CHAPTER

7

Dielectric Materials and Insulation

The familiar parallel plate capacitor equation with free space as an insulator is given by

$$C = \frac{\varepsilon_o A}{d}$$

where ε_o is the absolute permittivity, A is the plate area, and d is the separation between the plates. If there is a material medium between the plates, then the capacitance, the charge storage ability per unit voltage, increases by a factor of ε_r, where ε_r is called the **dielectric constant** of the medium or its **relative permittivity.** The increase in the capacitance is due to the **polarization** of the medium in which positive and negative charges are displaced with respect to their equilibrium positions. The opposite surfaces of the dielectric medium acquire opposite surface charge densities that are related to the amount of polarization in the material. An important concept in dielectric theory is that of an **electric dipole moment** p, which is a measure of the electrostatic effects of a pair of opposite charges $+Q$ and $-Q$ separated by a finite distance a, and so is defined by

$$p = Qa$$

Although the net charge is zero, this entity still gives rise to an electric field in space and also interacts with an electric field from other sources. The relative permittivity is a material property that is frequency dependent. Some capacitors are designed to work at low frequencies, whereas others have a wide frequency range. Furthermore, even though they are regarded as energy storage devices, all practical capacitors exhibit some losses when used in an electric circuit. These losses are no different than I^2R losses in a resistor carrying a current. The power dissipation in a practical capacitor depends on the frequency, and for some applications it can be an important factor. A defining property of a dielectric medium is not only its ability to increase capacitance but also, and equally important, its insulating behavior or low conductivity so that the charges are not conducted from one plate of the capacitor to the other through the dielectric. Dielectric materials often serve to insulate current-carrying conductors or conductors at different voltages. Why can we not simply use air as insulation between

high-voltage conductors? When the electric field inside an insulator exceeds a critical field called the **dielectric strength,** the medium suffers dielectric breakdown and a large discharge current flows through the dielectric. Some 40 percent of utility generator failures are linked to insulation failures in the generator. Dielectric breakdown is probably one of the oldest electrical engineering problems and that which has been most widely studied and never fully explained.

7.1 MATTER POLARIZATION AND RELATIVE PERMITTIVITY

7.1.1 RELATIVE PERMITTIVITY: DEFINITION

We first consider a parallel plate capacitor with vacuum as the dielectric medium between the plates, as shown in Figure 7.1a. The plates are connected to a constant voltage supply V. Let Q_o be the charge on the plates. This charge can be easily measured. The capacitance C_o of the parallel plate capacitor in free space, as in Figure 7.1a, is defined by

Definition of capacitance

$$C_o = \frac{Q_o}{V} \qquad [7.1]$$

The electric field, directed from high to low potential, is defined by the gradient of the potential $\mathcal{E} = -dV/dx$. Thus, the electric field $\mathcal{E}$ between the plates is just V/d where d is the separation of the plates.

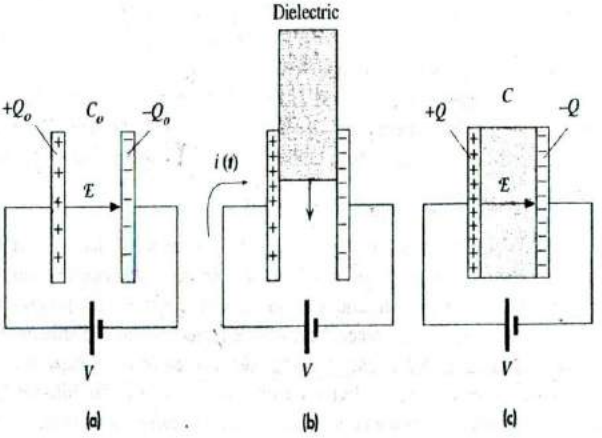


Figure 7.1

(a) Parallel plate capacitor with free space between the plates.

(b) As a slab of insulating material is inserted between the plates, there is an external current flow indicating that more charge is stored on the plates.

(c) The capacitance has been increased due to the insertion of a medium between the plates.

Consider now what happens when a dielectric slab (a slab of any nonconducting material) is inserted into this parallel plate capacitor, as shown in Figure 7.1b and c with V kept the same. During the insertion of the dielectric slab, there is an external current flow that indicates that there is additional charge being stored on the plates. The charge on the electrodes increases from Q_o to Q. We can easily measure the extra charge $Q - Q_o$ flowing from the battery to the plates by integrating the observed current in the circuit during the process of insertion, as shown in Figure 7.1b. Because there is now a greater amount of charge stored on the plates, the capacitance of the system in Figure 7.1c is larger than that in Figure 7.1a by the ratio Q to Q_o. The **relative permittivity** (or the **dielectric constant**) ε_r is defined to reflect this increase in the capacitance or the charge storage ability by virtue of having a dielectric medium. If C is the capacitance with the dielectric medium as in Figure 7.1c, then by definition

Definition of relative permittivity

$$\varepsilon_r = \frac{Q}{Q_o} = \frac{C}{C_o} \qquad [7.2]$$

The increase in the stored charge is due to the polarization of the dielectric by the applied field, as explained below. It is important to remember that when the dielectric medium is inserted, the electric field remains unchanged, provided that the insulator fills the whole space between the plates as shown in Figure 7.1c. The voltage V remains the same and therefore so does the gradient V/d, which means that $\mathcal{E}$ remains constant.

7.1.2 DIPOLE MOMENT AND ELECTRONIC POLARIZATION

An electrical dipole moment is simply a separation between a negative and positive charge of equal magnitude Q as shown in Figure 7.2. If $\mathbf{a}$ is the vector from the negative to the positive charge, the **electric dipole moment** is defined as a vector by

Definition of dipole moment

$$\mathbf{p} = Q\mathbf{a} \qquad [7.3]$$

The region that contains the $+Q$ and $-Q$ charges has zero net charge. Unless the two charge centers coincide, this region will nonetheless, by virtue of the definition in Equation 7.3, contain a dipole moment.

The net charge within a neutral atom is zero. Furthermore, on average, the center of negative charge of the electrons coincides with the positive nuclear charge, which means that the atom has no net dipole moment, as indicated in Figure 7.3a. However, when this atom is placed in an external electric field, it will develop an induced dipole moment. The electrons, being much lighter than the positive nucleus, become easily displaced by the field, which results in the separation of the negative charge center

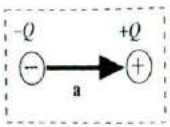


Figure 7.2 The definition of electric dipole moment.

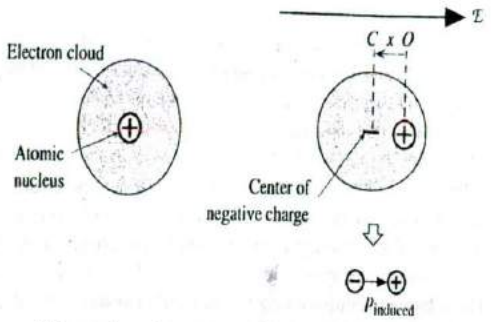


Figure 7.3 The origin of electronic polarization.

from the positive charge center, as shown in Figure 7.3b. This separation of negative and positive charges and the resulting induced dipole moment are termed **polarization.** An atom is said to be **polarized** if it possesses an effective dipole moment, that is, if there is a separation between the centers of negative and positive charge distributions.

The induced dipole moment depends on the electric field causing it. We define a quantity called the **polarizability** α to relate the induced dipole moment $p_{induced}$ to the field $\mathcal{E}$ causing it,

Definition of polarizability

$$p_{induced} = \alpha \mathcal{E} \qquad [7.4]$$

where α is a coefficient called the polarizability of the atom. It depends on the polarization mechanism. Since the polarization of a neutral atom involves the displacement of electrons, α is called **electronic polarization** and denoted as α_e. Inasmuch as the electrons in an atom are not rigidly fixed, all atoms possess a certain amount of electronic polarizability.

