CHAPTER

1

A Review

Science is an edifice, with the latest brick almost always laid on an earlier one. The physics described in this book depends on classical concepts and techniques described in an earlier course in physics. You have taken such a course, one that uses calculus, but while you have not forgotten everything, in all likelihood you have not perfectly retained all the material presented in that course. Our aim in this chapter is to provide you with a kind of road map of the most important material from your introductory course. We emphasize those things that are important in a course in modern physics. We also remind you of some of the history of this material, because we think it is useful for you to be aware of the way things have developed, not just the results.

We do assume that you have retained *some* things. In particular, we do not review the mathematics that was necessary for your first physics course. Thus, you are expected to know that vectors are not the same as scalars, that the acceleration is a vector given by the second derivative of the displacement, and that partial derivatives may enter into some equations, as well as what the area integral in Gauss' law means.

Our review is no substitute for a first course. You learn physics by understanding the material well enough to solve problems, and we make no attempt to do examples or present problems in this chapter. That will come later. The material we present here is so brief that it is at best a reminder; if you want details, consult the text from your first physics course, which is a far better reference.

1–1 Newton's Laws

Isaac Newton was one of the singular geniuses in all human history. However, he had predecessors in the development of physics—Copernicus, Galileo, and Kepler come to mind—and contemporaries, of whom Christian Huygens and Newton's great rival Robert Hooke were perhaps the most illustrious. Gottfried Leibniz, also a contemporary of Newton, independently invented the differential calculus, something that Newton would never admit. While Newton was aware of such earlier work, it was rare that he acknowledged it. But he transformed it in such a way that it would have been all but unrecognizable to his predecessors. Newton created what we think of as theoretical physics—the idea of describing nature in terms of equations. The gravitational force was the inspiration for this process: In order to solve the problem of gravitation, Newton invented the differential and integral calculus and introduced the three basic laws of mechanics that we know today as Newton's laws. These three laws are true foundation stones of physics.

Newton's second law describes the motion of an object of mass m that is acted on by a set of forces that add *vectorially* to a net force:

$$\vec{\mathbf{F}}_{\text{net}} = m\vec{\mathbf{a}}. \tag{1–1}$$

Here, $\vec{\mathbf{a}}$ is the acceleration of the object. For a given net force, $\vec{\mathbf{a}}$ is inversely proportional to the mass—the smaller the mass, the larger is the acceleration and vice versa. The meaning of this mass, more properly termed the **inertial mass**, is provided by Newton's second law itself. In classical physics, mass is conserved, although it can be spread around when an object breaks up or conjoined when an object is formed by amalgamation. We sometimes see the second law written in the form

$$\vec{\mathbf{F}}_{\text{net}} = \frac{d\vec{\mathbf{p}}}{dt}, \tag{1–2}$$

where the **momentum** of the object is $\vec{\mathbf{p}} = m\vec{\mathbf{v}}$. This form allows us to account for an object with a changing mass.

In order for Newton's second law to be useful, we must know something about the forces produced in different circumstances; that is, we must have **force laws.** When we do know about the forces, Newton's second law becomes a dynamical equation—an equation of motion—that allows us to predict the motion of the object in question. Some of the forces you will have studied are gravity, electrical and magnetic forces, and contact forces such as friction and the tension in a rope. An important example is the spring force, governed by the law which states that, when a spring is stretched or compressed by an amount x from its equilibrium position, it exerts a force given by $-kx$ on a mass attached to its end. Here, k is the spring constant, a quantity that varies from spring to spring. This particular force law is known as **Hooke's law**. A great deal of information is packed into Hooke's law when it is taken in conjunction with Newton's second law. The minus sign in Hooke's law tells us that the force is a restoring force, tending to move the mass back to the equilibrium position of the spring. When an object of a given mass is attached to the end of a spring and no other forces act, then the second law becomes a differential equation for the position of the mass, relating the position to its second derivative, the acceleration. In this context, we refer to the second law as the *equation of motion* of the spring, and the equation can be solved—meaning that we can find the function of time that represents the position. We will actually solve an equation of motion given by Hooke's law shortly. The importance of this particular case is that it applies to nearly any situation in which there is stable equilibrium—that is, a situation in which a slight displacement of an object from an equilibrium position produces a net force that pulls the object back to that position. But no matter how important this example is, it is only an example, one that fits into the framework of Newton's second law along with all the other examples we can imagine.

Newton's first law is superficially[1] a special case of the second law:

$$\text{If } \vec{\mathbf{F}}_{\text{net}} = \vec{0}, \text{ the motion is uniform.} \tag{1–3}$$

---

[1] We use this word because the first law is really more than a special case—it has to do with the existence of inertial frames—reference frames in which all forces have identifiable sources.

Here, when we say the motion is uniform, we mean that the acceleration is zero, or that the velocity is constant. This includes the case where the object in question is at rest. The novelty in Newton's first law is that it asserts that the role of a force is only to *accelerate* an object. Galileo's predecessors—especially Aristotle and all those who continued to quote him—assumed that it took a force to keep an object in motion. They could not imagine a frictionless world in which an object would simply keep going once it got started. This law is often used to deduce the presence, magnitude, and direction of an unknown force. For example, if a falling object reaches a constant terminal velocity, then we can deduce the magnitude of the drag force on it if we know that of the force of gravity that acts on it. Of greater significance, the first law tells us that there is no way of distinguishing between a "uniformly moving" and a "stationary" observer.

▮ The lack of distinction between observers moving relative to each other with constant velocities will play a crucial role in the development of the theory of relativity; see Chapters 2 and 3.

Gravity

We have said that when a given net force acts on an object, the acceleration of the object is inversely proportional to its mass. But we seem to have one conspicuous exception: the case in which a body falls free of all influence but that of gravity. One of the common errors of pre-Galilean thinking was the idea that heavier (more massive) objects fall more quickly (have larger accelerations) than lighter ones—something that contradicts the more careful observation that *all* objects falling under the (sole) influence of gravitation at the surface of the Earth do so with the same acceleration. Galileo claimed to have dropped objects from the Leaning Tower of Pisa to determine whether they all had the same acceleration. If he had actually done that, he would have found out that air drag spoils the experiment. As you might have learned in your introductory course, you can eliminate air drag by creating a laboratory vacuum in which a feather falls as rapidly as a penny. What, then, has happened to Newton's second law, which seems to imply that, because $\vec{a} = \vec{F}/m$, the acceleration *does* depend on the mass? The only way that the second law can be consistent is for the force law for gravitation to be proportional to the mass itself. In this way, the mass cancels from both sides of Newton's second law, and the acceleration is independent of the mass.

▮ When we discuss Einstein's theory of gravitation at the end of the book, we will see that this cancellation involves some very deep physics.

Hooke's Law

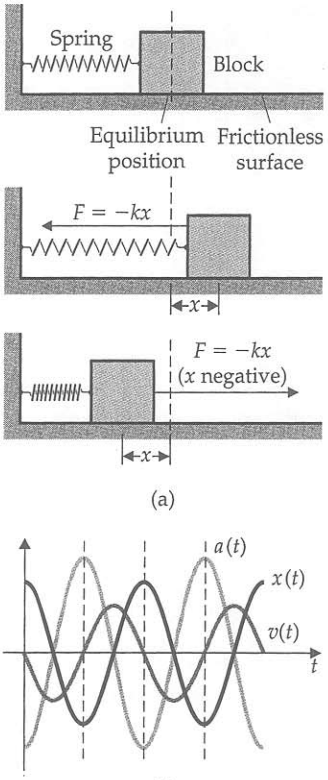
Let us examine how the equation expressing Newton's second law can be solved by reviewing an important example: the spring force, $F = -kx$ (•Fig. 1–1a). Since we are working in one dimension, vectorial aspects do not matter, and we can dispense with vector notation in the equation of motion:

$$-kx = m\frac{d^2x}{dt^2}. \tag{1–4}$$

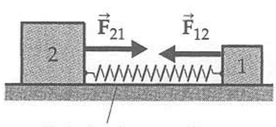
This equation relates a position, a function of time, to its second derivative. Not just any function can satisfy the equation for all time. Finding the function or functions that can do this is what we mean by solving the differential equation. The first derivative of the sine function is the cosine, and the derivative of the cosine is the sine again, but with a minus sign. Thus, the solution takes the form

$$x(t) = A\sin(\omega t + \varphi). \tag{1–5}$$

The quantities A, the *amplitude* of the motion, and φ, the *phase*, are determined by initial conditions; for example, specifying the initial position and velocity would determine A and φ. The quantity ω, on the other hand, is determined by



• Figure 1–1 (a) The force due to an idealized spring, which is proportional to the deviation of the length of the spring from its equilibrium length, leads to simple harmonic motion. (b) Plots of displacement, velocity, and acceleration as functions of time of a mass at the end of a spring.



