tanks; scaffolding	47	37	3080
Tupelo			Uniform texture; moderate in strength, hardness, shock resistance; high shrinkage; interlocked grain makes splitting difficult†	Flooring; planking; crates; furniture			
Walnut, black	0.51	0.55	Moderate shrinkage; high weight, strength, hardness, and shock resistance; easily worked and glued**	Gun stocks; cabinets; plywood; furniture; veneer	58	38	3170
White cedar	0.31	0.32	Low shrinkage, weight, shock resistance, and strength; soft; easily worked**	Poles; posts; ties; tanks; ships	24	23	1920
Willow, black			High strength and shock resistance; low beam strength and weight; interlocked grain	Lumber; veneer; charcoal; furniture; sub-flooring; studding			

†Decay resistance low.
‡Decay resistance medium.
**Decay resistance high.
*From: "Materials Data Book", E.R. Parker, McGraw-Hill Book Company, 1967, pp. 252–255.

Note: For weight-density in kg/m^3, multiply value in lb/ft^3 by 16.02.

TABLE C.14 Properties of Natural Fibers*

Because there are great variations within a given fiber class, average properties may be misleading. The following typical values are only a rough comparative guide.

Name	*Specific gravity*	*Tenacity, g/denier*	*Tensile strength, 10^3 psi*	*Elongation at break (dry), %*	*Standard regain, % of dry*[b]	*Fiber diameter, microns*	*Fiber length, in.*	*Fiber shape and kind*	*Resistant to*
ANIMAL ORIGIN									
Wool	1.32	1.0–1.7	17–29	23–35	15–18	17–40	1.5–5	Oval, crimped, scales	Age, weak acids, solvents
Silk	1.25	3.5–5	90	20–25	10	10–13		Flexible, soft, smooth	Heat, solvents, weak acids, wear
Cashmere						15–16	1–4	Round, scales, soft	
Mohair	1.32	1.2–1.5		30	13	24–50	6–12	Round, silky	Wear, age, solvents, weak acids
Camel hair	1.32	1.8		40	13	10–40	1–6	Oval, striated	Age, solvents
VEGETABLE ORIGIN									
Cotton	1.54	2–5	30–120	5–11	7.5–8.5	10–20	0.5–2	Flat, convoluted, ribbon	Age, heat, washing, wear, solvents, alkalies, insects
Jute (bast)	1.5		50	1–1.5	14	15–20		Woody, rough, polygon	
Sisal (leaf)	1.49	2.2	75	2–2.5	13	10–30	Strand 30–40	Stiff, straight	
Flax (bast)	1.52	4–7		2–3	12	15–18	Strand 40–50	Soft, fine	Age, solvents, washing, insects, weak acids, and alkalies
Kenaf (bast)			45			15–30		Polygon or oval	
Hemp (bast)	1.48			2		18–25	Strand 30–70	Polygon or oval, irregular	
Henequen (leaf)			60				Strand 30–60	Finer than sisal	
Abaca (leaf) (Manila)	1.48	2.3–2.9	100	2–3	13		Strand 30–120		
MINERAL ORIGIN									
Asbestos	2.5		40–200			Various	0.5–10	Smooth, straight	Heat to 400 deg C, acids, chemicals, organisms
Glass[a]	2.5	7–12	200–500	3–4.5	0	Various		Circular, smooth	Chemicals, insects
Silicate[a] (Ca, Al, Mg)	2.85				0				Heat to 900 deg C, most chemicals, insects, rot

Note: Wide variations may be expected, especially for different grades of cotton. Wet strength is lower (for rayon, very much lower), but it depends on the duration of soaking. The strength of yarn is only a fraction of the cumulative strength of all individual fibers.

Most fibers exhibit relaxation of stress at constant strain and also increase in elongation at constant load (creep). The stress-strain curve is greatly affected by the rate of extension. When the stress is removed, there is a quick elastic recovery, a delayed recovery, and a permanent set. Hence the elastic behavior of any fiber depends on its stress-strain history. The elastic recoveries of nylon and wool are high; those of cotton, flax, and rayon are much lower.

The heat capacity (specific heat) of most fibers is about one-third that of water.

Other fibers: Fur hair is slightly coarser than silk fibers. Camel and llama hairs are almost as coarse as wool but only about one-third the size of human hair. Horse hair is over 100 microns; hog bristles, over 200 microns. Jute, sisal, and hemp are intermediate between cotton and wool. These are rough average sizes, and many natural fibers range 50% above or below such averages.

[a]Here classified as natural fibers for convenience, although they are man-made by processing.
[b]Expected equilibrium moisture regain of dry fiber, in percent of dry weight, when exposed in air at 70 deg F, 65% relative humidity.

*Compiled from several sources.

TABLE C.15 Properties of Manufactured Fibers*

Chemical class; common name (sources)	*Specific gravity*	*Tenacity, g/denier*	*Tensile strength, 10^3 psi*	*Elongation at break, %*	*Regain (standard)*	*Softening point, deg C*	*Melting point, deg C*	*Flammability*	*Brittleness temp, deg C*
CELLULOSE FIBERS (NATURAL)									
Acetate	1.30	1.–1.3	18–25	20–30	6.5	140	230	Melts and burns	
Triacetate	1.32	1.2–1.4	20–28	25–30	3–4.5	225	300	Melts and burns	
Viscose rayon	1.51	2–2.6	30–46	17–25	13.		200[a]	Burns readily	
High-tenacity viscose	1.53	3–5	60–80	10–12	10		200[a]	Burns readily	< −114
Polynosic viscose	1.53	3–5	60–80	8–20	7		200[a]	Burns readily	
Cuprammonium rayon (cupro)	11.52	1.7–2.3	30–45	10–17	12.5		250[a]	Burns readily	
PROTEIN FIBERS (NATURAL)									
Animal: casein (milk)	1.3	1.0	15	60–70	14	100	150	Slow	
Vegetable—seed: soybeans, peanuts, corn	1.3	0.7–0.9	11–14	40–60	11–15	150	250	Slow	
Vegetable—latex: rubber (vulcanized)	1.0	0.4–0.6	4–7	700–900	0	300		Burns	−60
SYNTHETIC FIBERS									
Polyacrylonitrile (acrylic)	1.17	2–5	50–75	25–40	2	190	260	Burns	
Polyamide (nylon)	1.14	4–9	70–120	20–40	4	200	215–250	Slow	< −100
Polyester (PET dacron)	1.38	4–8	70–120	10–50	0.4	225	250–290	Low	
Polyethylene (olefin, low density)	0.92	3–6	40–70	25–40	0.15	90–120	120	Slow	−114
Polyethylene (olefin, high density)	0.95	5–7	60–80	10–20	0.01	120–130	140	Slow	−114
Polypropylene (olefin)	0.91	4.5–8	45–80	15–30	0–0.5	145	160–170	Self-ext. low	−70
Polyurethane (spandex)	1.1	0.5–1.0	7–16	500–700	1.0	190	250	Burns	
Polyvinyl chloride (PVC)	1.38	0.7–2	12–17	100–125	0.1	70	140[a]	No; chars	< −100
Polyvinyl alcohol (PVA)	1.3	3–7	60–90	15–28	5	230	240	Slow	
Polyvinylidene chloride (saran)	1.7	2	40	20–30	0.1	115–135	170	No	
Polytetrafluoroethylene (PTFE)	2.1	1.2–1.4	33	15–30	0	225	300[a]	No	

Note: Mechanical properties are for room temperature and humidity and based on unstressed cross section.