In the absence of an electric field, the center of mass of the orbital motions of the electrons coincides with the positively charged nucleus and the electronic dipole moment is zero. Suppose that the atom has Z number of electrons orbiting the nucleus and all the electrons are contained within a certain spherical region. When an electric field $\mathcal{E}$ is applied, the light electrons become displaced in the opposite direction to $\mathcal{E}$, so their center of mass C is shifted by some distance x with respect to the nucleus O, which we take to be the origin, as shown in Figure 7.3b. As the electrons are "pushed" away by the applied field, the Coulombic attraction between the electrons and nuclear charge "pulls in" the electrons. The force on the electrons, due to $\mathcal{E}$, trying to separate them away from the nuclear charge is $Ze\mathcal{E}$. The restoring force F_r, which is the Coulombic attractive force between the electrons and the nucleus, can be taken to be proportional to the displacement x, provided that the latter is small.[1] The restoring force F_r is obviously zero when C coincides with $O(x = 0)$. We can write

Restoring force

$$F_r = -\beta x$$

[1] It may be noticed that even if F_r is a complicated function of x, it can still be expanded in a series in terms of powers of x, that is, x, x^2, x^3, and so on, and for small x only the x term is significant, $F_r = -\beta x$.

where β is a constant and the negative sign indicates that F_r is always directed toward the nucleus O (Figure 7.3b). In equilibrium, the net force on the negative charge is zero or

$$Ze\mathcal{E} = \beta x$$

from which x is known. Therefore the **magnitude** of the induced electronic dipole moment p_e is given by

$$p_e = (Ze)x = \left(\frac{Z^2e^2}{\beta}\right)\mathcal{E} \qquad [7.5]$$

Electronic polarization

As expected, p_e is proportional to the applied field. The electronic dipole moment in Equation 7.5 is valid under static conditions, that is, when the electric field is a dc field. Suppose that we suddenly remove the applied electric field polarizing the atom. There is then only the restoring force $-\beta x$, which always acts to pull the electrons toward the nucleus O. The equation of motion of the negative charge center is then (from force = mass × acceleration)

Equation for simple harmonic motion

$$-\beta x = Zm_e\frac{d^2x}{dt^2}$$

Thus the displacement at any time is

$$x(t) = x_0\cos(\omega_o t)$$

where

$$\omega_o = \left(\frac{\beta}{Zm_e}\right)^{1/2} \qquad [7.6]$$

Electronic polarization resonance frequency

is the oscillation frequency of the center of mass of the electron cloud about the nucleus and x_0 is the displacement before the removal of the field. After the removal of the field, the electronic charge cloud executes simple harmonic motion about the nucleus with a natural frequency determined by Equation 7.6; ω_o is called the **electronic polarization resonance frequency.**[2] It is analogous to a mass on a spring being pulled and let go. The system then executes simple harmonic motion. The oscillations of course die out with time. In the atomic case, a sinusoidal displacement implies that the electronic charge cloud has an acceleration

$$\frac{d^2x}{dt^2} = -x_0\omega_o^2\cos(\omega_o t)$$

It is well known from classical electromagnetism that an accelerating charge radiates electromagnetic energy just like a radio antenna. Consequently the oscillating charge

[2] The term *natural frequency* refers to a system's characteristic frequency of oscillation when it is excited. A mass attached to a spring and then let go will execute simple harmonic motion with a certain natural frequency ω_o. If we then decide to oscillate this mass with an applied force, the maximum energy transfer will occur when the applied force has the same frequency as ω_o; the system will be put in resonance. ω_o is also a *resonant frequency*. Strictly, $\omega = 2\pi f$ is the angular frequency and f is the frequency. It is quite common to simply refer to ω as a frequency because the literature is dominated by ω; the meaning should be obvious within context.

cloud loses energy, and thus its amplitude of oscillation decreases. (Recall that the average energy is proportional to the square of the amplitude of the displacement.)

From the expression derived for p_e in Equation 7.5, we can find the electronic polarizability α_e from Equation 7.4,

Static electronic polarizability

$$\alpha_e = \frac{Ze^2}{m_e\omega_o^2} \qquad [7.7]$$

EXAMPLE 7.1 **ELECTRONIC POLARIZABILITY** Consider the electronic polarizability of inert gas atoms. These atoms have closed shells. Their electronic polarizabilities are listed in Table 7.1. For each type of atom calculate the electronic polarization resonant frequency $f_o = \omega_o/2\pi$, and plot α_e and f_o against the number of electrons Z in the atom. What is your conclusion?

SOLUTION

We can use Equation 7.7 to calculate the resonant frequency $f_o = \omega_o/2\pi$. Taking Ar,

$$\omega_o = \left(\frac{Ze^2}{\alpha_e m_e}\right)^{1/2} = \left[\frac{(18)(1.6 \times 10^{-19})^2}{(1.7 \times 10^{-40})(9.1 \times 10^{-31})}\right]^{1/2} = 5.46 \times 10^{16} \text{ rad s}^{-1}$$

Table 7.1 Electronic polarizability α_e dependence on Z for the inert element atoms

	Atom					
	He	Ne	Ar	Kr	Xe	Rn*
Z	2	10	18	36	56	
$\alpha_e \times 10^{-40}$ (F m^2)	0.18	0.45	1.7	2.7	4.4	5.9
$f_o \times 10^{15}$ (Hz)	8.90	12.6	8.69	9.76	9.36	10.2

*Rn (radon) gas is radioactive.

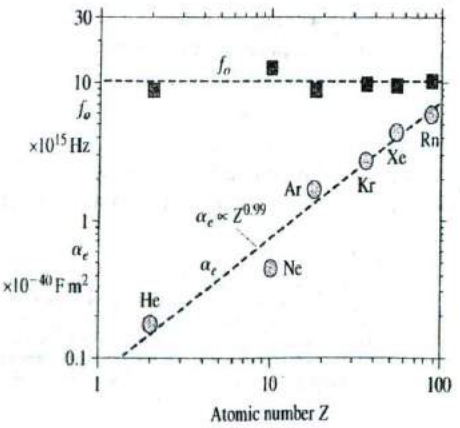


Figure 7.4 Electronic polarizability and its resonance frequency versus the number of electrons in the atom (Z).
The dashed line is the best-fit line.

so that

$$f_o = \frac{\omega_o}{2\pi} = 8.69 \times 10^{15} \text{ Hz}$$

which is listed in Table 7.1, among other f_o calculations for the other atoms. Such frequencies correspond to the field oscillations in UV light, that is, at optical frequencies. For all practical purposes, electronic polarization occurs very rapidly, that is, on a time scale $1/f_o$ or 10^{-15} s, and we can take the static polarizability α_e to remain the same up to optical frequencies.[3]

Figure 7.4 shows the dependence of α_e and f_o on the number of electrons Z. It is apparent that α_e is nearly linearly proportional to Z, whereas f_o is very roughly constant. It is left as an exercise to show that β increases with Z, which is reasonable since the restoring force was defined as the total force between *all* the electrons and the nucleus when the electrons are displaced.

7.1.3 POLARIZATION VECTOR P

When a material is placed in an electric field, the atoms and the molecules of the material become polarized, so we have a distribution of dipole moments in the material. We can visualize this effect with the insertion of the dielectric slab into the parallel plate capacitor, as depicted in Figure 7.5a. The placement of the dielectric slab into an electric field polarizes the molecules in the material. The induced dipole moments all point in the direction of the field. Consider the polarized medium alone, as shown in Figure 7.5b. In the bulk of the material, the dipoles are aligned head to tail. Every positive charge has a negative charge next to it and vice versa. There is therefore no net charge within the bulk. But the positive charges of the dipoles appearing at the right-hand face are not canceled by negative charges of any dipoles at this face. There is therefore a surface charge $+Q_P$ on the right-hand face that results from the polarization of the medium. Similarly, there is a negative charge $-Q_P$ with the same magnitude appearing on the left-hand face due to the negative charges of the dipoles at this face. We see that charges $+Q_P$ and $-Q_P$ appear on the opposite surfaces of a material when it becomes polarized in an electric field, as shown in Figure 7.5c. These charges are **bound** and are a direct result of the polarization of the molecules. They are termed **surface polarization charges.** Figure 7.5c emphasizes this aspect of dielectric behavior in an electric field by showing the dielectric and its polarization charges only.