• Figure 1–2 The force $\vec{\mathbf{F}}_{12}$ exerted on block 1 by block 2 is equal and opposite to the force $\vec{\mathbf{F}}_{21}$ exerted on block 2 by block 1.

Eq. (1–4) itself; we say that the dynamics determines ω. We find ω simply by insisting that Eq. (1–5) satisfy Eq. (1–4), with the result that

$$\omega = \sqrt{k/m}. \qquad (1\text{–}6)$$

The quantity ω, which is termed the *angular frequency*, is closely related to the repeat time, or *period* T, of the solution. The period and the *frequency* f of the repeating motion are related by

$$T = \frac{1}{f}, \qquad (1\text{–}7)$$

where $f = \omega/2\pi$. Note, finally, that by taking the first and second time derivatives of the position, Eq. (1–5), we find the velocity and acceleration, respectively (•Fig. 1–1b).

The last of Newton's three laws states that forces act between pairs of objects. That is, a force $\vec{\mathbf{F}}_{21}$ acts on object 2 *due to object 1*, then a force $\vec{\mathbf{F}}_{12}$ acts on object 1 *due to object* 2. For example, when a hand exerts a force on a piano key, the piano key exerts a force on the hand, or when Earth exerts a force on a tennis ball, the tennis ball exerts a force on Earth. **Newton's third law** states that these pairs of forces are represented by equal and opposite vectors (•Fig. 1–2):

$$\vec{\mathbf{F}}_{12} = -\vec{\mathbf{F}}_{21}. \qquad (1\text{–}8)$$

This law is sometimes called the law of equal action and reaction. Of course, that is somewhat of a misnomer if you think of the word "reaction" as a description of the motion of the objects on which forces act. When a less massive object interacts with a more massive one, the forces may be equal and opposite, but Newton's second law ensures us that the motions of the two objects will be quite different. The tennis ball reacts visibly when the force due to Earth acts on it, but the motion of Earth due to the tennis ball is not visible even to our best instruments.

The third law may be restated in terms of momentum. Imagine that you have two objects labeled 1 and 2, isolated from the outside world and interacting with each other. From Eq. (1–2), the force on object 1 is the rate of change of its momentum, $\vec{\mathbf{p}}_1$; the force on object 2 is the rate of change of $\vec{\mathbf{p}}_2$. Then Eq. (1–8) becomes

$$\frac{d\vec{\mathbf{p}}_1}{dt} = -\frac{d\vec{\mathbf{p}}_2}{dt}, \qquad \text{or} \qquad \frac{d}{dt}(\vec{\mathbf{p}}_1 + \vec{\mathbf{p}}_2) = \vec{0}.$$

Thus, the vector $\vec{\mathbf{p}}_1 + \vec{\mathbf{p}}_2$ is constant in time, a great simplification when Newton's laws are applied to what may otherwise be a complicated interaction between two objects. This result extends to more than one interacting object, taking the form

$$\vec{\mathbf{P}}_{\text{tot}} = \text{a constant vector}, \qquad (1\text{–}9)$$

where $\vec{\mathbf{P}}_{\text{tot}}$ is the total momentum. We refer to Eq. (1–9) as the law of **conservation of momentum**. It may be taken as an alternative to Newton's third law. Conservation laws play a central role in modern ideas about the operation of physical laws.

1–2 Work, Energy, and the Conservation of Energy

Energy is an extremely useful and fundamental quantity. The concept of energy was developed by early 19th-century engineers studying the operation of steam engines. Today, we can solve Newton's law numerically with computers if we know the forces involved. However, this was not true in the 18th and 19th centuries, and the search for simplifications led to a whole world of ideas encompassed under the general term "energy."

Consider a constant net force F acting in one dimension on an object of mass m. In this case the motion is entirely determined, with the velocity v changing linearly with time; that is $v \propto t$. If we eliminate the time and find the speed as a function of the distance moved, x, we obtain the relation

$$F \times (x_f - x_i) = \frac{1}{2}mv_f^2 - \frac{1}{2}mv_i^2, \tag{1–10}$$

where the initial point of the motion is labeled with the subscript i and the final point with the subscript f. It is customary to call the quantity $mv^2/2$ the **kinetic energy** K. Then the right side of Eq. (1–10) is the *change* in kinetic energy of the object. The left side is referred to as the **work** W done by the force on the object. When many forces act, each can perform work, and it is the work done by the net force, or the net work, that enters into Eq. (1–10). The work can be positive or negative, according to whether the kinetic energy increases or decreases. The sign is implicit in the definition of work and is determined by whether the force lies parallel [F and $x_f - x_i$ have the same sign] or antiparallel [F and $x_f - x_i$ have the opposite sign] to the displacement in this one-dimensional example.

Of course, not every force is constant, and not every force acts only in one dimension. The more general form of Eq. (1–10) consists in finding the definition of work suitable for the more general form of the net force. As far as forces that vary in space are concerned, we note that for a sufficiently small interval, the force can be regarded as constant, and we can sum the constant forces over all the small intervals. This is in fact, the definition of an integral, so that, for a one-dimensional *non*constant force (a force whose magnitude might vary), the work is defined as

$$W = \int_{x_i}^{x_f} F(x)\, dx. \tag{1–11}$$

For a force that acts in more than one dimension, or, what is the same thing, for motion in more than one dimension, only the component of the force that lies along the motion changes the speed, and hence the kinetic energy, and we pick out this component through the dot, or scalar, product of the force vector and the displacement vector. The result of including both effects is that a suitable definition of the work done on an object as it moves from position $\vec{\mathbf{r}}_i$ to $\vec{\mathbf{r}}_f$ is

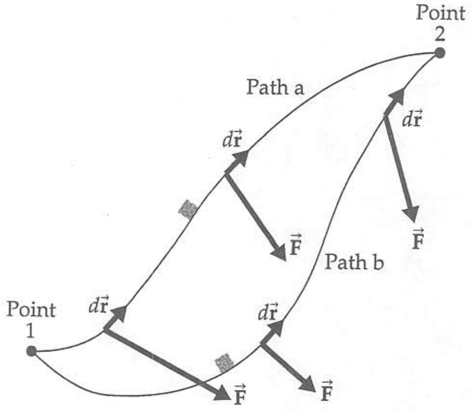
$$W = \int_{\vec{\mathbf{r}}_i}^{\vec{\mathbf{r}}_f} \vec{\mathbf{F}} \cdot d\vec{\mathbf{r}}. \tag{1–12}$$

With these definitions, Eq. (1–10) remains

$$W_{\text{net}} = \Delta K = K_f - K_i. \tag{1–13}$$

This equation is known as the **work–energy theorem.**

The work–energy theorem is useful if you are not interested in the time dependence of the motion. The equation can be applied to any force, although whether it is easy to use depends on whether the integration in Eq. (1–11) or Eq. (1–12) can be performed analytically. However, the work–energy theorem can be recast into another form for so-called conservative forces. Then it takes on a significance beyond convenience. A conservative force is a force for which the integration that expresses the work, Eq. (1–12), is independent of the path between the initial and final points. In •Fig. 1–3, we show initial and final points



• **Figure 1–3** When the work done by a force that acts on an object which moves from point 1 to a second point 2 is the same for any two paths between the points, the force is said to be conservative.

together with two paths, A and B, between them. A **conservative force** is a force for which values of the work done over the two paths—and indeed, over any paths connecting the two points—are *identical*. For such forces, the work depends only on the two endpoints:

$$\int_{\vec{r}_i}^{\vec{r}_f} \vec{F} \cdot d\vec{r} = f(\vec{r}_f) - f(\vec{r}_i).$$

We get the simplest form of the work–energy theorem if, instead of the function $f(\vec{r})$, we use its negative,

$$-\int_{\vec{r}_i}^{\vec{r}_f} \vec{F} \cdot d\vec{r} = U(\vec{r}_f) - U(\vec{r}_i). \tag{1–14}$$

With this notation, we can recast the work–energy theorem as

$$U(\vec{r}_i) + K_i = U(\vec{r}_f) + K_f. \tag{1–15}$$

This equation is in the form of a conservation law for the **total energy**, defined by

$$E \equiv U(\vec{r}) + K, \tag{1–16}$$

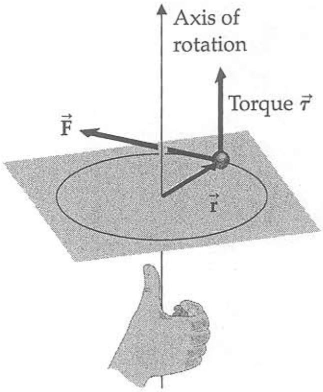
where the **potential energy**—also called the energy of position—is defined by

$$U(\vec{r}) = -\int_{\vec{r}_0}^{\vec{r}} \vec{F} \cdot d\vec{r} + U(\vec{r}_0). \tag{1–17}$$

$U(\vec{r}_0)$ is a constant that drops out of Eq. (1–14). The potential energy is thus defined only up to this constant, which plays no role in the connection between the force and the potential energy, viz., $\vec{F} = -\vec{\nabla}U(\vec{r})$, a relation discussed shortly.