[a]Decomposition; does not melt.

*Compiled from several sources.

TABLE C.16 Properties of Rubbers and Elastomers*

Elastomers cannot be classified in any brief and simple manner, nor are they well characterized by the usual mechanical tests. The terms *rubber* and *synthetic rubber* are loosely applied to a great variety of elastic materials, from pure gum natural rubber and pure synthetics to cured, compounded, filled, and even reinforced products.

ASTM designations (D1418) by chemical polymer description are used in the following table; yet within each class the properties can vary widely, depending on the exact composition, heat treatment service temperature, and application. Typical uses, such as rubber springs and cushioning, permit an almost unlimited number of combinations of design variables.

Mechanically, rubbers may be expected to lose strength rapidly with increase in temperature, to show a large hysteresis in stress-strain behavior, to exhibit marked creep and set, and to be greatly affected by rates of load application or frequency of repeated stress. "Heat build-up", i.e., increase in temperature in service, as well as deterioration from environment (sunlight, oils, ozone, etc.) will reduce the valuable properties of many rubbers, both natural and synthetic.

The following data apply to typical samples of commercial elastomers for common uses.

KEY:

A—Acetone
B—Benzene
C—Carbon tetrachloride
D—Carbon disulfide
E—Phenol
F—Sulfur compounds
G—Glycerol or glycol
H—Hexane
I—Acids
J—Alkalies
K—Ketones
L—Alcohols
M—Ammonia
N—Turpentine
O—Coal derivatives; bitumens
P—Petroleum products
R—Aromatics
S—Salts
T—Heat or high temperature
U—Ultraviolet
V—Vegetable oils
W—Weathering
X—Oxidation
Y—Aging
Z—Ozone

Chemical name	*Polyisoprene*	*Butadiene*	*Styrene-butadiene*	*Acrylonitrile butadiene*
Other names	*Natural (or synthetic) rubber NR (IR)*	*BR Cis 4*	*Buna S Styrene SBR, GR-S*	*Nitrile, Buna N Hycar NBR, GR-A*
CHEMICAL AND PHYSICAL				
Specific gravity	0.93	1.0	1.0	1.0
Specific heat	0.40	0.45	0.40	0.47
Thermal conductivity				
W/cm·K	0.001 7	0.002 5	0.002 6	0.002 5
Btu/hr·ft·deg F	0.10	0.14	0.15	0.14
Service temperature, deg C				
min	−25	−40	−20	−20
max	90	90	75	110
Solvents, softeners	D,K,P,V	D,H,N,P	K,P,R,V	C,K,O,R
Resistant to	A,I,J,L	G,I,J,W,Y	G,I,L,S,X	G,I,K,L,P,S, T,V,W
Swelled by	D,P,V	A,P,V	P,V	A,E,N
MECHANICAL AND ELECTRICAL				
Tensile strength				
kg/cm² (max)	300.	210.	210.	295.
kpsi (max)	4.3	3.0	3.0	4.2
Elongation at break, %	600.	700.	600.	600.
Vol. resistivity, ohm-cm	10^{15}	10^{15}	10^{14}	10^{10}
Dielectric strength				
kV/cm	235		235	185
V/mil	600.		600.	475.
Dielectric constant	3.0	2.3	2.8	3.0
Power factor (50–100 Hz)	0.003	0.005	0.005	0.007
Rebound	Good	Good	Fair	Good
COMPARATIVE RATINGS—RESISTANCE TO				
Abrasion	Good	Excellent	Good	Excellent
Cold flow (set)	Excellent		Good	Good
Tearing	Good		Poor	Fair
Air permeability	Fair	Good	Fair	Excellent
Oxidation	Fair	Fair	Fair	Fair
Flame	Poor		Poor	Poor

*Compiled from several sources.

TABLE C.16 (continued) Properties of Rubbers and Elastomers*

Chemical name	*Polychloro-prene*	*Isobutylene-isoprene*	*Polysulfide*	*Polymethane*
Other names	*Neoprene[a] CR, GR-M*	*Butyl IIR, GR-I*	*Thiokol[a] PS, GR-P*	*Adiprene[a] PU*
CHEMICAL AND PHYSICAL				
Specific gravity	1.25	0.95	1.4	1.2
Specific heat	0.5	0.45	0.31	0.45
Thermal conductivity				
W/cm·K	0.002 1	0.001 3	0.003	0.001 3
Btu/hr·ft·deg F	0.12	0.075	0.17	0.075
Service temperature, deg C				
min	−20	−40	−15	−35
max	100	120	90	120
Solvents, softeners	A,B,C,D,I,N,R	D,P	C	
Resistant to	G,L,P,S,T,U,V, W,Y,Z	E,G,J,S,U,V, W,X,Y,Z	L,P,U,Z	P,V,X,Z
Swelled by	C,D,N,R	D,H,P	C,R	B,C,K,R
MECHANICAL AND ELECTRICAL				
Tensile strength				
kg/cm^2 (max)	240.	175.	90.	350.
kpsi (max)	3.5	2.5	1.3	5.0
Elongation at break, %	800.	700.	500.	550.
Vol. resistivity, ohm-cm	10^{11}	10^{17}	10^{8}	10^{11}
Dielectric strength				
kV/cm	195	295	125	195
V/mil	500	750	325	500
Dielectric constant	7.	2.4	8.	7.
Power factor (50–100 Hz)	.04	0.004	0.02	0.04
Rebound	Good	Poor	Poor	
COMPARATIVE RATINGS—RESISTANCE TO				
Abrasion	Excellent	Fair	Poor	Excellent
Cold flow (set)	Excellent	Fair	Poor	Poor
Tearing	Good	Good	Poor	Excellent
Air permeability	Good	Excellent	Good	Excellent
Oxidation	Good	Good	Good	Good
Flame	Excellent	Poor	Poor	Poor

[a]Proprietary.

Appendix D. Gases and Vapors

TABLE D.1 SI Units — Definitions, Abbreviations and Prefixes

BASIC UNITS—MKS

Length	meter	m	Electric current	ampere	A
Mass	kilogram	kg	Thermodynamic temperature	kelvin	K
Time	second	s	Luminous intensity	candela	cd

DERIVED UNITS

Property	*Units†*	*Abbreviations and dimensions*	
Acceleration	meter per second squared	m/s^2	
Activity (of radioactive source)	1 per second	s^{-1}	
Angular acceleration	radian per second squared	rad/s^{-1}	
Angular velocity	radian per second	rad/s	
Area	square meter	m^2	
Density	kilogram per cubic meter	kg/m^3	
Dynamic viscosity	newton-second per sq meter	$N{\cdot}s/m^2$	
Electric capacitance	farad	F	$(A{\cdot}s/V)$
Electric charge	coulomb	C	$(A{\cdot}s)$
Electric field strength	volt per meter	V/m	
Electric resistance	ohm		(V/A)
Entropy	joule per kelvin	J/K	
Force	newton	N	$(kg{\cdot}m/s^2)$
Frequency	hertz	hz	(s^{-1})
Illumination	lux	lx	(lm/m^2)
Inductance	henry	H	$(V{\cdot}s/A)$
Kinematic viscosity	sq meter per second	m^2/s	
Luminance	candela per sq meter	cd/m^2	
Luminous flux	lumen	lm	$(cd{\cdot}sr)$
Magnetomotive force	ampere	A	
Magnetic field strength	ampere per meter	A/m	
Magnetic flux	weber	Wb	$(V{\cdot}s)$
Magnetic flux density	tesla	T	(Wb/m^2)
Power	watt	W	(J/s)
Pressure	newton per square meter	N/m^2	
Radiant intensity	watt per steradian	W/sr	
Specific heat	joule per kilogram kelvin	J/kg K	
Thermal conductivity	watt per meter kelvin	W/m K	
Velocity	meter per second	m/s	
Volume	cubic meter	m^3	
Voltage, potential difference, electromotive force	volt	V	(W/A)
Wave number	1 per meter	m^{-1}	
Work, energy, quantity of heat	joule	J	$(N{\cdot}m)$