We represent the polarization of a medium by a quantity called **polarization P,** which is defined as the total dipole moment per unit volume,

Definition of polarization vector

$$\mathbf{P} = \frac{1}{\text{Volume}}[\mathbf{p}_1 + \mathbf{p}_2 + \cdots + \mathbf{p}_N] \qquad [7.8a]$$

where $\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_N$ are the dipole moments induced at N molecules in the volume. If $\mathbf{p}_{av}$ is the average dipole moment per molecule, then an equivalent definition of $\mathbf{P}$ is

Definition of polarization vector

$$\mathbf{P} = N\mathbf{p}_{av} \qquad [7.8b]$$

[3] Electronic polarization at optical frequencies controls the optical properties such as the refractive index, as will be covered in Chapter 9.

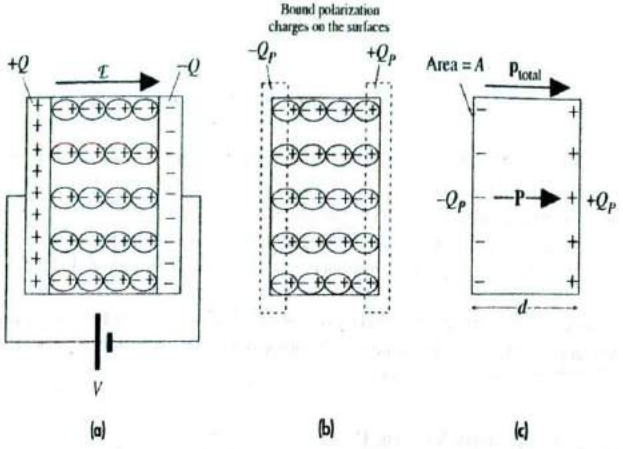


Figure 7.5

(a) When a dielectric is placed in an electric field, bound polarization charges appear on the opposite surfaces.

(b) The origin of these polarization charges is the polarization of the molecules of the medium.

(c) We can represent the whole dielectric in terms of its surface polarization charges $+Q_P$ and $-Q_P$.

where N is the number of molecules per unit volume. There is an important relationship, given below, between **P** and the polarization charges Q_P on the surfaces of the dielectric. It should be emphasized for future discussions that if polarization arises from the effect of the applied field, as shown in Figure 7.5a, which is usually the case, $\mathbf{p}_{av}$ must be the *average dipole moment per atom in the direction of the applied field.* In that case we often also denote $\mathbf{p}_{av}$ as the induced average dipole moment per molecule $\mathbf{p}_{induced}$.

To calculate the polarization **P** for the polarized dielectric in Figure 7.5b, we need to sum all the dipoles in the medium and divide by the volume Ad, as in Equation 7.8a. However, the polarized medium can be simply represented as in Figure 7.5c in terms of surface charge $+Q_P$ and $-Q_P$ which are separated by the thickness distance d. We can view this arrangement as one big dipole moment p_{total} from $-Q_P$ to $+Q_P$. Thus

$$p_{total} = Q_P d$$

Since the polarization is defined as the total dipole moment per unit volume, the magnitude of **P** is

Polarization and bound surface charge density

$$P = \frac{p_{total}}{\text{Volume}} = \frac{Q_P d}{Ad} = \frac{Q_P}{A}$$

But Q_P/A is the **surface polarization charge density** σ_P, so

$$P = \sigma_P \qquad [7.9a]$$

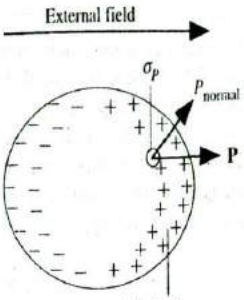


Figure 7.6 Polarization charge density on the surface of a polarized medium is related to the normal component of the polarization vector.

Polarization is a vector and Equation 7.9a only gives its magnitude. For the rectangular slab in Figure 7.5c, the direction of P is normal to the surface. For $+\sigma_p$ (right face), it comes out from the surface and for $-\sigma_p$ (left face), it is directed into the surface. Although Equation 7.9a is derived for one specific geometry, the rectangular slab, it can be generalized as follows. *The charge per unit area appearing on the surface of a polarized medium is equal to the component of the polarization vector normal to this surface.* If P_{normal} is the component of **P** normal to the surface where the polarization charge density is σ_p, as shown in Figure 7.6, then,

Polarization and bound surface charge density

$$P_{\text{normal}} = \sigma_p \qquad [7.9b]$$

The polarization P induced in a dielectric medium when it is placed in an electric field depends on the field itself. The induced dipole moment per molecule within the medium depends on the electric field by virtue of Equation 7.4. To express the dependence of P on the field $\mathcal{E}$, we define a quantity called the **electric susceptibility** χ_e by

Definition of electric susceptibility

$$P = \chi_e \varepsilon_o \mathcal{E} \qquad [7.10]$$

Equation 7.10 shows an *effect* P due to a *cause* $\mathcal{E}$ and the quantity χ_e relates the effect to its cause. Put differently, χ_e acts as a proportionality constant. It may depend on the field itself, in which case the effect is nonlinearly related to the cause. Further, electronic polarizability is defined by

$$p_{\text{induced}} = \alpha_e \mathcal{E}$$

so

$$P = N p_{\text{induced}} = N \alpha_e \mathcal{E}$$

where N is the number of molecules per unit volume. Then from Equation 7.10, χ_e and α_e are related by

Electric susceptibility and polarization

$$\chi_e = \frac{1}{\varepsilon_o} N \alpha_e \qquad [7.11]$$

It is important to recognize the difference between *free* and *polarization* (or *bound*) charges. The charges stored on the metal plates in Figure 7.5a are free because they result from the motion of free electrons in the metal. For example both Q_o and Q, before and after the dielectric insertion in Figure 7.1, are free charges that arrive on the plates from the battery. The polarization charges $+Q_P$ and $-Q_P$, on the other hand, are bound to the molecules. They cannot move within the dielectric or on its surface.

The field $\mathcal{E}$ *before* the dielectric was inserted (Figure 7.1a) is given by

$$\mathcal{E} = \frac{V}{d} = \frac{Q_o}{C_o d} = \frac{Q_o}{\varepsilon_o A} = \frac{\sigma_o}{\varepsilon_o} \qquad [7.12]$$

where $\sigma_o = Q_o/A$ is the **free surface charge density** without any dielectric medium between the plates, as in Figure 7.1a.

After the insertion of the dielectric, this field remains the same V/d, but the free charges on the plates are different. The free surface charge on the plates is now Q. In addition there are bound polarization charges on the dielectric surfaces next to the plates, as shown in Figure 7.5a. It is apparent that the flow of current during the insertion of the dielectric, Figure 7.1b, is due to the additional free charges $Q - Q_o$ needed on the capacitor plates to neutralize the opposite polarity polarization charges Q_P appearing on the dielectric surfaces. The total charge (see Figure 7.5a) due to that on the plate plus that appearing on the dielectric surface, $Q - Q_P$, must be the same as before, Q_o, so that the field, as given by Equation 7.12, does not change inside the dielectric, that is,

$$Q - Q_P = Q_o$$

or

$$Q = Q_o + Q_P$$

Dividing by A, defining $\sigma = Q/A$ as the free surface charge density on the plates with the dielectric inserted, and using Equation 7.12, we obtain

$$\sigma = \varepsilon_o \mathcal{E} + \sigma_P$$

Since $\sigma_P = P$ and $P = \chi_e \varepsilon_o \mathcal{E}$, Equations 7.9 and 7.10, we can eliminate σ_P to obtain

$$\sigma = \varepsilon_o(1 + \chi_e)\mathcal{E}$$

From the definition of the relative permittivity in Equation 7.2 we have

$$\varepsilon_r = \frac{Q}{Q_o} = \frac{\sigma}{\sigma_o}$$

Relative permittivity and electric susceptibility

so substituting for σ and using Equation 7.12 we obtain

$$\varepsilon_r = 1 + \chi_e \qquad [7.13]$$

In terms of electronic polarization, from Equation 7.11, this is

Relative permittivity and polarizability

$$\varepsilon_r = 1 + \frac{N\alpha_e}{\varepsilon_o} \qquad [7.14]$$

The significance of Equation 7.14 is that it relates the microscopic polarization mechanism that determines α_e to the macroscopic property ε_r.