A conservation principle such as the law of conservation of energy is both a central principle of physical laws and a powerful problem-solving tool. Energy conservation states that there is no change in the numerical value of E in the course of time. This principle is not affected by the presence of an arbitrary constant in the definition of U, namely, its value at an arbitrary initial point $\vec{r}_0$ in the integral. That point and the value of U there can be chosen for convenience. We usually express this arbitrariness by setting U equal to zero at some convenient point; we will see some examples later. Another thing to keep in mind is that energy is a scalar, not a vector, quantity. Of course, scalar quantities have signs. Thus, while K is positive, U and hence E can perfectly well be negative; again, only *changes* in conserved quantities matter.

What forces are conservative? In other words, what forces are associated with a potential energy, so that we can express the work–energy theorem as the



• **Figure 1–6** The cause of rotation is torque, specified by a cross product (vector product) of the position vector of the point of application of a force with the force itself. Here, too, a right-hand rule is involved.

equations for linear motion—Newton's second law—involve the linear acceleration. There is an equivalent to mass, the **rotational inertia** I, and an equivalent to force, the **torque** $\vec{\tau}$, and the relation is the same, namely,

$$\vec{\tau} = I\vec{\alpha} = \frac{d\vec{L}}{dt}, \tag{1–23}$$

where $\vec{L}$ is the **angular momentum**. The torque is expressed with respect to an axis and is given by

$$\vec{\tau} \equiv \vec{r} \times \vec{F}, \tag{1–24}$$

where $\vec{r}$ is a vector running perpendicularly from the axis of rotation to the point where the force is applied (•Fig. 1–6). The rotational inertia I is associated with the way the mass of the body is distributed. Like all the other rotational quantities that appear here, it is defined with respect to the axis of rotation. If we imagine breaking up the object in question, whose total mass is M, into many discrete portions of mass Δm_i, then

$$I \equiv \sum_i (\Delta m_i) r_i^2, \tag{1–25}$$

where r_i is the distance from the mass point labeled i to the axis of rotation. In the limit of a continuously distributed mass, the sum becomes an integration.

The work–energy theorem for rotational motion takes a form that extends the parallels we see in the dynamical equations. The work done in rotating a rigid body through an angle $\Delta\theta = \theta_f - \theta_i$ is

$$W = \int_{\theta_i}^{\theta_f} \tau \, d\theta, \tag{1–26}$$

and the kinetic energy of an object rotating with angular speed ω is

$$K = \frac{1}{2} I\omega^2 \tag{1–27}$$

The relation between these quantities is the familiar work–energy theorem, Eq. (1–13).

There is, in fact, nothing special about the axis of rotation: No physical effect depends on the choice of axis, any more than any physical effect depends on choice of origin for Newton's laws. Any axis can be employed as a reference, and the preceding relations remain true as long as the same axis is used for all of the quantities employed.

The quantities just described can be extended to nonrigid objects. To do so, we need only sum over them for a set of point objects that make up a suitable description of the entire system. We can also describe the rotational quantities with respect to an origin, not an axis. If $\vec{r}$ describes the vector from this origin to a given point object, then the angular momentum of the object is

$$\vec{L} = \vec{r} \times \vec{p}, \tag{1–28}$$

and the torque $\vec{\tau}$ remains as in Eq. (1–24) and describes the rate of change of $\vec{L}$. The angular momentum of a system is the sum of the angular momenta of the system's individual mass components. In particular, if there is no external torque acting on the system, and if the internal forces are central—directed along the lines between the internal masses—then Newton's third law assures us that *the angular momentum of the system is conserved.*

The study of rotations is a study of extended objects—that is, of mass distributions, rigid or otherwise. This study reveals the presence of a location with special properties within a mass distribution, the **center of mass**. If our object has a total mass M, the center of mass is given by

$$\vec{\mathbf{R}} = \frac{\sum_i m_i \vec{\mathbf{r}}_i}{M}. \qquad (1\text{–}29)$$

Note the important property that when a net external force acts on an extended object, the object's center of mass moves as a point mass with mass M according to Newton's second law:

$$\vec{\mathbf{F}}_{\text{net, external}} = M\vec{\mathbf{A}} \qquad (1\text{–}30)$$

where $\vec{\mathbf{A}} = d^2\vec{\mathbf{R}}/dt^2$. This is a great simplification.

1–4 Elastic Media and Waves

The collections of interacting masses that make up matter are capable of a type of collective motion that we refer to as **waves**. Waves are some of the most obvious features of our physical environment; they are also pervasive, if less obvious, in light, in sound, and within all types of matter. Because light waves have some special features—in particular, they do not require matter in which to propagate—in this section we refer to all types of waves *except* light or other electromagnetic waves. (See Section 1–8.)

Internal forces that resemble the force in Hooke's law lead to wave motion within materials. This is not so unlikely as it may sound: It really means only that there is a stable equilibrium and that the internal forces tend to lead back to that equilibrium. For example, the pressure in air is a stable quantity. A fluctuation within the air such that the pressure varies from its average is subject to forces that tend to bring it back to the average, and a *small* deformation within many materials is subject to forces that tend to remove the deformation. Let us imagine that there is some quantity h that represents a variation from equilibrium. This quantity could be the pressure in a gas, or a displacement from an equilibrium position along a string or within a material. The quantity h varies throughout the medium, as well as with time, so that $h = h(x, t)$. Here, we have let the position within the medium be represented by the single variable x, as if the medium were one dimensional, but the position could more generally be a multidimensional position vector $\vec{\mathbf{r}}$. The variable h satisfies what is called the wave equation,

$$\frac{\partial^2 h}{\partial x^2} = \frac{1}{v^2}\frac{\partial^2 h}{\partial t^2}. \qquad (1\text{–}31)$$

This equation is a direct result of applying Newton's second law to the elastic medium. The equation has the important property that it is **linear**: If we have two solutions to the equation, then the sum of those two solutions is also a solution. This is sometimes referred to as the **superposition principle**. The quantity v, whose interpretation is the **wave speed**, is a function of the internal restoring forces and of an inertial factor that describes how nimbly the medium responds to the forces:

$$v = \sqrt{\frac{\text{restoring force factor}}{\text{inertial factor}}}. \qquad (1\text{–}32)$$

For example, the speed of sound waves in a solid of mass density ρ is given by the expression $\sqrt{Y/\rho}$, where Y is Young's modulus, a quantity that determines how the solid responds to a stretching force. Analogously, the speed of a wave moving along a string of mass per unit length μ under a tension T is $\sqrt{T/\mu}$.

The solution of Eq. (1–31) is given by $h(x, t) = g(x - vt)$, where g is *any* function of the combination $x - vt$. We call the wave denoted by h a **traveling wave**. It represents some shape described by the function g moving in the x-direction with (positive or negative) velocity v. The wave could be a pulse of some kind; alternatively, it could repeat regularly. In particular, if g takes a sinusoidal form, h is known as a **harmonic wave**:

$$h(x, t) = H\sin\left[k(x - vt)\right] = H\sin(kx - \omega t). \qquad (1\text{–}33)$$

The parameters that appear here have simple interpretations. The **amplitude** of the wave—its maximum departure from equilibrium—is H, while k and ω specify the repeat length and repeat time, respectively, for the wave. In particular, if the repeat length is the **wavelength** λ, the repeat time is the period T, and the repeat frequency is $f = 1/T$, then k, termed the **wave number**, and ω, termed the **angular frequency**, are related to λ and T by

$$\omega = 2\pi f = \frac{2\pi}{T} \qquad \text{and} \qquad k = \frac{2\pi}{\lambda}. \qquad (1\text{–}34)$$

The wave speed links all these quantities:

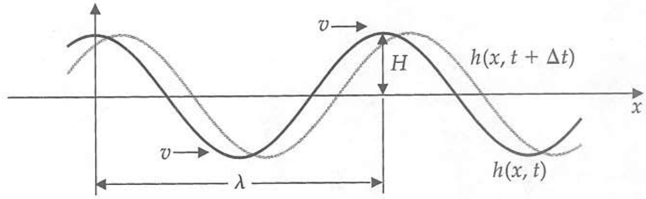
$$v = \lambda f. \qquad (1\text{–}35)$$

•Fig. 1–7 illustrates some of these points. Harmonic waves are of special importance because of Fourier's theorem, which states that any wave can be broken down into (or constructed using) harmonic waves of different frequencies. In some media, the wave speed is a function of frequency. Such media are said to be **dispersive**; the treatment of waves in dispersive media is more complicated.