PREFIX NAMES OF MULTIPLES AND SUBMULTIPLES OF UNITS

Decimal equivalent	*Prefix*	*Pronunciation*	*Symbol*	*Exponential expression*
1,000,000,000,000	tera	tĕr′ȧ	T	10^{+12}
1,000,000,000	giga	jĭ′gȧ	G	10^{+9}
1,000,000	mega	mĕg′ȧ	M	10^{+6}
1,000	kilo	kĭl′ō	k	10^{+3}
100	hecto	hĕk′tō	h	10^{+2}
10	deka	dĕk′ȧ	da	10
0.1	deci	dĕs′ĭ	d	10^{-1}
0.01	centi	sĕn′tĭ	c	10^{-2}
0.001	milli	mĭl′ĭ	m	10^{-3}
0.000 001	micro	mī′krō	μ	10^{-6}
0.000 000 001	nano	năn′ō	n	10^{-9}
0.000 000 000 001	pico	pē′kō	p	10^{-12}
0.000 000 000 000 001	femto	fĕm′tō	f	10^{-15}
0.000 000 000 000 000 001	atto	ăt′tō	a	10^{-18}

Appendix E. Miscellaneous

TABLE E.1 Sizes and Allowable Unit Stresses for Softwood Lumber

American Softwood Lumber Standard. A voluntary standard for softwood lumber has been developing since 1922. Five editions of Simplified Practice Recommendation R16 were issued from 1924–53 by the Department of Commerce; the present NBS voluntary Product Standard PS 20-70, "American Softwood Lumber Standard", was issued in 1970. It was supported by the American Lumber Standards Committee, which functions through a widely representative National Grading Rule Committee.

Part a. Nominal and Minimum-Dressed Sizes of Lumber*

Item	*Thicknesses*			*Face widths*		
	Nominal	*Minimum-dressed*		*Nominal*	*Minimum-dressed*	
		Dry,[a] inches	*Green, inches*		*Dry,[a] inches*	*Green, inches*
Boards[b]				2	$1\frac{1}{2}$	$1\frac{9}{16}$
				3	$2\frac{1}{2}$	$2\frac{9}{16}$
				4	$3\frac{1}{2}$	$3\frac{9}{16}$
				5	$4\frac{1}{2}$	$4\frac{5}{8}$
	1	$\frac{3}{4}$	$\frac{25}{32}$	6	$5\frac{1}{2}$	$5\frac{5}{8}$
				7	$6\frac{1}{2}$	$6\frac{5}{8}$
	$1\frac{1}{4}$	1	$1\frac{1}{32}$	8	$7\frac{1}{4}$	$7\frac{1}{2}$
				9	$8\frac{1}{4}$	$8\frac{1}{2}$
	$1\frac{1}{2}$	$1\frac{1}{4}$	$1\frac{9}{32}$	10	$9\frac{1}{4}$	$9\frac{1}{2}$
				11	$10\frac{1}{4}$	$10\frac{1}{2}$
				12	$11\frac{1}{4}$	$11\frac{1}{2}$
				14	$13\frac{1}{4}$	$13\frac{1}{2}$
				16	$15\frac{1}{4}$	$15\frac{1}{2}$
Dimension				2	$1\frac{1}{2}$	$1\frac{9}{16}$
				3	$2\frac{1}{2}$	$2\frac{9}{16}$
				4	$3\frac{1}{2}$	$3\frac{9}{16}$
	2	$1\frac{1}{2}$	$1\frac{9}{16}$	5	$4\frac{1}{2}$	$4\frac{5}{8}$
	$2\frac{1}{2}$	2	$2\frac{1}{16}$	6	$5\frac{1}{2}$	$5\frac{5}{8}$
	3	$2\frac{1}{2}$	$2\frac{9}{16}$	8	$7\frac{1}{4}$	$7\frac{1}{2}$
	$3\frac{1}{2}$	3	$3\frac{1}{16}$	10	$9\frac{1}{4}$	$9\frac{1}{2}$
				12	$11\frac{1}{4}$	$11\frac{1}{2}$
				14	$13\frac{1}{4}$	$13\frac{1}{2}$
				16	$15\frac{1}{4}$	$15\frac{1}{2}$
Dimension				2	$1\frac{1}{2}$	$1\frac{9}{16}$
				3	$2\frac{1}{2}$	$2\frac{9}{16}$
				4	$3\frac{1}{2}$	$3\frac{9}{16}$
				5	$4\frac{1}{2}$	$4\frac{5}{8}$
	4	$3\frac{1}{2}$	$3\frac{9}{16}$	6	$5\frac{1}{2}$	$5\frac{5}{8}$
	$4\frac{1}{2}$	4	$4\frac{1}{16}$	8	$7\frac{1}{4}$	$7\frac{1}{2}$
				10	$9\frac{1}{4}$	$9\frac{1}{2}$
				12	$11\frac{1}{4}$	$11\frac{1}{2}$
				14		$13\frac{1}{2}$
				16		$15\frac{1}{2}$
Timbers	5 and thicker		$\frac{1}{2}$ off	5 and wider		$\frac{1}{2}$ off

[a] Maximum moisture content of 19% or less.

[b] Boards less than the minimum thickness for 1 in. nominal but $\frac{5}{8}$ in. or greater thickness dry ($\frac{11}{16}$ in. green) may be regarded as American Standard Lumber, but such boards shall be marked to show the size and condition of seasoning at the time of dressing. They shall also be distinguished from 1-in. boards on invoices and certificates.

*Reprinted from: "American Softwood Lumber Standard", NBS PS 20-70, National Bureau of Standards, 1970; available from Superintendent of Documents.

Note: This table applies to boards, dimensional lumber, and timbers. The thicknesses apply to all widths and all widths to all thicknesses.

TABLE E.1 (continued) Sizes and Allowable Unit Stresses for Softwood Lumber

The "American Softwood Lumber Standard", PS 20-70, gives the size and grade provisions for American Standard lumber and describes the organization and procedures for compliance enforcement and review. It lists commercial name classifications and complete definitions of terms and abbreviations.

Eleven softwood species are listed in PS 20-70, viz., cedar, cypress, fir, hemlock, juniper, larch, pine, redwood, spruce, tamarack, and yew. Five dimensional tables show the standard dressed (surface planed) sizes for almost all types of lumber, including matched tongue-and-grooved and shiplapped flooring, decking, siding, etc. Dry or seasoned lumber must have 19% or less moisture content, with an allowance for shrinkage of 0.7–1.0% for each four points of moisture content below the maximum. Green lumber has more than 19% moisture. Table A illustrates the relation between nominal size and dressed or green sizes.