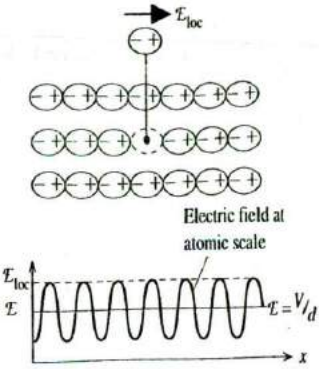


Figure 7.7 The electric field inside a polarized dielectric at the atomic scale is not uniform. The local field is the actual field that acts on a molecule. It can be calculated by removing that molecule and evaluating the field at that point from the charges on the plates and the dipoles surrounding the point.

7.1.4 LOCAL FIELD $\mathcal{E}_{loc}$ AND CLAUSIUS–MOSSOTTI EQUATION

Equation 7.14, which relates ε_r to electronic polarizability α_e is only approximate because it assumes that the field acting on an individual atom or molecule is the field $\mathcal{E}$, which is assumed to be uniform within the dielectric. In other words, the induced polarization, $p_{induced} \propto \mathcal{E}$. However, the induced polarization depends on the actual field experienced by the molecule. It is apparent from Figure 7.5a that there are polarized molecules within the dielectric with their negative and positive charges separated so that the field is not constant *on the atomic scale* as we move through the dielectric. This is depicted in Figure 7.7. The field experienced by an individual molecule is actually different than $\mathcal{E}$, which represents the average field in the dielectric. As soon as the dielectric becomes polarized, the field at some arbitrary point depends not only on the charges on the plates (Q) but also on the orientations of all the other dipoles around this point in the dielectric. When averaged over some distance, say a few thousand molecules, this field becomes $\mathcal{E}$, as shown in Figure 7.7.

The actual field experienced by a molecule in a dielectric is defined as the **local field** and denoted by $\mathcal{E}_{loc}$. It depends not only on the free charges on the plates but also on the arrangement of all the polarized molecules around this point. In evaluating $\mathcal{E}_{loc}$ we simply remove the molecule from this point and calculate the field at this point coming from all sources, including neighboring polarized molecules, as visualized in Figure 7.7. $\mathcal{E}_{loc}$ will depend on the amount of polarization the material has experienced. The greater the polarization, the greater is the local field because there are bigger dipoles around this point. $\mathcal{E}_{loc}$ depends on the arrangement of polarized molecules around the point of interest and hence depends on the crystal structure. In the simplest case of a material with a cubic crystal structure, or a liquid (no crystal structure), the local field $\mathcal{E}_{loc}$ acting on a molecule increases with polarization as[4]

Lorentz local field in dielectrics

$$\mathcal{E}_{loc} = \mathcal{E} + \frac{1}{3\varepsilon_o} P \qquad [7.15]$$

[4] This field is called the **Lorentz field** and the proof, though not difficult, is not necessary for the present introductory treatment of dielectrics. This local field expression does not apply to dipolar dielectrics discussed in Section 7.3.2.

Equation 7.15 is called the **Lorentz field.** The induced polarization in the molecule now depends on this local field $\mathcal{E}_{loc}$ rather than the average field $\mathcal{E}$. Thus

$$p_{induced} = \alpha_e \mathcal{E}_{loc}$$

The fundamental definition of electric susceptibility by the equation

$$P = \chi_e \varepsilon_o \mathcal{E}$$

is unchanged, which means that $\varepsilon_r = 1 + \chi_e$, Equation 7.13, remains intact. The polarization is defined by $P = Np_{induced}$, and $p_{induced}$ can be related to $\mathcal{E}_{loc}$ and hence to $\mathcal{E}$ and P. Then

$$P = (\varepsilon_r - 1)\varepsilon_o \mathcal{E}$$

can be used to eliminate $\mathcal{E}$ and P and obtain a relationship between ε_r and α_e. This is the **Clausius–Mossotti equation,**

Clausius–Mossotti equation

$$\frac{\varepsilon_r - 1}{\varepsilon_r + 2} = \frac{N\alpha_e}{3\varepsilon_o} \qquad [7.16]$$

This equation allows the calculation of the macroscopic property ε_r from microscopic polarization phenomena, namely, α_e.

EXAMPLE 7.2 **ELECTRONIC POLARIZABILITY OF A VAN DER WAALS SOLID** The electronic polarizability of the Ar atom is 1.7×10^{-40} F m^2. What is the static dielectric constant of solid Ar (below 84 K) if its density is 1.8 g cm^{-3}?

SOLUTION

To calculate ε_r, we need the number of Ar atoms per unit volume N from the density d. If $M_{at} = 39.95$ is the relative atomic mass of Ar and N_A is Avogadro's number, then

$$N = \frac{N_A d}{M_{at}} = \frac{(6.02 \times 10^{23}\ \text{mol}^{-1})(1.8\ \text{g cm}^{-3})}{(39.95\ \text{g mol}^{-1})} = 2.71 \times 10^{22}\ \text{cm}^{-3}$$

with $N = 2.71 \times 10^{28}$ m^{-3} and $\alpha_e = 1.7 \times 10^{-40}$ F m^2, we have

$$\varepsilon_r = 1 + \frac{N\alpha_e}{\varepsilon_o} = 1 + \frac{(2.71 \times 10^{28})(1.7 \times 10^{-40})}{(8.85 \times 10^{-12})} = 1.52$$

If we use the Clausius–Mossotti equation, we get

$$\varepsilon_r = \frac{1 + \dfrac{2N\alpha_e}{3\varepsilon_o}}{1 - \dfrac{N\alpha_e}{3\varepsilon_o}} = 1.63$$

The two values are different by about 7 percent. The simple relationship in Equation 7.14 underestimates the relative permittivity.

7.2 ELECTRONIC POLARIZATION: COVALENT SOLIDS

When a field is applied to a solid substance, the constituent atoms or molecules become polarized, as we visualized in Figure 7.5a. The electron clouds within each atom become shifted by the field, and this gives rise to **electronic polarization.** This type of electronic polarization within an atom, however, is quite small compared with the polarization due to the valence electrons in the covalent bonds within the solid. For example, in crystalline silicon, there are electrons shared with neighboring Si atoms in covalent bonds, as shown in Figure 7.8a. These valence electrons form bonds (*i.e.*, become shared) between the Si atoms because they are already loosely bound to their parent atoms. If this were not the case, the solid would be a van der Waals solid with atoms held together by secondary bonds (*e.g.*, solid Ar below 83.8 K). In the covalent solid, the valence electrons therefore are not rigidly tied to the ionic cores left in the Si atoms. Although intuitively we often view these valence electrons as living in covalent bonds between the ionic Si cores, they nonetheless belong to the whole crystal because they can tunnel from bond to bond and exchange places with each other. We refer to their wavefunctions as delocalized, that is, not localized to any particular Si atom. When an electric field is applied, the negative charge distribution associated with these valence electrons becomes readily shifted with respect to the positive charges of the ionic Si cores, as depicted in Figure 7.8b and the crystal exhibits polarization, or develops a polarization vector. One can appreciate the greater flexibility of electrons in covalent bonds compared with those in individual ionic cores by comparing the energy involved in freeing each. It takes perhaps 1–2 eV to break a covalent bond to free the valence electron, but it takes more than 10 eV to free an electron from an individual ionic Si core. Thus, the valence electrons in the bonds readily respond to an applied field and become displaced. This type of electronic polarization, due to the displacement of electrons in covalent bonds, is responsible for the large dielectric constants of covalent crystals. For example $\varepsilon_r = 11.9$ for the Si crystal and $\varepsilon_r = 16$ for the Ge crystal.

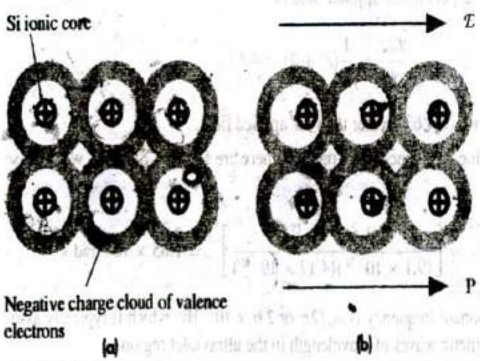


Figure 7.8
(a) Valence electrons in covalent bonds in the absence of an applied field.
(b) When an electric field is applied to a covalent solid, the valence electrons in the covalent bonds are shifted very easily with respect to the positive ionic cores. The whole solid becomes polarized due to the collective shift in the negative charge distribution of the valence electrons.