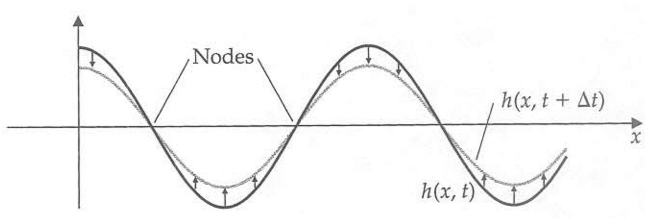
The wave equation has a second type of solution, corresponding to **standing waves**. This type of solution separates the time and space dependence:

$$h(x; t) = H\sin(kx + \varphi)\cos(\omega t + \delta). \qquad (1\text{–}36)$$

These waves oscillate in place, in contrast to traveling waves, which represent a moving disturbance (•Fig. 1–8). Note the presence of **nodes**, places where there is no disturbance. Here, the frequency $f = \omega/2\pi$ is determined by the wave equation. But what determines the wavelength? The answer is that the wavelength is determined by boundary conditions. For example, a common boundary condition is that the endpoints of a string of length L on which the



• **Figure 1–7** A harmonic wave has a repetition length called the *wavelength* λ and a maximum value of the variable it describes called the *amplitude*. The time necessary for one wavelength to pass a given point x is the *period*. The wavelength and period are connected by the wave speed.

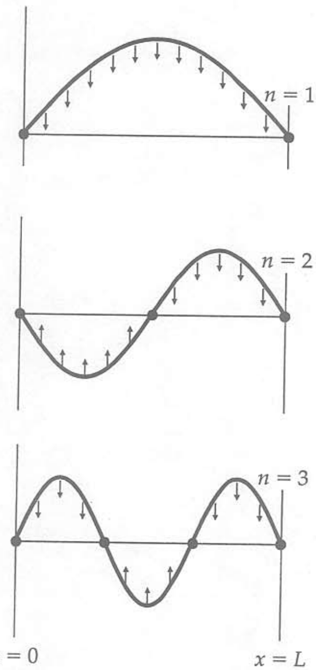


• **Figure 1–8** In harmonic standing waves, the changing variable oscillates in place. Points at which the variable vanishes are fixed in space at all times and are called *nodes*.

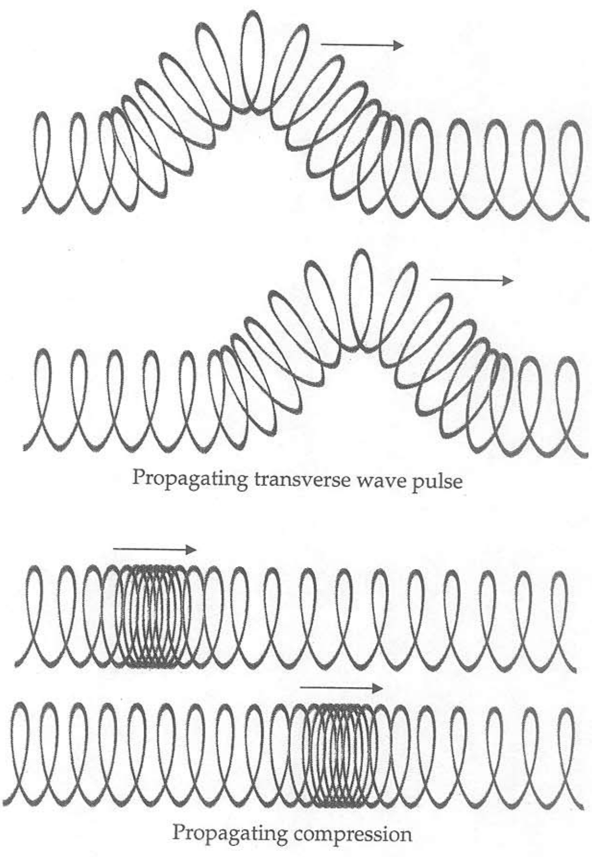
waves are set up must be fixed; that is, the quantity h must vanish at the two ends, $x = 0$ and $x = L$. In this case, $\varphi = 0$ and $kL = n\pi$; that is, $\lambda = 2L/n$, where n is a positive integer. These conditions correspond to a half-integer number of wavelengths fitting the string, as in •Fig. 1–9. The appearance of integers is often cited as unique to quantum physics; here, we see that it is perfectly possible in a classical context. It is actually typical of wavelike phenomena, and these phenomena are in many ways common to classical and quantum physics.

▌We discuss quantum numbers throughout this book.

We have not been very specific about the characteristic wave variable h, which is generally a displacement from equilibrium. That is because a large variety of physical variables satisfy wave equations, depending on the medium. There are two main categories of waves: **transverse** and **longitudinal**. This distinction depends on whether the displacement is perpendicular to, or along, the direction of propagation of the wave itself. As an example of a medium that supports both types of wave, we could take a long spring such as a Slinky™. If the spring is stretched along the x-direction, then an initial pulse formed by moving one end of the spring in the y-direction will form a propagating transverse wave, as in •Fig. 1–10a. If, instead, an initial pulse is formed by pushing the spring along its length, in the x-direction, then this pulse will propagate as a longitudinal wave, as in •Fig. 1–10b. Perhaps the most familiar longitudinal wave is that of **sound**, which is a wave in which the density of the air varies about its equilibrium value in the direction of the propagation of the sound wave.



• **Figure 1–9** For a harmonic standing wave on a string with boundary conditions that fix the ends, only certain wavelengths are allowed. This is a "quantization" phenomenon.



• **Figure 1–10** Waves can be characterized according to whether the disturbance is (a) perpendicular to the direction of the propagation, in which case the wave is transverse, or (b) along the direction of the propagation, in which case the wave is longitudinal. Sometimes a combination of the two occurs. The two types of waves are illustrated here by a Slinky™.

Power and Energy in Waves

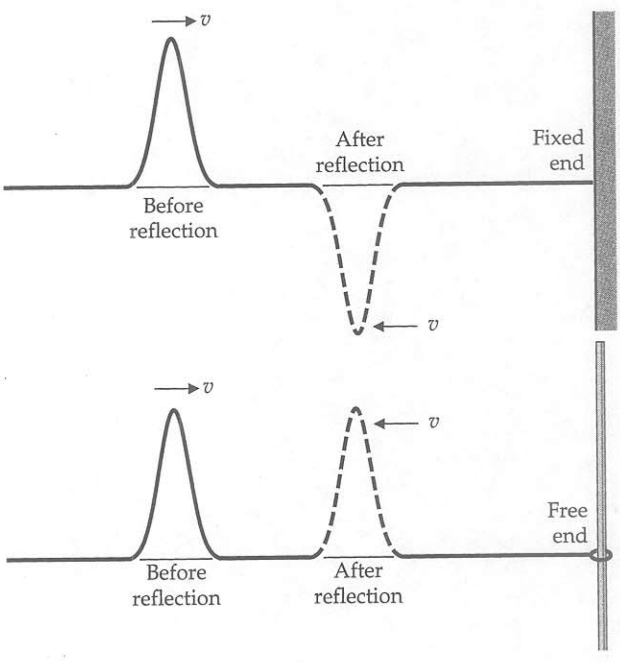
In a wave, a disturbance propagates from one place to another; there is no overall motion of the medium. Nevertheless, a traveling wave can transport the energy that is put into the formation of a pulse or into a periodic wave from one end of a medium to another. The harmonic wave of Eq. (1–33) propagating on a string of mass density μ provides the essential features: The power—that is, the rate at which energy is delivered through a unit area perpendicular to the direction of wave propagation—in this wave is

$$P = \mu v \omega^2 H^2 \cos^2(kx - \omega t).$$

The energy density in the harmonic wave is the power divided by the wave speed v. The power and energy density are themselves traveling waves, and each of these quantities is proportional to the *amplitude squared* and the *frequency squared*.

Reflection and Refraction

Waves that reach boundaries between different media reflect back into the medium upon which they were incident and refract into the medium on the far side of the boundary. (We also refer to the refracted wave as a **transmitted** wave.) For waves on one-dimensional systems such as strings, we need only note the possibility of a phase change at the boundary. Thus, a wave that has reflected from a boundary at which the string is fixed undergoes a phase change of 180°; that is, the wave is inverted (•Fig. 1–11a). If the string is free at the end, the wave is reflected upright, without a phase change (•Fig. 1–11b). In these cases no wave is transmitted, but we can arrange for one by connecting the string to another string of different mass density. In a two- or three-dimensional system (such as light), we have angles of reflection and refraction to deal with. In any case, the amplitudes of reflected and refracted waves are constrained by the requirement that energy be conserved.