National Design Specification. Part b is condensed from the 1971 edition of "National Design Specification for Stress-Grade Lumber and Its Fastenings," as recommended and published by the National Forest Products Association, Washington, D.C. This specification was first issued by the National Lumber Manufacturers Association in 1944; subsequent editions have been issued as recommended by the Technical Advisory Committee. The 1971 edition is a 65-page bulletin with a 20-page supplement giving "Allowable Unit Stresses, Structural Lumber," from which Part b has been condensed. The data on working stresses in this Supplement have been determined in accordance with the corresponding ASTM Standards, D245-70 and D2555-70.

Part b. Species, Sizes, Allowable Stresses, and Modulus of Elasticity of Lumber

Normal loading conditions: Moisture content not over 19%, No. 1 grade, visual grading. To convert psi to N/m^2, multiply by 6 895.

Species[a]	*Sizes, nominal*	*Typical grading agency, 1971*[b]	*Allowable unit stresses, psi*[d]				*Modulus of elasticity, psi*
			Extreme fiber in bending[c]	*Tension parallel to grain*	*Compression perpendicular*	*Compression parallel*	
CEDAR							
Northern white	2 × 4	NL, NH	1 100	600	205	675	800 000
	2 or 4 × 6+	NL, NH	1 000	575	205	675	800 000
Western	2 × 4	NC	1 450	725	285	975	1 100 000
	2 or 4 × 6+	NC, WW	1 250	725	285	975	1 100 000
FIR							
Balsam	2 × 4	NL, NH	1 300	675	170	825	1 200 000
	2 or 4 × 6+	NL, NH	1 150	650	170	825	1 200 000
Douglas (larch)	2 × 4	WC, NC	2 400	1 200	385	1 250	1 800 000
	2 or 4 × 6+	WC, NC	1 750	1 000	385	1 250	1 800 000
HEMLOCK							
Eastern (tamarack)	2 × 4	NL, NH	1 750	900	365	1 050	1 300 000
	2 or 4 × 6+	NL, NH	1 500	875	365	1 050	1 300 000
Hem-fir	2 × 4	WC, NC	1 600	825	245	1 000	1 500 000
	2 or 4 × 6+	WC, NC	1 400	800	245	1 000	1 500 000
Mountain	2 × 4	WC, WW	1 700	850	370	1 000	1 300 000
	2 or 4 × 6+	WC, WW	1 450	850	370	1 000	1 300 000
PINE							
Idaho white	2 × 4	WW	1 400	725	240	925	1 400 000
	2 or 4 × 6+	WW	1 200	700	240	925	1 400 000
Lodgepole	2 × 4	WW	1 500	750	250	900	1 300 000
	2 or 4 × 6+	WW	1 300	750	250	900	1 300 000

TABLE E.1 (continued) Sizes and Allowable Unit Stresses for Softwood Lumber

Species[a]	*Sizes, nominal*	*Typical grading agency, 1971*[b]	*Allowable unit stresses, psi*[d]				*Modulus of elasticity, psi*
			Extreme fiber in bending[c]	*Tension parallel to grain*	*Compression perpendicular*	*Compression parallel*	
PINE *(continued)*							
Northern	2×4	NL, NH	1 600	825	280	975	1 400 000
	2 or 4×6+	NL, NH	1 400	800	280	975	1 400 000
Ponderosa (sugar)	2×4	WW, NC	1 400	700	250	850	1 200 000
	2 or 4×6+	WW, NC	1 200	700	250	850	1 200 000
Red	2×4	NC	1 350	700	280	825	1 300 000
	2 or 4×6+	NC	1 150	675	280	825	1 300 000
Southern	2×4	SP	2 000	1 000	405	1 250	1 800 000
	2 or 4×6+	SP	1 750	1 000	405	1 250	1 800 000
REDWOOD							
California	2 or 4×2 or 4	RI	1 950	1 000	425	1 250	1 400 000
	2 or 4×6 to 12	RI	1 700	1 000	425	1 250	1 400 000
SPRUCE							
Eastern	2×4	NL, NH	1 500	750	255	900	1 400 000
	2 or 4×6+	NL, NH	1 250	750	255	900	1 400 000
Engelmann	2×4	WW	1 300	675	195	725	1 200 000
	2 or 4×6+	WW	1 150	650	195	725	1 200 000
Sitka	2×4	WC	1 550	775	280	925	1 500 000
	2 or 4×6+	WC	1 300	775	280	925	1 500 000

Note: Allowable unit stresses in horizontal shear are in the range of 60–100 psi for No. 1 grade.

[a]Grade designations are not entirely uniform. Values in the table apply approximately to "No. 1." There is seldom more than one better grade than No. 1, and this may be designated as select, select structural, dense, or heavy. In addition to lower grades 2 and 3, there may be other lower grades, designated as construction, standard, stud, and utility. In bending and tension the allowable unit stresses in the lowest recognized grade (utility) are of the order of $\frac{1}{8}$ to $\frac{1}{6}$ of the allowable stresses for grade No. 1. The tabular values for allowable bending stress are for the extreme fiber in "repetitive member uses," and edgewise use. The original tables give correction factors, which are less than unity for moist locations and for short-time loading; they are greater than unity if the moisture content of the wood in service is 15% or less. In general, all data apply to uses within covered structures. From the extensive tables, only the No. 1 grade in nominal 2 × 4 size and 2-in. or 4-in. planks, 6 in., and wider have been selected for illustration.

In a few cases the allowable stresses specified for the Canadian products will vary slightly from those given here for the same species by the U.S. agencies.

[b]Grading agencies represented by letters in this column are as follows:

NC = National Lumber Grades Authority (a Canadian agency)
NH = Northern Hardwood and Pine Manufacturers Association
NL = Northern Lumber Manufacturers Association
RI = Redwood Inspection Service
SP = Southern Pine Inspection Bureau
WC = West Coast Lumber Inspection Bureau
WW = Western Wood Products Association

[c]It is assumed that all members are so framed, anchored, tied, and braced that they have the necessary rigidity.

[d]For short term loads, these values may be increased: add 15% for 2-month snow load; add 33% for wind or earthquake; add 100% for impact load.

REFERENCES

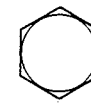
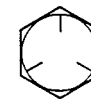
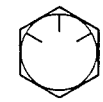
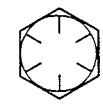
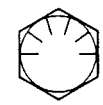
"Wood Handbook", Handbook No. 72, U.S. Department of Agriculture, 1955.

"Timber Construction Manual", American Institute of Timber Construction, John Wiley & Sons, Inc., 1966.

"National Design Specification for Stress-Grade Lumber and Its Fastenings", National Forest Products Association, Washington D.C., 1971.