EXAMPLE 7.3 **ELECTRONIC POLARIZABILITY OF COVALENT SOLIDS** Consider a pure Si crystal that has $\varepsilon_r = 11.9$.

a. What is the electronic polarizability due to valence electrons per Si atom (if one could portion the observed crystal polarization to individual atoms)?

b. Suppose that a Si crystal sample is electroded on opposite faces and has a voltage applied across it. By how much is the local field greater than the applied field?

c. What is the resonant frequency f_o corresponding to ω_o?

From the density of the Si crystal, the number of Si atoms per unit volume, N, is given as $5 \times 10^{28}\ \text{m}^{-3}$.

SOLUTION

a. Given the number of Si atoms, we can apply the Clausius–Mossotti equation to find α_e

$$\alpha_e = \frac{3\varepsilon_o}{N}\frac{\varepsilon_r - 1}{\varepsilon_r + 2} = \frac{3(8.85 \times 10^{-12})}{(5 \times 10^{28})}\frac{11.9 - 1}{11.9 + 2} = 4.17 \times 10^{-40}\ \text{F m}^2$$

This is larger, for example, than the electronic polarizability of an isolated Ar atom, which has more electrons. If we were to take the inner electrons in each Si atom as very roughly representing Ne, we would expect their contribution to the overall electronic polarizability to be roughly the same as the Ne atom, which is $0.45 \times 10^{-40}\ \text{F m}^2$.

b. The local field is

$$\mathcal{E}_{\text{loc}} = \mathcal{E} + \frac{1}{3\varepsilon_o}P$$

But, by definition,

$$P = \chi_e \varepsilon_o \mathcal{E} = (\varepsilon_r - 1)\varepsilon_o \mathcal{E}$$

Substituting for P,

$$\mathcal{E}_{\text{loc}} = \mathcal{E} + \frac{1}{3}(\varepsilon_r - 1)\mathcal{E}$$

so the local field with respect to the applied field is

$$\frac{\mathcal{E}_{\text{loc}}}{\mathcal{E}} = \frac{1}{3}(\varepsilon_r + 2) = 4.63$$

The local field is a factor of 4.63 greater than the applied field.

c. Since polarization is due to valence electrons and there are four per Si atom, we can use Equation 7.7,

$$\omega_o = \left(\frac{Ze^2}{m_e \alpha_e}\right)^{1/2} = \left[\frac{4(1.6 \times 10^{-19})^2}{(9.1 \times 10^{-31})(4.17 \times 10^{-40})}\right]^{1/2} = 1.65 \times 10^{16}\ \text{rad s}^{-1}$$

The corresponding resonant frequency is $\omega_o/2\pi$ or 2.6×10^{15} Hz, which is typically associated with electromagnetic waves of wavelength in the ultraviolet region.

7.3 POLARIZATION MECHANISMS

In addition to electronic polarization, we can identify a number of other polarization mechanisms that may also contribute to the relative permittivity.

7.3.1 IONIC POLARIZATION

This type of polarization occurs in ionic crystals such as NaCl, KCl, and LiBr. The ionic crystal has distinctly identifiable ions, for example, Na^+ and Cl^-, located at well-defined lattice sites, so each pair of oppositely charged neighboring ions has a dipole moment. As an example, we consider the one-dimensional NaCl crystal depicted as a chain of alternating Na^+ and Cl^- ions in Figure 7.9a. In the absence of an applied field, the solid has no net polarization because the dipole moments of equal magnitude are lined up head to head and tail to tail so that the net dipole moment is zero. The dipole moment p_+ in the positive x direction has the same magnitude as p_- in the negative x direction, so the net dipole moment

$$p_{\text{net}} = p_+ - p_- = 0$$

In the presence of a field $\mathcal{E}$ along the x direction, however, the Cl^- ions are pushed in the $-x$ direction and the Na^+ ions in the $+x$ direction about their equilibrium positions. Consequently, the dipole moment p_+ in the $+x$ direction *increases* to p'_+ and the dipole moment p_- *decreases* to p'_-, as shown in Figure 7.9b. The net dipole moment is now no longer zero. The net dipole moment, or the average dipole moment, per ion pair is now $(p'_+ - p'_-)$, which depends on the electric field $\mathcal{E}$. Thus the induced average dipole moment per ion pair p_{av} depends on the field $\mathcal{E}$. The ionic polarizability α_i is defined in terms of the local field experienced by the ions,

$$p_{\text{av}} = \alpha_i \mathcal{E}_{\text{loc}} \qquad [7.17]$$

Ionic polarizability

The larger the α_i, the greater the induced dipole moment. Generally, α_i is larger than the electronic polarizability α_e by a factor of 10 or more, which leads to ionic solids having large dielectric constants. The polarization P exhibited by the ionic solid

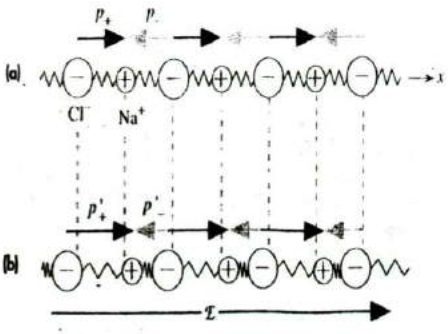


Figure 7.9
(a) A NaCl chain in the NaCl crystal without an applied field. Average or net dipole moment per ion is zero.
(b) In the presence of an applied field, the ions become slightly displaced, which leads to a net average dipole moment per ion.

is therefore given by

$$P = N_i p_{av} = N_i \alpha_i \mathcal{E}_{loc}$$

where N_i is the number of ion pairs per unit volume. By relating the local field to $\mathcal{E}$ and using

$$P = (\varepsilon_r - 1)\varepsilon_o \mathcal{E}$$

Clausius–Mossotti equation for ionic polarization

we can again obtain the Clausius–Mossotti equation, but now due to ionic polarization,

$$\frac{\varepsilon_r - 1}{\varepsilon_r + 2} = \frac{1}{3\varepsilon_o} N_i \alpha_i \qquad [7.18]$$

Each ion also has a core of electrons that become displaced in the presence of an applied field with respect to their positive nuclei and therefore also contribute to the polarization of the solid. This electronic polarization simply adds to the ionic polarization. Its magnitude is invariably much smaller than the ionic contribution in these solids.

7.3.2 ORIENTATIONAL (DIPOLAR) POLARIZATION

Certain molecules possess permanent dipole moments. For example, the HCl molecule shown in Figure 7.10a has a permanent dipole moment p_o from the Cl^- ion to the H^+ ion. In the liquid or gas phases, these molecules, in the absence of an electric field, are randomly oriented as a result of thermal agitation, as shown in Figure 7.10b. When an electric field $\mathcal{E}$ is applied, $\mathcal{E}$ tries to align the dipoles parallel to itself, as depicted in Figure 7.10c. The Cl^- and H^+ charges experience forces in opposite directions. But the nearly rigid bond between Cl^- and H^+ holds them together, which means that the

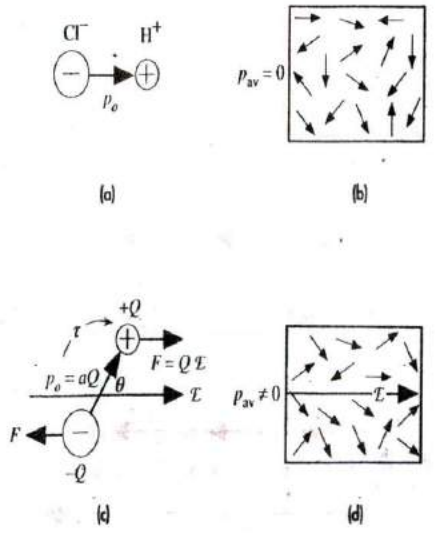


Figure 7.10

(a) A HCl molecule possesses a permanent dipole moment p_o.

(b) In the absence of a field, thermal agitation of the molecules results in zero net average dipole moment per molecule.