• **Figure 1–11** How waves reflect depends on boundary conditions. (a) If one end of a string is fixed, an incoming pulse will be inverted on reflection. (b) If the reflecting end is free, the reflected pulse will remain upright.

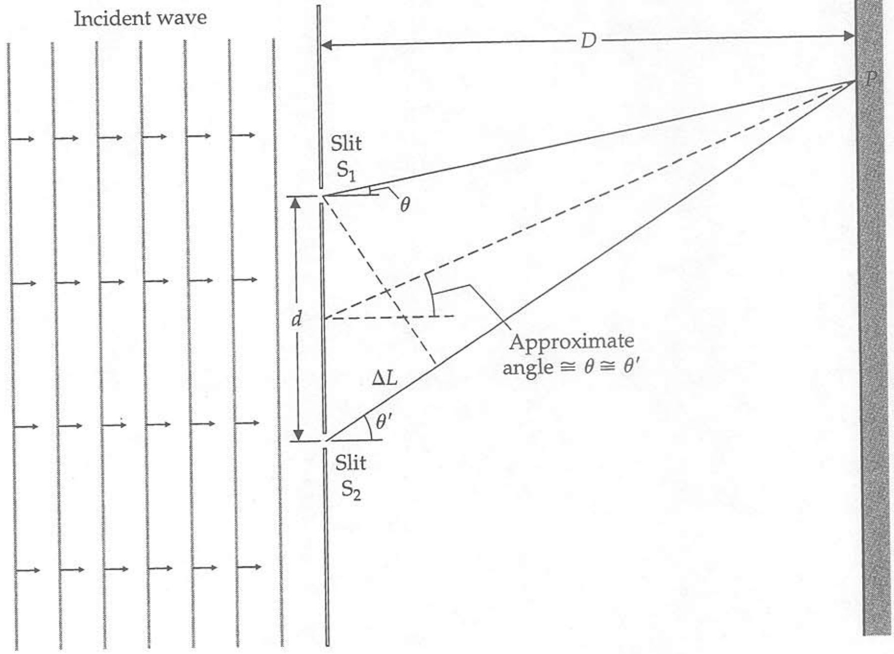
Coherence, Interference, and Diffraction

The superposition principle tells us that any sum of harmonic waves is also an acceptable solution of the wave equation. We have already remarked that this principle relates pulses to harmonic waves. In addition, a standing wave of the type in Eq. (1–36) is a superposition of two traveling waves like those of Eq. (1–33), one moving to the right and one moving to the left. Now, when two waves superpose, interesting things can happen. If, at a certain point in space and time, there are two waves with h-values that are equal and opposite, then at that point the net displacement cancels; we say the waves interfere destructively. If, in contrast, we have two waves with h-values that are equal and of the same sign, then at that point in space the two waves interfere constructively. An interference pattern between two or more waves can be set up throughout space and time, and this pattern can be regular if the two waves are themselves regular. We say that harmonic waves are **coherent** if there is a definite relation between their frequencies and phases. The possibility of coherent waves can be realized in various ways. Let us enumerate several of the typical patterns that result.

- *Beats.* Imagine two interfering waves of slightly different angular frequencies ω_1 and ω_2 propagating in the same medium, so that they have the same wave speed. (For sound, this can easily be arranged with two slightly different tuning forks.) Then if they each have the same amplitude, their algebraic sum takes the form

$$H\sin(k_1x - \omega_1t) + H\sin(k_2x - \omega_2t) = 2H\sin(Kx - \Omega t)\cos\left(\frac{\delta k}{2}x - \frac{\delta\omega}{2}t\right),$$

  where K is the average of k_1 and k_2, Ω is the average of ω_1 and ω_2, δk is the difference of k_1 and k_2, and $\delta\omega$ is the difference of ω_1 and ω_2. The interfering waves produce the product of two waves, one with a very small wave number $\delta k/2$—that is, a very long wavelength—and a very small angular frequency. The part with the small frequency is termed the **beat**.
- *Interference patterns in space with two sources.* •Figure 1–12 shows one way to set up an interference pattern in space. A single wave is sent through two slits that act as two sources of the same frequency. A point P is located at different distances from the two slits, and hence there is a definite phase difference between the waves at that point. The two waves have the same amplitude and wavelength λ, but because one has to travel farther than the other to get to point P, they arrive at P with different phases. Depending on the difference in path length, ΔL, the two waves could interfere constructively or destructively. The condition for constructive interference is that $\Delta L = n\lambda$, while for destructive interference, the condition is $\Delta L = (n + 1/2)\lambda$. In each case, n is either zero or a positive or negative integer. To find how these conditions correspond to a particular set of points in space, an exercise in geometry is necessary. For instance, •Fig. 1–12 shows a situation in which the distance D from the two sources to an observing screen is much larger than the distance d between the sources. In that case, the light rays leaving from S_1 and S_2 are nearly parallel, and we have $\theta \approx \theta'$. Thus, the difference between the two path lengths is given approximately by $d\sin\theta$, and the condition for fully constructive interference becomes $d\sin\theta = n\lambda$, where n is any integer, positive or negative. A pattern of maxima and minima is the result. Another way to get two coherent sources is through the use of split



• **Figure 1–12** An arrangement for producing an interference pattern. The two apertures produce two coherent waves whose disturbances systematically add or subtract on a screen, here assumed to be distant compared with the separation between the sources. The waves could be sound, light, or water waves in a ripple tank. For $L \gg d, \theta' \cong \theta$.

beams that are partially reflected and partially transmitted and then are rejoined by means of mirrors.

- *Gratings* involve spatial interference patterns similar to those produced by two sources. However, in a grating, N sources are used, and if N is very large, it has the effect of greatly sharpening the interference pattern. Gratings are particularly important when the wave involved is light. (See Section 1–8.)

Christian Huygens demonstrated how a wave front is due to interference. He pictured the propagation of a wave front as a continual regeneration of "wavelets" along the front. The straight-line propagation of the front is due to the constructive interference of the wavelets all along the front. If, on the other hand, the front is broken by, say, the presence of a barrier, then the constructive interference is no longer present on the far side of the barrier, and the wave front will bend around the edge of the barrier. This phenomenon is known as **diffraction**.

The Doppler Shift

This phenomenon describes how the movement of the receiver or emitter of a wave or the medium itself affects the frequency or wavelength of the wave. The Doppler shift plays an important role in relativity; accordingly, we shall reserve treatment of it in some detail in that discussion.

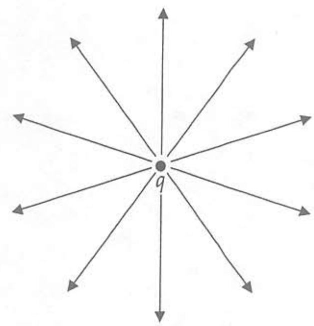
1–7 Electricity and Magnetism

The story of the discovery of the laws of electromagnetism, in the nineteenth century, is one of the glorious tales of science. A full recounting of this history would take us too far afield; but if you are not familiar with it, we cannot recommend too strongly that you take the trouble to read about it.[4]

Electrical forces occur between stationary electric charges; **magnetic forces** occur between moving electric charges, or, equivalently, **electric currents**. If one maps the forces acting on a tiny test charge in the presence of a given distribution of charges, moving or otherwise, and then divides out the effects of the test charge itself, one is left with something that depends on the distribution. As the test charge moves through space, it traces the **electric and magnetic fields** that are due to the original distribution. These fields are vectors—the forces have directions—that have a value at every point in space. We label the electric field $\vec{\mathbf{E}}(\vec{\mathbf{r}})$ and the magnetic field $\vec{\mathbf{B}}(\vec{\mathbf{r}})$. As a first approximation, electric charges are associated with electric fields, and electric currents are associated with magnetic fields. (The direction of positive current is, by definition, the direction of the movement of positive charges. We emphasize, however, that you can have a current in an electrically neutral situation, as long as equal amounts of positive and negative charge are present—all that is necessary is that *one* of the two charge components be moving. This is the situation in real wires. In addition, note that a current moving to the right could correspond to positive charges moving to the right or negative charges moving to the left.) As soon as a dependence on time is included, the sources of electric and magnetic forces become more complicated.