TABLE E.2 Standard Grades of Bolts

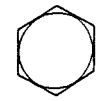
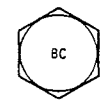
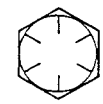
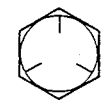
Part a: SAE Grades for Steel Bolts

SAE grade no.	Size range incl.	Proof strength,† kpsi	Tensile strength,† kpsi	Material	Head marking
1	¼–1½			Low- or medium-carbon steel	
2	¼–¾ ⅞–1½	55 33	74 60		
5	¼–1	85	120	Medium-carbon steel, Q & T	
	1⅛–1½	74	105		
5.2	¼–1	85	120	Low-carbon martensite steel, Q & T	
7	¼–1½	105	133	Medium-carbon alloy steel, Q & T‡	
8	¼–1½	120	150	Medium-carbon alloy steel, Q & T	
8.2	¼–1	120	150	Low-carbon martensite steel, Q & T	

†Minimum values.
‡Roll threaded after heat treatment.
SOURCES: See "Helpful Hints," by Russell, Burdsall & Ward Corp., Mentor, Ohio 44060; and Chap. 23.

TABLE E.2 (continued) Standard Grades of Bolts

Part b: ASTM Grades for Steel Bolts

ASTM designation	Size range incl.	Proof strength,† kpsi	Tensile strength,† kpsi	Material	Head marking
A307	¼ to 4			Low-carbon steel	
A325 type 1	½ to 1 1⅛ to 1½	85 74	120 105	Medium-carbon steel, Q & T	
A325 type 2	½ to 1 1⅛ to 1½	85 74	120 105	Low-carbon martensite steel, Q & T	
A325 type 3	½ to 1 1⅛ to 1½	85 74	120 105	Weathering steel, Q & T	
A354 grade BC				Alloy steel, Q & T	
A354 grade BD	¼ to 4	120	150	Alloy steel, Q & T	
A449	¼ to 1 1⅛ to 1½ 1¾ to 3	85 74 55	120 105 90	Medium-carbon steel, Q & T	
A490type	½ to 1½	120	150	Alloy steel, Q & T	
A490type 3				Weathering steel, Q & T	

† Minimum value.

Sources: See "Helpful Hints," by Russell,Burdsall&Ward Corp.,Mentor,Ohio 44060; and Chapter 23.

TABLE E.2 (continued) Standard Grades of Bolts

Part c: Metric Mechanical Property Classes for Steel Bolts, Screws, and Studs

Property class	Size range incl.	Proof strength, MPa	Tensile strength, MPa	Material	Head marking
4.6	M5–M36	225	400	Low- or medium-carbon steel	
4.8	M1.6–M16	310	420	Low- or medium-carbon steel	
5.8	M5–M24	380	520	Low- or medium-carbon steel	
8.8	M16–M36	600	830	Medium-carbon steel, Q & T	
9.8	M1.6–M16	650	900	Medium-carbon steel, Q & T	
10.9	M5–M36	830	1040	Low-carbon martensite steel, Q & T	
12.9	M1.6–M36	970	1220	Alloy steel, Q & T	

sources: "Helpful Hints," by Russell, Burdsall & Waard Corp,. Mentor, Ohio 44060; see also Chapter 23 and SAEstandard J1199, and ASTM standard F568.

TABLE E.3 Steel Pipe Sizes

Nominal Pipe Size, in.	*Outside Diameter, in.*	*Schedule Number or Weight*	*Wall Thickness, in.*	*Inside Diameter, in.*	*Surface Area*		*Areas and Weights Cross-sectional*		*Weight*
					Outside, ft²/ft	*Inside, ft²/ft*	*Metal Area, in.²*	*Flow Area, in.²*	*Pipe lb/ft*
¾	1.05	40	0.113	0.824	0.275	0.216	0.333	0.533	1.131
		80	0.154	0.742	0.275	0.194	0.434	0.432	1.474
1	1.315	40	0.133	1.049	0.344	0.275	0.494	0.864	1.679
		80	0.179	0.957	0.344	0.250	0.639	0.719	2.172
1¼	1.660	40	0.140	1.38	0.434	0.361	0.668	1.496	2.273
		80	0.191	1.278	0.434	0.334	0.881	1.283	2.997
1½	1.900	40	0.145	1.61	0.497	0.421	0.799	2.036	2.718
		80	0.200	1.50	0.497	0.393	1.068	1.767	3.632
2	2.375	40	0.154	2.067	0.622	0.541	1.074	3.356	3.653
		80	0.218	1.939	0.622	0.508	1.477	2.953	5.022
2½	2.875	40	0.203	2.469	0.753	0.646	1.704	4.79	5.794
		80	0.276	2.323	0.753	0.608	2.254	4.24	7.662
3	3.5	40	0.216	3.068	0.916	0.803	2.228	7.30	7.58
		80	0.300	2.900	0.916	0.759	3.016	6.60	10.25
3½	4.0	40	0.226	3.548	1.047	0.929	2.680	9.89	9.11
		80	0.318	3.364	1.047	0.881	3.678	8.89	12.51
4	4.5	40	0.237	4.026	1.178	1.054	3.17	12.73	10.79
		80	0.337	3.826	1.178	1.002	4.41	11.50	14.99
5	5.563	10 S	0.134	5.295	1.456	1.386	2.29	22.02	7.77
		40	0.258	5.047	1.456	1.321	4.30	20.01	14.62
		80	0.375	4.813	1.456	1.260	6.11	18.19	20.78
6	6.625	10 S	0.134	6.357	1.734	1.664	2.73	31.7	9.29
		40	0.280	6.065	1.734	1.588	5.58	28.9	18.98
		80	0.432	5.761	1.734	1.508	8.40	26.1	28.58
8	8.625	10 S	0.148	8.329	2.258	2.180	3.94	54.5	13.40
		30	0.277	8.071	2.258	2.113	7.26	51.2	24.7
		80	0.500	7.625	2.258	1.996	12.76	45.7	43.4
10	10.75	10 S	0.165	10.420	2.81	2.73	5.49	85.3	18.7
		30	0.279	10.192	2.81	2.67	9.18	81.6	31.2
		Extra heavy	0.500	9.750	2.81	2.55	16.10	74.7	54.7
12	12.75	10 S	0.180	12.390	3.34	3.24	7.11	120.6	24.2
		30	0.330	12.09	3.34	3.17	12.88	114.8	43.8
		Extra heavy	0.500	11.75	3.34	3.08	19.24	108.4	65.4
14	14.0	10	0.250	13.5	3.67	3.53	10.80	143.1	36.7
		Standard	0.375	13.25	3.67	3.47	16.05	137.9	54.6
		extra heavy	0.500	13.00	3.67	3.40	21.21	132.7	72.1
16	16.0	10	0.250	15.50	4.19	4.06	12.37	188.7	42.1
		Standard	0.375	15.25	4.19	3.99	18.41	182.7	62.6
		extra heavy	0.500	15.00	4.19	3.93	24.35	176.7	82.8
18	18.0	10 S	0.188	17.624	4.71	4.61	10.52	243.9	35.8
		Standard	0.375	17.25	4.71	4.52	20.76	233.7	70.6
		extra heavy	0.500	17.00	4.71	4.45	27.49	227.0	93.5
20	20.0	10 S	0.218	19.564	5.24	5.12	13.55	300.6	46.1
		Standard	0.375	19.25	5.24	5.04	23.12	291	78.6
		extra heavy	0.500	19.00	5.24	4.97	30.6	283.5	104.1
22	22.0	10	0.250	21.50	5.76	5.63	17.1	363	58.1
		Standard	0.375	21.25	5.76	5.56	25.5	355	86.6
		extra heavy	0.500	21.00	5.76	5.50	33.8	346	114.8
24	24.0	10	0.250	23.50	6.28	6.15	18.7	434	63.4
		Standard	0.375	23.25	6.28	6.09	27.8	425	94.6
		extra heavy	0.500	23.00	6.28	6.02	36.9	415	125.5
26	26.0	Standard	0.375	25.25	6.81	6.61	30.2	501	102.6
		extra heavy	0.500	25.00	6.81	6.54	40.1	491	136.2
30	30.0	10	0.312	29.376	7.85	7.69	29.1	678	98.9
		Standard	0.375	29.250	7.85	7.66	34.9	672	118.7
		extra heavy	0.500	29.00	7.85	7.59	46.3	661	157.6
34	34.0	Standard	0.375	33.250	8.90	8.70	39.6	868	134.7
		extra heavy	0.500	33.00	8.90	8.64	52.6	855	178.9
36	36.0	Standard	0.375	35.25	9.42	9.23	42.0	976	142.7
		extra heavy	0.500	35.00	9.42	9.16	55.8	962	189.6
42	42.0	Standard	0.375	41.25	11.0	10.8	49.0	1336	166.7
		extra heavy	0.500	41.00	11.0	10.73	65.2	1320	221.6