(c) A dipole such as HCl placed in a field experiences a torque that tries to rotate it to align p_o with the field $\mathcal{E}$.

(d) In the presence of an applied field, the dipoles try to rotate to align with the field against thermal agitation. There is now a net average dipole moment per molecule along the field.

molecule experiences a torque τ about its center of mass.[5] This torque acts to rotate the molecule to align p_o with $\mathcal{E}$. If all the molecules were to simply rotate and align with the field, the polarization of the solid would be

$$P = Np_o$$

where N is the number of molecules per unit volume. However, due to their thermal energy, the molecules move around randomly and collide with each other and with the walls of the container. These collisions destroy the dipole alignments. Thus the thermal energy tries to randomize the orientations of the dipole moments. A snapshot of the dipoles in the material in the presence of a field can be pictured as in Figure 7.10d in which the dipoles have different orientations. There is, nonetheless, a net average dipole moment per molecule p_{av} that is finite and directed along the field. Thus the material exhibits net polarization, which leads to a dielectric constant that is determined by this **orientational polarization.**

To find the induced average dipole moment p_{av} along $\mathcal{E}$, we need to know the average potential energy E_{dip} of a dipole placed in a field $\mathcal{E}$ and how this compares with the average thermal energy $\frac{5}{2}kT$ per molecule as in the present case of five degrees of freedom. E_{dip} represents the average external work done by the field in aligning the dipoles with the field. If $\frac{5}{2}kT$ is much greater than E_{dip}, then the average thermal energy of collisions will prevent any dipole alignment with the field. If, however, E_{dip} is much greater than $\frac{5}{2}kT$, then the thermal energy is insufficient to destroy the dipole alignments.

A dipole at an angle θ to the field experiences a torque τ that tries to rotate it, as shown in Figure 7.10c. Work done dW by the field in rotating the dipole by $d\theta$ is $\tau\, d\theta$ (as in $F\, dx$). This work dW represents a small change dE in the potential energy of the dipole. No work is done if the dipole is already aligned with $\mathcal{E}$, when $\theta = 0$, which corresponds to the minimum in PE. On the other hand, maximum work is done when the torque has to rotate the dipole from $\theta = 180°$ to $\theta = 0°$ (either clockwise or counterclockwise, it doesn't matter). The torque experienced by the dipole, according to Figure 7.10c, is given by

Torque on a dipole

$$\tau = (F \sin\theta)a \qquad \text{or} \qquad \mathcal{E}p_o \sin\theta$$

where

$$p_o = aQ$$

If we take $PE = 0$ when $\theta = 0$, then the maximum PE is when $\theta = 180°$, or

$$E_{max} = \int_0^{\pi} p_o \mathcal{E} \sin\theta\, d\theta = 2p_o\mathcal{E}$$

The average dipole potential energy is then $\frac{1}{2}E_{max}$ or $p_o\mathcal{E}$. For orientational polarization to be effective, this energy must be greater than the average thermal energy. The average dipole moment p_{av} along $\mathcal{E}$ is directly proportional to the magnitude of p_o itself and also proportional to the average dipole energy to average thermal energy

[5] The oppositely directed forces also slightly stretch the Cl –H⁺ bond, but we neglect this effect.

ratio, that is,

$$p_{av} \propto p_o \frac{p_o \mathcal{E}}{\frac{5}{2} kT}$$

If we were to do the calculation properly using Boltzmann statistics for the distribution of dipole energies among the molecules, that is, the probability that the dipole has an energy E is proportional to $\exp(-E/kT)$, then we would find that when $p_o E < kT$ (generally the case),

Average dipole moment in orientational polarization

$$p_{av} = \frac{1}{3} \frac{p_o^2 \mathcal{E}}{kT} \qquad [7.19]$$

It turns out that the intuitively derived expression for p_{av} is roughly the same as Equation 7.19. Strictly, of course, we should use the local field acting on each molecule, in which case $\mathcal{E}$ is simply replaced by $\mathcal{E}_{loc}$. From Equation 7.19 we can define a **dipolar orientational polarizability** α_d per molecule by

Dipolar orientational polarizability

$$\alpha_d = \frac{1}{3} \frac{p_o^2}{kT} \qquad [7.20]$$

It is apparent that, in contrast to the electronic and ionic polarization, dipolar orientational polarization is strongly temperature dependent. α_d decreases with temperature, which means that the relative permittivity ε_r also decreases with temperature. Dipolar orientational polarization is normally exhibited by polar liquids (*e.g.*, water, alcohol, acetone, and various electrolytes) and polar gases (*e.g.*, gaseous HCl and steam). It can also occur in solids if there are permanent dipoles within the solid structure, even if dipolar rotation involves a discrete jump of an ion from one site to another, such as in various glasses.

7.3.3 INTERFACIAL POLARIZATION

Interfacial polarization occurs whenever there is an accumulation of charge at an interface between two materials or between two regions within a material. The simplest example is interfacial polarization due to the accumulation of charges in the dielectric near one of the electrodes, as depicted in Figure 7.11a and b. Invariably materials, however perfect, contain crystal defects, impurities, and various mobile charge carriers such as electrons (*e.g.*, from donor-type impurities), holes, or ionized host or impurity ions. In the particular example in Figure 7.11a, the material has an equal number of positive ions and negative ions, but the positive ions are assumed to be far more mobile. For example, if present, the H^+ ion (which is a proton) and the Li^+ ion in ceramics and glasses are more mobile than negative ions in the structure because they are relatively small. Under the presence of an applied field, these positive ions migrate to the negative electrode. The positive ions, however, cannot leave the dielectric and enter the crystal structure of the metal electrode. They therefore simply pile up at the interface and give rise to a positive space charge near the electrode. These positive charges at the interface attract more electrons to the negative electrode. This additional charge on the electrode,

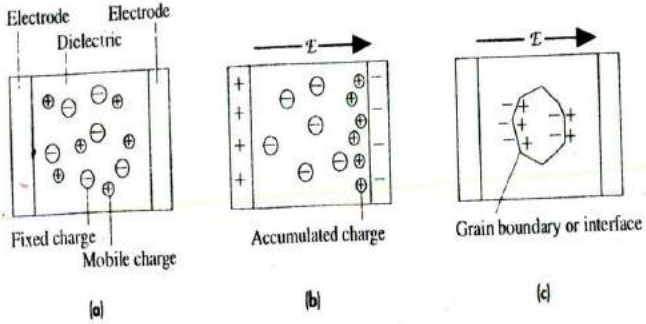


Figure 7.11

(a) A crystal with equal number of mobile positive ions and fixed negative ions. In the absence of a field, there is no net separation between all the positive charges and all the negative charges.

(b) In the presence of an applied field, the mobile positive ions migrate toward the negative electrode and accumulate there. There is now an overall separation between the negative charges and positive charges in the dielectric. The dielectric therefore exhibits interfacial polarization.

(c) Grain boundaries and interfaces between different materials frequently give rise to interfacial polarization.

of course, appears as an increase in the dielectric constant. The term **interfacial polarization** arises because the positive charges accumulating at the interface and the remainder of negative charges in the bulk together constitute dipole moments that appear in the polarization vector **P** (**P** sums all the dipoles within the material per unit volume).

Another typical interfacial polarization mechanism is the trapping of electrons or holes at defects at the crystal surface, at the interface between the crystal and the electrode. In this case we can view the positive charges in Figure 7.11a as holes and negative charges as immobile ionized acceptors. We assume that the contacts are blocking and do not allow electrons or holes to be injected, that is, exchanged between the electrodes and the dielectric. In the presence of a field, the holes drift to the negative electrode and become trapped in defects at the interface, as in Figure 7.11b.

Grain boundaries frequently lead to interfacial polarization as they can trap charges migrating under the influence of an applied field, as indicated in Figure 7.11c. Dipoles between the trapped charges increase the polarization vector. Interfaces also arise in heterogeneous dielectric materials, for example, when there is a dispersed phase within a continuous phase. The principle is then the same as schematically illustrated in Figure 7.11c.