The electric field, at least as it is set up by electric charges (•Fig. 1–18), is associated with conservative forces. Therefore, there is a potential energy, a function of position, for these electric forces. This potential energy is that of a (test) charge in the presence of some given distribution of charge that sets up the field, and by dividing out the test charge, we are left with something that depends only on the charge distribution, just as the electric field does. We refer to this quantity as the **electric potential**, and it is given by

$$V(\vec{\mathbf{r}}) = -\int_{\vec{\mathbf{r}}_0}^{\vec{\mathbf{r}}} \vec{\mathbf{E}} \cdot d\vec{\mathbf{r}} + V(\vec{\mathbf{r}}_0). \tag{1–51}$$



$\vec{\mathbf{E}} = \hat{\mathbf{r}}\dfrac{q}{4\pi\epsilon_0 r^2}$

• Figure 1–18 The electric field due to a positive point charge points away from the charge and falls as $1/r^2$, where r is the distance from the charge.

As with potential energy, we are free to choose the location $\vec{\mathbf{r}}_0$ where the potential is zero, typically at infinity. Generally speaking, the potential is positive near positive charges and negative near negative charges.

The fields represent more than a simple way to summarize the forces: The fundamental laws of electricity and magnetism can be formulated in terms of them. The most important of these laws are known as **Maxwell's equations**, after James Clerk Maxwell, who formulated them in 1867. The four Maxwell equations describe the behavior of electric and magnetic fields, which are associated with electric charges and electric currents. The equations are as follows:

1. Gauss' law for electric fields:

$$\iint_{\text{closed surface}} \vec{\mathbf{E}} \cdot d\vec{\mathbf{A}} = \frac{Q}{\varepsilon_0}. \tag{1–52}$$

Here, Q is the net electric charge (the algebraic sum of the positive and negative charges) contained within the closed surface over which the electric field is integrated, while ε_0 is a constant called the **permittivity of free space**. This constant is associated with the units of charge. The meaning of the surface integration is as follows: $d\vec{\mathbf{A}}$ is a surface element that is oriented outward from the closed surface, perpendicular to it. The integral is known as the **electric flux** through the surface; although the flux is in this case over a closed surface, it is a concept that has meaning even when the surface is not closed. While Gauss' law holds even when a dependence on time is present, it is useful to note that in static situations it is equivalent to **Coulomb's law** for the force between charges, which, in the context of the electric field and for a point charge q, takes the form

$$\vec{\mathbf{E}} = \frac{q}{4\pi\varepsilon_0 r^2}\hat{\mathbf{r}}. \tag{1–53}$$

Here, r is the distance from the point charge. (Recall that the force on a test charge is the electric field times the magnitude of the charge.) From Coulomb's law, we note that electric fields point away from positive charges and toward negative charges.

2. **Gauss' law for magnetic fields:**

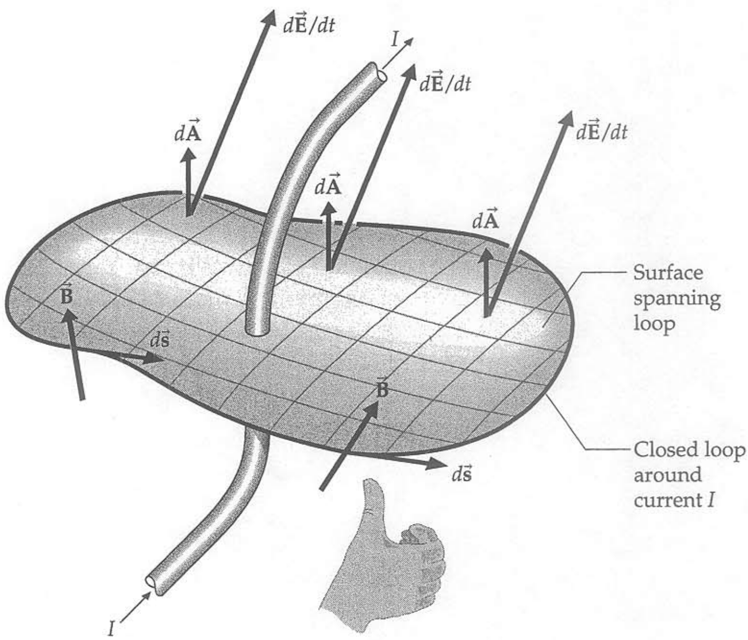
$$\iint_{\text{closed surface}} \vec{\mathbf{B}} \cdot d\vec{\mathbf{A}} = 0. \tag{1–54}$$

In this case, it is the **magnetic flux** through a closed surface that is involved. The fact that it is zero means that magnetic fields trace out closed loops. As far as we know, there are no analogues of electric charges for magnetic fields. Although we know of no deep reason that forbids such analogues, called magnetic monopoles, they simply have never been detected.

3. The **generalized Ampère's law** relates magnetic fields to the currents and changing electric fields that produce them by the equation

$$\oint \vec{\mathbf{B}} \cdot d\vec{\mathbf{s}} = \mu_0 I + \mu_0 \varepsilon_0 \frac{d}{dt} \iint_{\text{surface}} \vec{\mathbf{E}} \cdot d\vec{\mathbf{A}}. \tag{1–55}$$

The left side of this equation is a line integral around a closed loop (•Fig. 1–19). If one breaks the closed loop into tiny segments, then the line element $d\vec{\mathbf{s}}$ is a vector of length ds pointing in the direction followed by the loop. This direction, clockwise or counterclockwise, sets information for the quantities on the right-hand side of the equation. The quantity I is the electric current—the rate at which charge passes—through the closed loop. The second term on the right contains a surface integral that is the electric flux through the surface—any surface—that spans the loop. This term is known as the **Maxwell displacement current** term, after Maxwell, who noted that the meaning of the electric current passing through the loop is ambiguous without this second term. The direction of positive current, or of positive



• **Figure 1–19** Ampère's law relates the integral of the magnetic field around a closed loop to the current through the loop and the rate of change of the electric flux crossing a surface spanning the loop. A right-hand rule is involved.

flux (meaning positive area elements $d\vec{\mathbf{A}}$ in the surface integral), is defined by a right-hand rule: Curl the fingers of the right hand in the direction in which the loop is followed, and the right thumb defines the positive direction. Finally, a new constant, μ_0, appears here, the **permeability of free space**. This constant is associated with units of current and the strength of magnetic forces. The generalized Ampere's law pinpoints the origins of magnetic fields: currents or changing electric fields.

4. **Faraday's law:**

$$\oint \vec{\mathbf{E}} \cdot d\vec{\mathbf{s}} = -\frac{d}{dt} \iint_{\text{surface}} \vec{\mathbf{B}} \cdot d\vec{\mathbf{A}}. \tag{1–56}$$

Here, as in the generalized Ampère's law, a loop integral occurs, this time over the electric field; conventions about directions are the same as for Ampère's law. On the right-hand side, we see the rate of change of magnetic flux for any surface spanning the loop that appears on the left-hand side. Faraday's law describes how a changing magnetic flux can generate an electric field, as do electric charges (by Gauss' law). This electric field is quite different from the field generated by charges, however, because it can form closed loops, while the electric field generated by charges must begin or end on those charges. The electric field described by Eq. (1–54) generates its own secondary magnetic field by the generalized Ampère's law. The minus sign that appears in Eq. (1–56) implies that the secondary magnetic field has a magnetic flux that tends to oppose the original change in the magnetic flux; in this form, the law is also known as **Lenz's law**. Note the similarity between Faraday's law and the generalized Ampère's law. If magnetic monopoles existed, then one might expect a term in Faraday's law analogous to the electric charge in Ampère's law, representing the movement of magnetic charges.

To the four Maxwell equations, we add the force laws that describe how electric and magnetic fields affect charges and currents. The force on a charge q moving with velocity $\vec{\mathbf{v}}$ in an electric field $\vec{\mathbf{E}}$ and a magnetic field $\vec{\mathbf{B}}$ is given by the **Lorentz force law:**

$$\vec{\mathbf{F}} = q(\vec{\mathbf{E}} + \vec{\mathbf{v}} \times \vec{\mathbf{B}}). \tag{1–57}$$

Recognizing that a current is no more than moving charges, we can sometimes write the velocity-dependent part of this expression as a force on a length of wire $d\vec{\ell}$ carrying a current element in a magnetic field. We then have

$$d\vec{\mathbf{F}} = I d\vec{\ell} \times \vec{\mathbf{B}}. \tag{1–58}$$

The net magnetic force on a wire of finite length is found by integrating elements like those of Eq. (1–58).

By adding (integrating) the magnetic force on a succession of wire elements, we can find forces on finite wire segments. A particularly important example is the case of a current loop (•Fig. 1–20). If such a loop is placed in a uniform magnetic field $\vec{\mathbf{B}}$, there is no net force, but there is a torque that tends to twist the loop. This torque is given by

$$\vec{\boldsymbol{\tau}} = \vec{\boldsymbol{\mu}} \times \vec{\mathbf{B}}. \tag{1–59}$$

The magnetic moment $\vec{\boldsymbol{\mu}}$ has magnitude IA, where I is the current carried in the loop and A is the area of the loop. The direction of $\vec{\boldsymbol{\mu}}$ is defined by a right-hand rule, as in •Fig. 1–19. Thus, the torque tends to rotate the loop so that flux (meaning positive area elements $d\vec{\mathbf{A}}$ in the surface integral), is defined by a right-hand rule: Curl the fingers of the right hand in the direction in which the loop is followed, and the right thumb defines the positive direction. Finally, a new constant, μ_0, appears here, the **permeability of free space**. This constant is associated with units of current and the strength of magnetic forces. The generalized Ampere's law pinpoints the origins of magnetic fields: currents or changing electric fields.