*Reprinted with permission, from: "Design Properties of Pipe", ©1958, Chemetron Corporation.

TABLE E.4 Commercial Copper Tubing*

The following table gives dimensional data and weights of copper tubing used for automotive, plumbing, refrigeration, and heat exchanger services. For additional data see the standards handbooks of the Copper Development Association, Inc., the ASTM standards, and the "SAE Handbook."

Dimensions in this table are actual specified measurements, subject to accepted tolerances. Trade size designations are usually by actual OD, except for water and drainage tube (plumbing), which measures 1/8-in. larger OD. A 1/2-in. plumbing tube, for example, measures 5/8-in. OD, and 2-in. plumbing tube measures 2 1/8-in. OD.

KEY TO GAGE SIZES

Standard-gage wall thicknesses are listed by numerical designation (14 to 21), BWG or Stubs gage. These gage sizes are standard for tubular heat exchangers. The letter *A* designates SAE tubing sizes for automotive service. Letter designations *K* and *L* are the common sizes for plumbing services, soft or hard temper.

OTHER MATERIALS

These same dimensional sizes are also common for much of the commercial tubing available in aluminum, mild steel, brass, bronze, and other alloys. Tube weights in this table are based on copper at 0.323 lb/in^3. For other materials the weights should be multiplied by the following approximate factors:

aluminum	0.30	monel	0.96
mild steel	0.87	stainless steel	0.89
brass	0.95		

Size, OD		*Wall Thickness*			*Flow Area*		*Metal*	*Surface Area*		
in.	*mm*	*in.*	*mm*	*gage*	*in.*^{2}	*mm*^{2}	*Area, in.*^{2}	*Inside, ft*^{2}*/ft*	*Outside, ft*^{2}*/ft*	*Weight, lb/ft*
1/8	3.2	.030	0.76	A	0.003	1.9	0.012	0.017	0.033	0.035
3/16	4.76	.030	0.76	A	0.013	8.4	0.017	0.034	0.049	0.058
1/4	6.4	.030	0.76	A	0.028	18.1	0.021	0.050	0.066	0.080
1/4	6.4	.049	1.24	18	0.018	11.6	0.031	0.038	0.066	0.120
5/16	7.94	.032	0.81	21A	0.048	31.0	0.028	0.065	0.082	0.109
3/8	9.53	.032	0.81	21A	0.076	49.0	0.033	0.081	0.098	0.134
3/8	9.53	.049	1.24	18	0.060	38.7	0.050	0.072	0.098	0.195
1/2	12.7	.032	0.81	21A	0.149	96.1	0.047	0.114	0.131	0.182
1/2	12.7	.035	0.89	20L	0.145	93.6	0.051	0.113	0.131	0.198
1/2	12.7	.049	1.24	18K	0.127	81.9	0.069	0.105	0.131	0.269
1/2	12.7	.065	1.65	16	0.108	69.7	0.089	0.97	0.131	0.344
5/8	15.9	.035	0.89	20A	0.242	156	0.065	0.145	0.164	0.251
5/8	15.9	.040	1.02	L	0.233	150	0.074	0.143	0.164	0.285
5/8	15.9	.049	1.24	18K	0.215	139	0.089	0.138	0.164	0.344
3/4	19.1	.035	0.89	20A	0.363	234	0.079	0.178	0.196	0.305
3/4	19.1	.042	1.07	L	0.348	224	0.103	0.174	0.196	0.362
3/4	19.1	.049	1.24	18K	0.334	215	0.108	0.171	0.196	0.418
3/4	19.1	.065	1.65	16	0.302	195	0.140	0.162	0.196	0.542
3/4	19.1	.083	2.11	14	0.268	173	0.174	0.151	0.196	0.674
7/8	22.2	.045	1.14	L	0.484	312	0.117	0.206	0.229	0.455
7/8	22.2	.065	1.65	16K	0.436	281	0.165	0.195	0.229	0.641
7/8	22.2	.083	2.11	14	0.395	255	0.206	0.186	0.229	0.800
1	25.4	.065	1.65	16	0.594	383	0.181	0.228	0.262	0.740
1	25.4	.083	2.11	14	0.546	352	0.239	0.218	0.262	0.927
1 1/8	28.6	.050	1.27	L	0.825	532	0.176	0.268	0.294	0.655

*Compiled and computed.

TABLE E.4 (continued) Commercial Copper Tubing*

Size, OD		*Wall Thickness*			*Flow Area*		*Metal Area,*	*Surface Area*		*Weight,*
in.	*mm*	*in.*	*mm*	*gage*	*in.²*	*mm²*	*in.²*	*Inside, ft²/ft*	*Outside, ft²/ft*	*lb/ft*
1 1/8	28.6	.065	1.65	16K	0.778	502	0.216	0.261	0.294	0.839
1 1/4	31.8	.065	1.65	16	0.985	636	0.242	0.293	0.327	0.938
1 1/4	31.8	.083	2.11	14	0.923	596	0.304	0.284	0.327	1.18
1 3/8	34.9	.055	1.40	L	1.257	811	0.228	0.331	0.360	0.884
1 3/8	34.9	.065	1.65	16K	1.217	785	0.267	0.326	0.360	1.04
1 1/2	38.1	.065	1.65	16	1.474	951	0.294	0.359	0.393	1.14
1 1/2	38.1	.083	2.11	14	1.398	902	0.370	0.349	0.393	1.43
1 5/8	41.3	.060	1.52	L	1.779	1148	0.295	0.394	0.425	1.14
1 5/8	41.3	.072	1.83	K	1.722	1111	0.351	0.388	0.425	1.36
2	50.8	.083	2.11	14	2.642	1705	0.500	0.480	0.628	1.94
2	50.8	.109	2.76	12	2.494	1609	0.620	0.466	0.628	2.51
2 1/8	54.0	.070	1.78	L	3.095	1997	0.449	0.520	0.556	1.75
2 1/8	54.0	.083	2.11	14K	3.016	1946	0.529	0.513	0.556	2.06
2 5/8	66.7	.080	2.03	L	4.77	3078	0.645	0.645	0.687	2.48
2 5/8	66.7	.095	2.41	13K	4.66	3007	0.760	0.637	0.687	2.93
3 1/8	79.4	.090	2.29	L	6.81	4394	0.950	0.771	0.818	3.33
3 1/8	79.4	.109	2.77	12K	6.64	4284	1.034	0.761	0.818	4.00
3 5/8	92.1	.100	2.54	L	9.21	5942	1.154	0.897	0.949	4.29
3 5/8	92.1	.120	3.05	11K	9.00	5807	1.341	0.886	0.949	5.12
4 1/8	104.8	.110	2.79	L	11.92	7691	1.387	1.022	1.080	5.38
4 1/8	104.8	.134	3.40	10K	11.61	7491	1.682	1.009	1.080	6.51