7.3.4 TOTAL POLARIZATION

In the presence of electronic, ionic, and dipolar polarization mechanisms, the average induced dipole moment per molecule will be the sum of all the contributions in terms of the local field,

Total induced dipole moment

$$p_{av} = \alpha_e \mathcal{L}_{loc} + \alpha_i \mathcal{L}_{loc} + \alpha_d \mathcal{L}_{loc}$$

Table 7.2 Typical examples of polarization mechanisms

Example	Polarization	Static ε_r	Comment
Ar gas	Electronic	1.0005	Small N in gases: $\varepsilon_r \approx 1$
Ar liquid ($T < 87.3$ K)	Electronic	1.53	van der Waals bonding
Si crystal	Electronic polarization due to valence electrons	11.9	Covalent solid; bond polarization
NaCl crystal	Ionic	5.90	Ionic crystalline solid
CsCl crystal	Ionic	7.20	Ionic crystalline solid
Water	Orientational	80	Dipolar liquid
Nitromethane (27 °C)	Orientational	34	Dipolar liquid
PVC (polyvinyl chloride)	Orientational	7	Dipole orientations partly hindered in the solid

Each effect adds linearly to the net dipole moment per molecule, a fact verified by experiments. Interfacial polarization cannot be simply added to the above equation as $\alpha_{ij}\mathcal{E}_{loc}$ because it occurs at interfaces and cannot be put into an average polarization per molecule in the bulk. Further, the fields are not well defined at the interfaces. In addition, we *cannot* use the simple Lorentz local field approximation for dipolar materials. That is, the Clausius–Mossotti equation does not work with dipolar dielectrics and the calculation of the local field is quite complicated. The dielectric constant ε_r under **electronic** and **ionic polarizations,** however, can be obtained from

Clausius–Mossotti equation

$$\frac{\varepsilon_r - 1}{\varepsilon_r + 2} = \frac{1}{3\varepsilon_o}(N_e\alpha_e + N_i\alpha_i) \qquad [7.21]$$

Table 7.2 summarizes the various polarization mechanisms and the corresponding static (or very low frequency) dielectric constant. Typical examples where one mechanism dominates over others are also listed.

EXAMPLE 7.4 **IONIC AND ELECTRONIC POLARIZABILITY** Consider the CsCl crystal which has one Cs^+–Cl^- pair per unit cell and a lattice parameter a of 0.412 nm. The electronic polarizability of Cs^+ and Cl^- ions is 3.35×10^{-40} F m^2 and 3.40×10^{-40} F m^2, respectively, and the mean ionic polarizability per ion pair is 6×10^{-40} F m^2. What is the dielectric constant at low frequencies and that at optical frequencies?

SOLUTION

The CsCl structure has one cation (Cs^+) and one anion (Cl^-) in the unit cell. Given the lattice parameter $a = 0.412 \times 10^{-9}$ m, the number of ion pairs N_i per unit volume is $1/a^3 = 1/(0.412 \times 10^{-9}\text{ m})^3 = 1.43 \times 10^{28}$ m^{-3}. N_i is also the concentration of cations and anions individually. From the Clausius–Mossotti equation,

$$\frac{\varepsilon_r - 1}{\varepsilon_r + 2} = \frac{1}{3\varepsilon_o}[N_i\alpha_e(Cs^+) + N_i\alpha_e(Cl^-) + N_i\alpha_i]$$

That is,

$$\frac{\varepsilon_r - 1}{\varepsilon_r + 2} = \frac{(1.43 \times 10^{28}\ \text{m}^{-3})(3.35 \times 10^{-40} + 3.40 \times 10^{-40} + 6 \times 10^{-40}\ \text{F m}^2)}{3(8.85 \times 10^{-12}\ \text{F m}^{-1})}$$

Solving for ε_r, we find $\varepsilon_r = 7.56$.

At high frequencies—that is, near-optical frequencies—the ionic polarization is too sluggish to allow ionic polarization to contribute to ε_r. Thus, ε_{rop}, relative permittivity at optical frequencies, is given by

$$\frac{\varepsilon_{rop} - 1}{\varepsilon_{rop} + 2} = \frac{1}{3\varepsilon_o}[N_i\alpha_e(\text{Cs}^+) + N_i\alpha_e(\text{Cl}^-)]$$

That is,

$$\frac{\varepsilon_{rop} - 1}{\varepsilon_{rop} + 2} = \frac{(1.43 \times 10^{28}\ \text{m}^{-3})(3.35 \times 10^{-40} + 3.40 \times 10^{-40}\ \text{F m}^2)}{3(8.85 \times 10^{-12}\ \text{F m}^{-1})}$$

Solving for ε_{rop}, we find $\varepsilon_{rop} = 2.71$. Note that experimental values are $\varepsilon_r = 7.20$ at low frequencies and $\varepsilon_{rop} = 2.62$ at high frequencies, very close to calculated values.

7.4 FREQUENCY DEPENDENCE: DIELECTRIC CONSTANT AND DIELECTRIC LOSS

7.4.1 Dielectric Loss

The static dielectric constant is an effect of polarization under dc conditions. When the applied field, or the voltage across a parallel plate capacitor, is a sinusoidal signal, then the polarization of the medium under these ac conditions leads to an ac dielectric constant that is generally different than the static case. As an example we will consider orientational polarization involving dipolar molecules. The sinusoidally varying field changes magnitude and direction continuously, and it tries to line up the dipoles one way and then the other way and so on. If the instantaneous induced dipole moment p per molecule can instantaneously follow the field variations, then at any instant

$$p = \alpha_d \mathcal{E} \qquad [7.22]$$

and the polarizability α_d has its expected maximum value from dc conditions, that is,

$$\alpha_d = \frac{p_o^2}{3kT} \qquad [7.23]$$

There are two factors opposing the immediate alignment of the dipoles with the field. First is that thermal agitation tries to randomize the dipole orientations. Collisions in the gas phase, random jolting from lattice vibrations in the liquid and solid phases, for example, aid the randomization of the dipole orientations. Second, the molecules rotate in a viscous medium by virtue of their interactions with neighbors, which is particularly strong in the liquid and solid states and means that the dipoles cannot respond instantaneously to the changes in the applied field. If the field changes too

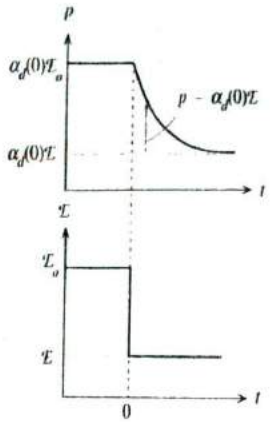


Figure 7.12 The applied dc field is suddenly changed from $\mathcal{E}_o$ to $\mathcal{E}$ at time $t = 0$. The induced dipole moment p has to decrease from $\alpha_d(0)\mathcal{E}_o$ to a final value of $\alpha_d(0)\mathcal{E}$. The decrease is achieved by random collisions of molecules in the gas.

rapidly, then the dipoles cannot follow the field and, as a consequence, remain randomly oriented. At high frequencies, therefore, α_d will be zero as the field cannot induce a dipole moment. At low frequencies, of course, the dipoles can respond rapidly to follow the field and α_d has its maximum value. It is clear that α_d changes from its maximum value in Equation 7.23 to zero as the frequency of the field is increased. We need to find the behavior of α_d as a function of frequency ω so that we can determine the dielectric constant ε_r by the Clausius–Mossotti equation.