4. **Faraday's law:**

$$\oint \vec{\mathbf{E}} \cdot d\vec{\mathbf{s}} = -\frac{d}{dt} \iint_{\text{surface}} \vec{\mathbf{B}} \cdot d\vec{\mathbf{A}}. \tag{1–56}$$

Here, as in the generalized Ampère's law, a loop integral occurs, this time over the electric field; conventions about directions are the same as for Ampère's law. On the right-hand side, we see the rate of change of magnetic flux for any surface spanning the loop that appears on the left-hand side. Faraday's law describes how a changing magnetic flux can generate an electric field, as do electric charges (by Gauss' law). This electric field is quite different from the field generated by charges, however, because it can form closed loops, while the electric field generated by charges must begin or end on those charges. The electric field described by Eq. (1–54) generates its own secondary magnetic field by the generalized Ampère's law. The minus sign that appears in Eq. (1–56) implies that the secondary magnetic field has a magnetic flux that tends to oppose the original change in the magnetic flux; in this form, the law is also known as **Lenz's law**. Note the similarity between Faraday's law and the generalized Ampère's law. If magnetic monopoles existed, then one might expect a term in Faraday's law analogous to the electric charge in Ampère's law, representing the movement of magnetic charges.

To the four Maxwell equations, we add the force laws that describe how electric and magnetic fields affect charges and currents. The force on a charge q moving with velocity $\vec{\mathbf{v}}$ in an electric field $\vec{\mathbf{E}}$ and a magnetic field $\vec{\mathbf{B}}$ is given by the **Lorentz force law**:

$$\vec{\mathbf{F}} = q(\vec{\mathbf{E}} + \vec{\mathbf{v}} \times \vec{\mathbf{B}}). \tag{1–57}$$

Recognizing that a current is no more than moving charges, we can sometimes write the velocity-dependent part of this expression as a force on a length of wire $d\vec{\ell}$ carrying a current element in a magnetic field. We then have

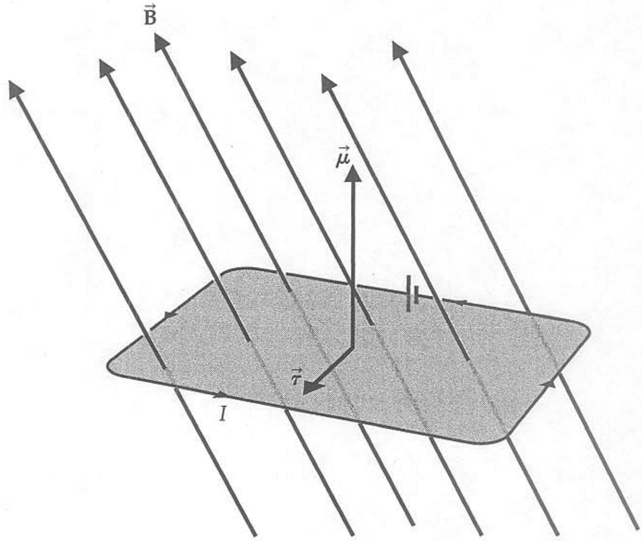
$$d\vec{\mathbf{F}} = I d\vec{\ell} \times \vec{\mathbf{B}}. \tag{1–58}$$

The net magnetic force on a wire of finite length is found by integrating elements like those of Eq. (1–58).

By adding (integrating) the magnetic force on a succession of wire elements, we can find forces on finite wire segments. A particularly important example is the case of a current loop (•Fig. 1–20). If such a loop is placed in a uniform magnetic field $\vec{\mathbf{B}}$, there is no net force, but there is a torque that tends to twist the loop. This torque is given by

$$\vec{\boldsymbol{\tau}} = \vec{\boldsymbol{\mu}} \times \vec{\mathbf{B}}. \tag{1–59}$$

The magnetic moment $\vec{\boldsymbol{\mu}}$ has magnitude IA, where I is the current carried in the loop and A is the area of the loop. The direction of $\vec{\boldsymbol{\mu}}$ is defined by a right-hand rule, as in •Fig. 1–19. Thus, the torque tends to rotate the loop so that



• **Figure 1–20** A current loop placed in a magnetic field undergoes a torque. In the figure, the loop is placed at an angle to a constant magnetic field. The magnetic moment points in a direction perpendicular to the loop, and the torque points in a direction that lies in the plane and is perpendicular to the magnetic field.

the magnetic moment is parallel to $\vec{\mathbf{B}}$. When it is placed in the field, the loop has a potential energy that is a minimum when $\vec{\boldsymbol{\mu}}$ and $\vec{\mathbf{B}}$ are aligned, namely,

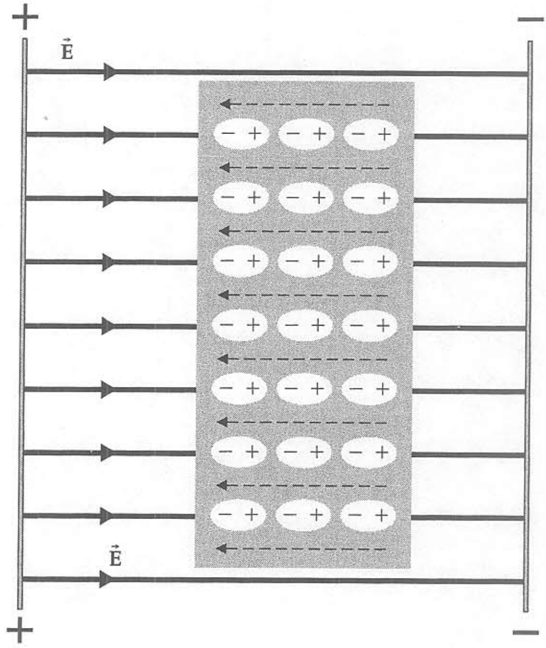
$$U = -\vec{\boldsymbol{\mu}} \cdot \vec{\mathbf{B}}. \tag{1–60}$$

The fact that electric and magnetic fields are associated with forces means that they are also associated with energy. In other words, *the fields themselves contain energy*. In a region of space with permittivity ε_0 and permeability μ_0 where the electric field has magnitude E and the magnetic field has magnitude B, the energy density, or energy per unit volume, is

$$u = \frac{1}{2\mu_0} B^2 + \frac{1}{2} \varepsilon_0 E^2. \tag{1–61}$$

The presence of matter has an effect on Maxwell's equations that can be accounted for in a direct way. The mixtures of moving charges that constitute matter react to the presence of fields and modify those fields. How they do so depends on how the atoms form that matter. For our purposes, we can divide materials into several classes. **Insulators** have the property that electric charges—either those supplied from within the atoms that make up the material or those introduced from the outside—cannot move through them very easily. At most, the charges within individual atoms can move from side to side within the atom. **Conductors** have the property that, in effect, electrons move freely within them. These materials are bulk metals; the electrons, which are also known as *valence electrons*, come from the very atoms that make up the conductor. **Semiconductors** lie somewhere between these two groups.

For insulators in the presence of an external electric field, the atoms align so that the electric field within the material is diminished (•Fig. 1–21). The effects of insulators, also known as dielectrics, can be summarized by replacing ε_0 by $\varepsilon = \kappa\varepsilon_0$, where κ, the *dielectric constant*, is greater than unity. Conductors react much more dramatically: The charges within them move until the electric field within is canceled, at least if the field is not dependent on time. Materials also can be classified according to their magnetic properties. The constant μ_0 is replaced by a modified constant μ in magnetic materials. The most important class of magnetic materials is known as **ferromagnetic**; these materials pro-



• **Figure 1–21** When an insulator is placed in an external electric field, dipoles align to produce a field that partially cancels the external field, so that the net field within the insulator is less than the external field.

duce a magnetic field on their own. The source of this field is the alignment of atoms, which, by virtue of the current loops formed by the circulating charge of the atomic electrons, themselves act like tiny magnets.

Other electric properties of materials are of importance in understanding electric circuits. *Ohm's law* describes how currents in a piece of conducting material are formed when an electric field is maintained across the material by a battery or its equivalent:

$$V = IR. \tag{1–62}$$

Here, V is the potential difference from one end of the wire to another. The proportionality constant R is the resistance, and when it is independent of V, we say that Ohm's law applies, or that the material is ohmic.