TABLE E.5 Standard Gages for Wire, Sheet, and Twist Drills

Gage	(1) Mfrs. steel sheet	(2) USS steel sheet (old)	(3) Birmingham or Stub	(4) W & M or Roebling steel wire	(5) AWG or B & S non-ferrous wire or sheet	Numbered twist drills	Copper wire (AWG) Circular mils	Ohms/1000 ft, 77°F	Lb/1000 ft	Sheet steel Lb/sq ft
0000000		0.500		0.4900						20.00
000000		0.469		0.4615	0.580					18.75
00000		0.438		0.4305	0.516					17.50
0000		0.406	.454	0.3938	0.460		212,000	0.0500	641.0	16.25
000		0.375	.425	0.3625	0.410		168,000	0.0630	508.0	15
00		0.344	.380	0.3310	0.365		133,000	0.0795	403.0	13.75
0		0.313	.340	0.3065	0.325		106,000	0.100	319.0	12.50
1		0.281	.300	0.2830	0.289	0.2280	83,700	0.126	253.0	11.25
2		0.266	.284	0.2625	0.258	0.2210	66,400	0.159	201.0	10.625
3	.2391	0.250	.259	0.2437	0.229	0.2130	52,600	0.201	159.0	10
4	.2242	0.234	.238	0.2253	0.204	0.2090	41,700	0.253	126.0	9.375
5	.2092	0.219	.220	0.2070	0.182	0.2055	33,100	0.319	100.0	8.75
6	.1943	0.203	.203	0.1920	0.162	0.2040	26,300	0.403	79.5	8.125
7	.1793	0.188	.180	0.1770	0.144	0.2010	20,800	0.508	63.0	7.5
8	.1644	0.172	.165	0.1620	0.128	0.1990	16,500	0.641	50.0	6.875
9	.1495	0.156	.148	0.1483	0.114	0.1960	13,100	0.808	39.6	6.25
10	.1345	0.141	.134	0.1350	0.102	0.1935	10,400	1.02	31.4	5.625
11	.1196	0.125	.120	0.1205	0.0907	0.1910	8,230	1.28	24.9	5
12	.1046	0.109	.109	0.1055	0.0808	0.1890	6,530	1.62	19.8	4.375
13	.0897	0.0937	.095	0.0915	0.0720	0.1850	5,180	2.04	15.7	3.75
14	.0747	0.0781	.083	0.0800	0.0641	0.1820	4,110	2.58	12.4	3.125
15	.0673	0.0703	.072	0.0720	0.0571	0.1800	3,260	3.25	9.86	2.813
16	.0598	0.0625	.065	0.0625	0.0508	0.1770	2,580	4.09	7.82	2.5
17	.0538	0.0562	.058	0.0540	0.0453	0.1730	2,050	5.16	6.20	2.25
18	.0478	0.0500	.049	0.0475	0.0403	0.1695	1,620	6.51	4.92	2
19	.0418	0.0437	.042	0.0410	0.0359	0.1660	1,290	8.21	3.90	1.75
20	.0359	0.0375	.035	0.0348	0.0320	0.1610	1,020	10.4	3.09	1.50
21	.0329	0.0344	.032	0.0318	0.0285	0.1590	810	13.1	2.45	1.375
22	.0299	0.0312	.028	0.0286	0.0253	0.1570	642	16.5	1.94	1.25
23	.0269	0.0281	.025	0.0258	0.0226	0.1540	509	20.8	1.54	1.125
24	.0239	0.0250	.022	0.0230	0.0201	0.1520	404	26.2	1.22	1
25	.0209	0.0219	.020	0.0204	0.0179	0.1495	320	33.0	0.970	0.875
26	.0179	0.0187	.018	0.0181	0.0159	0.1470	254	41.6	0.769	0.75
27	.0164	0.0172	.016	0.0173	0.0142	0.1440	202	52.5	0.610	0.6875
28	.0149	0.0156	.014	0.0162	0.0126	0.1405	160	66.2	0.484	0.625
29	.0135	0.0141	.013	0.0150	0.0113	0.1360	127	83.4	0.384	0.5625
30	.0120	0.0125	.012	0.0140	0.0100	0.1285	101	105	0.304	0.5
31	.0105	0.0109	.010	0.0132	0.0089	0.1200	79.7	133	0.241	0.4375
32	.0097	0.0102	.009	0.0128	0.0080	0.1160	63.2	167	0.191	0.4063
33	.0090	0.0094	.008	0.0118	0.0071	0.1130	50.1	211	0.152	0.375
34	.0082	0.0086	.007	0.0104	0.0063	0.1110	39.8	266	0.120	0.3438
35	.0075	0.0078	.005	0.0095	0.0056	0.1100	31.5	335	0.0954	0.3125
36	.0067	0.0070	.004	0.0090	0.0050	0.1065	25.0	423	0.0757	0.2813
37	.0064	0.0066		0.0085	0.0045	0.1040	19.8	533	0.0600	0.2656
38	.0060	0.0062		0.0080	0.0040	0.1015	15.7	673	0.0476	0.25
39				0.0075	0.0035	0.0995	12.5	848	0.0377	
40				0.0070	0.0031	0.0980	9.9	1070	0.0200	
41				0.0066	0.0028	0.0960				
42				0.0062	0.0025	0.0935				
43				0.0060	0.0022	0.0890				
44				0.0058	.0020	0.0860				
45				0.0055	.0018	0.0820				
46				0.0052	.0016	0.0810				
47				0.0050	.0014	0.0785				
48				0.0048	.0012	0.0760				
49				0.0046	.0011	0.0730				
50				0.0044	.0010	0.0700				

Note: The present trend, especially for sheet and strip, is to quote thickness as decimal or fraction of an inch rather than gage number. ANSI Standard preferred thicknesses have been adopted. These preferred sizes for thickness of uncoated sheet, strip, and plate under 0.25 in. are as follows: .224, .220, .180, .160, .140, .125, .112, .100, .090, .080, .071, .063, .056, .050, .045, .040, .036, .032, .028, .025, .022, .020, .018, .016, .014, .012, .011, .010, .009, .008, .007, .006, .005, .004.

KEY: (1) Manufacturer's standard for hot- and cold-rolled uncoated carbon steel sheet and most alloy steel sheet.
(2) U.S. Standard for cold-rolled steel strip and stainless and nickel alloy sheet.
(3) Birmingham or Stub for hot-rolled carbon and alloy steel strip and tubing.
(4) Washburn and Moen, Roebling, or U.S. Steel for steel wire.
(5) American wire gage or Brown and Sharpe for non-ferrous wire, sheet, and strip.

Dimensions in approximate decimals of an inch.

TABLE E.6 Properties of Typical Gaseous and Liquid Commercial Fuels*

Gaseous fuels	*Composition, percent by volume*								*Mol wt of fuel*	*Theor. air/fuel ratio by wt*	*Higher heating value, Btu/lb$_m$*	*Density, lb$_m$/ft^3*
	H_2	N_2	O_2	CH_4	CO	CO_2	C_2H_4	C_6H_6				
Blast furnace gas	1.0	60.0	—	—	27.5	11.5	—	—	29.6	0.667	1,170	.075 5[a]
Blue water gas	47.3	8.3	0.7	1.3	37.0	5.4	—	—	16.4	3.759	6,550	.042 2[a]
Carb. water gas	40.5	2.9	0.5	10.2	34.0	3.0	6.1	2.8	18.3	7.299	11,350	.046 6[a]
Coal gas	54.5	4.4	0.2	24.2	10.9	3.0	1.5	1.3	12.1	10.87	16,500	.031 1[a]
Coke-oven gas	46.5	8.1	0.8	32.1	6.3	2.2	3.5	0.5	13.7	17.24	17,000	.032 6[a]
Natural gas (15.8% C_2H_6)	—	0.8	—	83.4	—	—	—	—	18.3	17.24	24,100	.045 1[a]
Producer gas	14.0	50.9	0.6	3.0	27.0	4.5	—	—	24.7	14.29	2,470	.063 6[a]

Liquid commercial fuels	*Vapor*		*Gravity, API, 60°F*	*Distillation*			*Flash point, °F*	*Viscosity, centi-stokes, 100°F*	*Mol wt of fuel*	*Theor. air/fuel ratio by wt*	*Higher heating value, Btu/lb$_m$*	*Density, lb$_m$/ft^3*
	c_p, 60°F	*c_p/c_v, 60°F*		*10%, °F*	*90%, °F*	*End point, °F*						
	(approximately)											
Gasoline	0.4	1.05	63	121	320	397	0	—	113	14.93	20,460	43.8[b]
Gasoline	0.4	1.05	63	118	330	410	0	—	126[c]	14.97	20,260	46.1[b]
Kerosene	0.4	1.05	41.9	370	510	546	130	—	154[c]	14.99	19,750	51.5[b]
Diesel oil (1-D)	0.4	1.05	42	—	550	—	100	1.4–2.5	170	15.02	19,240	54.6[b]
Diesel oil (2-D)	0.4	1.05	36	—	540–576	—	125	2.0–5.8	184	15.06	19,110	57.4[b]
Diesel oil (4-D)	0.4	1.05	—	—	—	—	130	5.8–26.4	198	14.93	18,830	59.9[b]

[a]Based on dry air at 25°C and 760 mm Hg.
[b]Based on H_2O at 60°F, 1 atm (ρ = 62.367 lb$_m$/ft^3).
[c]Estimated.

*Abridged from: "Engineering Experimentation", G.L. Tuve and L.C. Domholdt, McGraw-Hill Book Company, 1966; and "The Internal Combustion Engine", 2nd ed., C.F. Taylor and E.S. Taylor, International Textbook Co., 1961.

Note: For heating value in J/kg, multiply the value in Btu/lb$_m$ by 2324. For density in kg/m^3, multiply the value in lb/ft^3 by 16.02.

TABLE E.7 Combustion Data for Hydrocarbons*

Hydrocarbon	*Formula*	*Higher heating value (vapor), Btu/lb$_m$*	*Theor. air/fuel ratio, by mass*	*Max flame speed, ft/sec*	*Adiabatic flame temp (in air), °F*	*Ignition temp (in air), °F*	*Flash point, °F*	*Flammability limits (in air), % by volume*	
PARAFFINS OR ALKANES									
Methane	CH_4	23875	17.195	1.1	3484	1301	gas	5.0	15.0
Ethane	C_2H_6	22323	15.899	1.3	3540	968–1166	gas	3.0	12.5
Propane	C_3H_8	21669	15.246	1.3	3573	871	gas	2.1	10.1
n-Butane	C_4H_{10}	21321	14.984	1.2	3583	761	−76	1.86	8.41
iso-Butane	C_4H_{10}	21271	14.984	1.2	3583	864	−117	1.80	8.44
n-Pentane	C_5H_{12}	21095	15.323	1.3	4050	588	< −40	1.40	7.80
iso-Pentane	C_5H_{12}	21047	15.323	1.2	4055	788	< −60	1.32	9.16
Neopentane	C_5H_{12}	20978	15.323	1.1	4060	842	gas	1.38	7.22
n-Hexane	C_6H_{14}	20966	15.238	1.3	4030	478	−7	1.25	7.0
Neohexane	C_6H_{14}	20931	15.238	1.2	4055	797	−54	1.19	7.58
n-Heptane	C_7H_{16}	20854	15.141	1.3	3985	433	25	1.00	6.00
Triptane	C_7H_{16}	20824	15.141	1.2	4035	849	—	1.08	6.69
n-Octane	C_8H_{18}	20796	15.093	—	—	428	56	0.95	3.20
iso-Octane	C_8H_{18}	20770	15.093	1.1	—	837	10	0.79	5.94
OLEFINS OR ALKENES									
Ethylene	C_2H_4	21636	14.807	2.2	4250	914	gas	2.75	28.6
Propylene	C_3H_6	21048	14.807	1.4	4090	856	gas	2.00	11.1
Butylene	C_4H_8	20854	14.807	1.4	4030	829	gas	1.98	9.65
iso-Butene	C_4H_8	20737	14.807	1.2	—	869	gas	1.8	9.0
n-Pentene	C_5H_{10}	20720	14.807	1.4	4165	569	—	1.65	7.70
AROMATICS									
Benzene	C_6H_6	18184	13.297	1.3	4110	1044	12	1.35	6.65
Toluene	C_7H_8	18501	13.503	1.2	4050	997	40	1.27	6.75
p-Xylene	C_8H_{10}	18663	13.663	—	4010	867	63	1.00	6.00
OTHER HYDROCARBONS									
Acetylene	C_2H_2	21502	13.297	4.6	4770	763–824	gas	2.50	81
Naphthalene	$C_{10}H_8$	17303	12.932	—	4100	959	174	0.90	5.9

*Based largely on: "Gas Engineers' Handbook", American Gas Association, Inc., Industrial Press, 1967.

REFERENCES

"American Institute of Physics Handbook", 2nd ed., D.E. Gray, Ed., McGraw-Hill Book Company, 1963.

"Chemical Engineers' Handbook", 4th ed., R.H. Perry, C.H. Chilton, and S.D. Kirkpatrick, Eds., McGraw-Hill Book Company, 1963.

"Handbook of Chemistry and Physics", 53rd ed., R.C. Weast, Ed., The Chemical Rubber Company, 1972; gives the heat of combustion of 500 organic compounds.

"Handbook of Laboratory Safety", 2nd ed., N.V. Steere, Ed., The Chemical Rubber Company, 1971.

"Physical Measurements in Gas Dynamics and Combustion", Princeton University Press, 1954.

Note: For heating value in J/kg, multiply the value in Btu/lb$_m$ by 2324. For flame speed in m/s, multiply the value in ft/s by 0.3048.