1–8 Electromagnetic Waves and Light

Taken together, Maxwell's equations imply two connected wave equations, one for the electric field and one for the magnetic field. As we shall see shortly, Maxwell recognized that the waves described by these equations had the right speed to represent light, and this understanding, together with the German physicist Heinrich Hertz's generation and detection of these waves in a direct electromagnetic context, closed a major chapter in the history of science.

The 17th century was rich in optical discoveries, beginning with those of Willebrord Snell, who quantified **refraction**, the bending of light as it propagates from one medium to another. Somewhat later, Francisco Grimaldi observed—and named—the phenomenon of diffraction, the bending of light as it goes past a barrier. Diffraction is manifested in the fact that light going through a narrow slit is spread on the other side and that a shadow has no perfect edge. This phenomenon had also been noted by Robert Hooke, who suspected that it might indicate the wave nature of light. In the second half of the century, Olaf Roemer made the first quantitative determination of the speed of light. He compared the calculated and the observed times of the eclipses of Jupiter's satellites.

The difference between these two times was due to the fact that light does not propagate with an infinite speed. We see the light after an event—the eclipse in this case—has taken place.

By the first half of the 17th century, Pierre de Fermat had conjectured the correct law of light propagation, namely, that light propagates between two points along the path that minimizes the time it takes to make the trip. Some decades later, Newton and Huygens each elaborated rather different ideas about the nature of light, although Newton's ideas were more subtle and complex than his followers and successors acknowledged. They assumed that Newton had in mind a simple "particle" theory of light, in which indivisible "atoms" of light moved in straight lines and were subject to the effects of gravitation. But Newton also investigated the colors seen in thin layers, such as those of soap bubbles. He noted that if one placed a slightly convex lens on a flat piece of glass such that the surfaces were not quite in contact, brilliantly colored rings, called *Newton's rings*, would be observed. This did not seem explicable by the propagation of particles, and indeed, Newton referred to light propagation in this effect as being in "fits" of reflection and transmission, which one might take to describe the motion of a longitudinal wave like a sound wave. Newton also knew about certain polarization phenomena—the fact that certain crystals will transmit light with various intensities, depending on how their axes are oriented—but he attributed these phenomena to the notion that his atoms of light had particular shapes.

At about the same time that Newton was doing his work, Christian Huygens developed a wave picture of light. The full generality of his construction was exploited only in the beginning of the 19th century, but Huygens was able to explain refraction using it. The idea is that when the leading edge of the wave hits, say, the denser medium, it is slowed down, and the wave is turned around this point as the faster moving portion of it catches up. This is a phenomenon you can observe as water waves hit a beach.

During the next century, the particle theory became the dominant theory of light propagation. Nonetheless, by the beginning of the 19th century, the next great period in the history of optical discovery, the particle theory of light had been swept aside in favor of the wave theory, chiefly due to the work of Thomas Young and Augustin Fresnel. The latter used the Huygens construction to give a complete theory of diffraction, including the discovery that in the center of any circular shadow there is a bright spot.

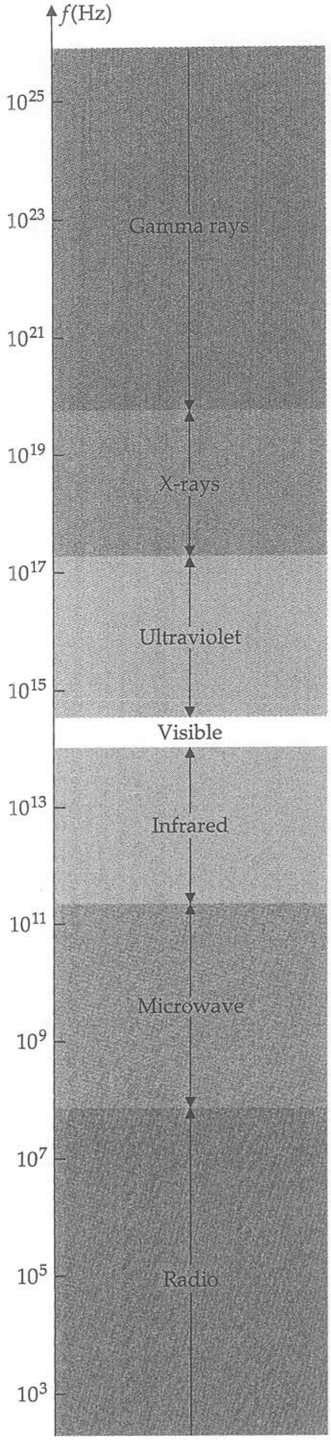
In Young's most famous experiment, early in the 19th century, the distance between two pinholes that supply coherent light that can interfere on a screen was set to be about a millimeter. The screen was about a meter away from the pinholes. As the pinholes are made smaller, the two patches of light they make on the screen get larger due to diffraction, and eventually they will overlap. At this point, an array of light and dark bands—the interference pattern of the two sources—appears. Apparently, these phenomena could be explained only if light was described by a wave. Using the conditions for constructive interference described in Section 1–4, Young could even determine the wavelength of the light. Such wavelengths λ are quite small by ordinary standards, on the order of several hundred nanometers.

This historical diversion brings us back to Maxwell's equations. For the electric field, the presence of a set of charges oscillating in the x-direction allows us to generate a wave equation of the form

$$\frac{\partial^2}{\partial z^2} E_x = \mu_0 \varepsilon_0 \frac{\partial^2}{\partial t^2} E_x. \tag{1–63}$$

If there are no mechanisms in these media that destroy the waves, then the media are transparent, and in line with what we stated in Section 1–7, the only change is that ε_0 is replaced by ε and μ_0 is replaced by μ. This means that the speed of light changes to $c' = c/n$, where the quantity n is the **index of refraction** of the material. Transparent materials have $\mu \approx \mu_0$, so that $n \approx \sqrt{\kappa}$, where κ is the dielectric constant, a quantity greater than unity. The reduced speed of waves in such media neatly explains refractive effects.

- The fact that the phase of the magnetic and electric fields in these waves is the same is a general feature. When the electric field is a maximum, so is the magnetic field; when one field is a minimum, so is the other.
- An additional restriction on the relative size of E_0 and B_0, and hence on the size of E and B, in electromagnetic waves is that $E = cB$.
- The electric and magnetic fields are perpendicular to the direction of propagation of the wave and to each other, so that electromagnetic waves are *transverse*. This is also a general feature.



• **Figure 1–23** The electromagnetic spectrum. All these waves have the same structure, determined by Maxwell's equations. Only the wavelengths differ.

Energy and Momentum Transport

Equation (1–61) shows that there is energy in electromagnetic waves, and the energy density in a region where they propagate is

$$u = \frac{1}{2}\varepsilon_0\left(E^2 + c^2B^2\right) = \varepsilon_0 E^2. \qquad (1\text{–}67)$$

Here we have used the relation between E and B in the wave, a relation that tells us that the energy density is equally split between the electric and the magnetic components of the wave. Moreover, the energy density is itself a propagating wave—it contains the factor $\cos^2(kz - \omega t + \varphi)$, which corresponds to something propagating with speed c. In other words, electromagnetic waves transport energy, and the speed of transport is the speed of light. The rate at which this energy arrives at a surface perpendicular to the direction of propagation of the waves is the **energy flux**, and it is given by the product cu. The energy flux can be more fully characterized by a vector that is along the direction of propagation—that is, proportional to $\vec{\mathbf{E}} \times \vec{\mathbf{B}}$—and of magnitude cu. This vector is known as the **Poynting vector**,

$$\vec{\mathbf{S}} = (\vec{\mathbf{E}} \times \vec{\mathbf{B}})/\mu_0. \qquad (1\text{–}68)$$

Another quantity used to measure energy transport is the **intensity** I, defined as the time average of the energy flux. Because the average of the cosine squared is $1/2$, $I = S/2$.

The fields of an electromagnetic wave exert a net force in the direction of propagation on a charge of any sign. This shows that the wave also transports momentum to the charge when the wave hits it. The momentum per unit volume, or momentum density, carried by the wave is $\vec{\mathbf{S}}/c^2$. One consequence of this fact is a pressure is exerted on a material when an electromagnetic wave is incident on it. This *radiation pressure* is given by $2u$ for a perfectly reflecting surface and by u for a perfectly absorbing surface.

Polarization

Although the electric field lies in a plane perpendicular to the direction of propagation of an electromagnetic wave (•Fig. 1–22), there is nothing that picks a direction in that plane. The particular direction chosen specifies the **polarization** of the wave. It is possible to measure this direction, as well as to "filter" the waves so that only certain directions are passed out of a beam that consists of