Suppose that after a prolonged application, corresponding to dc conditions, the applied field across the dipolar gaseous medium is suddenly decreased from $\mathcal{E}_o$ to $\mathcal{E}$ at a time we define as zero, as shown in Figure 7.12. The field $\mathcal{E}$ is smaller than $\mathcal{E}_o$, so the induced dc dipole moment per molecule should be smaller and given by $\alpha_d(0)\mathcal{E}$ where $\alpha_d(0)$ is α_d at $\omega = 0$, dc conditions. Therefore, the induced dipole moment per molecule has to decrease, or *relax*, from $\alpha_d(0)\mathcal{E}_o$ to $\alpha_d(0)\mathcal{E}$. In a gas medium the molecules would be moving around randomly and their collisions with each other and the walls of the container randomize the induced dipole per molecule. Thus the decrease, or the **relaxation process,** in the induced dipole moment is achieved by random collisions. Assuming that τ is the average time, called the **relaxation time,** between molecular collisions, then this is the mean time it takes per molecule to randomize the induced dipole moment. If p is the instantaneous induced dipole moment, then $p - \alpha_d(0)\mathcal{E}$ is the *excess* dipole moment, which must eventually disappear to zero through random collisions as $t \to \infty$. It would take an average τ seconds to eliminate the excess dipole moment $p - \alpha_d(0)\mathcal{E}$. The rate at which the induced dipole moment is changing is then $-|p - \alpha_d(0)\mathcal{E}|/\tau$, where the negative sign represents a decrease. Thus,

Dipolar relaxation equation

$$\frac{dp}{dt} = -\frac{p - \alpha_d(0)\mathcal{E}}{\tau} \qquad [7.24]$$

Although we did not derive Equation 7.24 rigorously, it is nonetheless a good first-order description of the behavior of the induced dipole moment per molecule in

a dipolar medium. Equation 7.24 can be used to obtain the dipolar polarizability under ac conditions. For an ac field, we would write

$$\mathcal{E} = \mathcal{E}_o \sin(\omega t)$$

and solve Equation 7.24, but in engineering we prefer to use an exponential representation for the field

$$\mathcal{E} = \mathcal{E}_o \exp(j\omega t) \qquad \textit{Applied field}$$

as in ac voltages. In this case the impedance of a capacitor C and an inductor L become $1/j\omega C$ and $j\omega L$, where j represents a phase shift of 90°. With $\mathcal{E} = \mathcal{E}_o \exp(j\omega t)$ in Equation 7.24, we have

$$\frac{dp}{dt} = -\frac{p}{\tau} + \frac{\alpha_d(0)}{\tau}\mathcal{E}_o \exp(j\omega t) \qquad [7.25]$$

Dipole relaxation equation

Solving this we find the induced dipole moment as

$$p = \alpha_d(\omega)\mathcal{E}_o \exp(j\omega t)$$

where $\alpha_d(\omega)$ is given by

$$\alpha_d(\omega) = \frac{\alpha_d(0)}{1 + j\omega\tau} \qquad [7.26]$$

Orientational polarizability and frequency

and represents the orientational polarizability under ac field conditions. Polarizability $\alpha_d(\omega)$ is a complex number that indicates that p and $\mathcal{E}$ are out of phase.[6] Put differently, if N is the number of molecules per unit volume, $P = Np$ and $\mathcal{E}$ are out of phase, as indicated in Figure 7.13a. At low frequencies, $\omega\tau \ll 1$, $\alpha_d(\omega)$ is nearly $\alpha_d(0)$, and p is in phase with $\mathcal{E}$. The rate of relaxation $1/\tau$ is much faster than the frequency of the field or the rate at which the polarization is being changed; p then closely follows $\mathcal{E}$. At very high frequencies, $\omega\tau \gg 1$, the rate of relaxation $1/\tau$ is much slower than the frequency of the field and p can no longer follow the variations in the field.

We can easily obtain the dielectric constant ε_r from $\alpha_d(\omega)$ by using Equation 7.14, which then leads to a complex number for ε_r since α_d itself is a complex number. By convention, we generally write the **complex dielectric constant** as

$$\varepsilon_r = \varepsilon_r' - j\varepsilon_r'' \qquad [7.27]$$

Complex relative permittivity

where ε_r' is the real part and ε_r'' is the imaginary part, both being frequency dependent, as shown in Figure 7.13b. The real part ε_r' decreases from its maximum value $\varepsilon_r'(0)$, corresponding to $\alpha_d(0)$, to 1 at high frequencies when $\alpha_d = 0$ as $\omega \to \infty$ in Equation 7.26. The imaginary part $\varepsilon_r''(\omega)$ is zero at low and high frequencies but peaks when $\omega\tau = 1$ or when $\omega = 1/\tau$. The real part ε_r' represents the relative permittivity that we would use in calculating the capacitance, as for example in $C = \varepsilon_r \varepsilon_o A/d$. The imaginary part $\varepsilon_r''(\omega)$ represents the energy lost in the dielectric medium as the dipoles are oriented against random collisions one way and then the other way and so on by the field. Consider

[6] The polarization P lags behind $\mathcal{E}$ by some angle ϕ, that is determined by Equation 7.26 as shown in Figure 7.13.

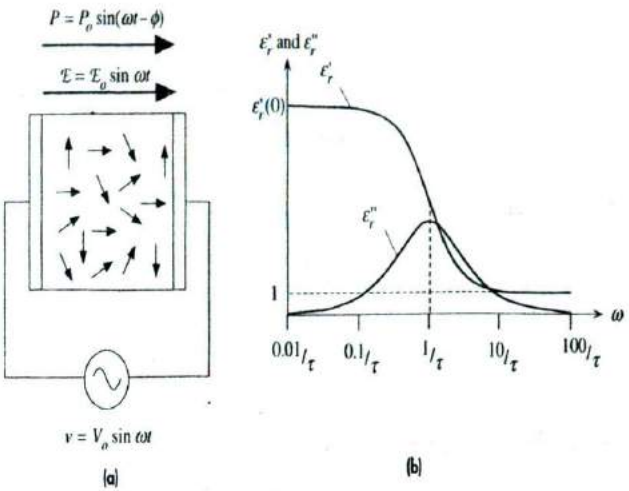


Figure 7.13
(a) An ac field is applied to a dipolar medium. The polarization P ($P = Np$) is out of phase with the ac field.
(b) The relative permittivity is a complex number with real (ε_r') and imaginary (ε_r'') parts that exhibit relaxation at $\omega \approx 1/\tau$.

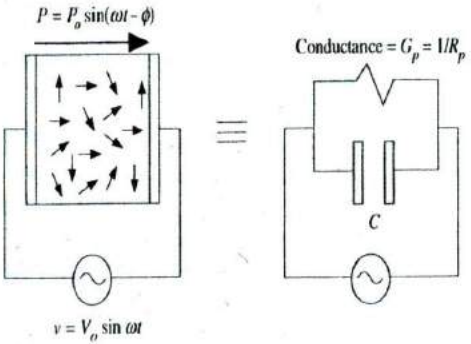


Figure 7.14 The dielectric medium behaves like an ideal (lossless) capacitor of capacitance C, which is in parallel with a conductance G_P.

the capacitor in Figure 7.14, which has this dielectric medium between the plates. Then the admittance Y, *i.e.*, the reciprocal of impedance of this capacitor, with ε_r given in Equation 7.27 is

Admittance of a parallel plate capacitor

$$Y = \frac{j\omega A\varepsilon_o\varepsilon_r(\omega)}{d} = \frac{j\omega A\varepsilon_o\varepsilon_r'(\omega)}{d} + \frac{\omega A\varepsilon_o\varepsilon_r''(\omega)}{d}$$

which can be written as

$$Y = j\omega C + G_P \qquad [7.28]$$

where

$$C = \frac{A\varepsilon_o\varepsilon_r'}{d} \qquad [7.29]$$

Equivalent ideal capacitance

and

$$G_P = \frac{\omega A\varepsilon_o\varepsilon_r''}{d} \qquad [7.30]$$

Equivalent parallel conductance

is a real number just as if we had a conductive medium with some conductance G_P or resistance $1/G_P$. The admittance of the dielectric medium according to Equation 7.28 is a parallel combination of an ideal, or lossless, capacitor C, with a relative permittivity ε_r', and a resistance of $R_P = 1/G_P$ as indicated in Figure 7.14. Thus the dielectric medium behaves as if C_o and R_P were in parallel. There is no real electric power dissipated in C, but there is indeed real power dissipated in R_P because

$$\text{Input power} = IV = YV^2 = j\omega CV^2 + \frac{V^2}{R_P}$$

and the second term is real. Thus the power dissipated in the dielectric medium is related to ε_r'' and peaks when $\omega = 1/\tau$. The rate of energy storage by the field is determined by ω whereas the rate of energy transfer to molecular collisions is determined by $1/\tau$. When $\omega = 1/\tau$, the two processes, energy storage by the field and energy transfer to random collisions, are then occurring at the same rate, and hence energy is being transferred to heat most efficiently. The peak in ε_r'' versus ω is called a **relaxation peak,** which is at a frequency when the dipole relaxations are at the right rate for maximum power dissipation. This process is known as **dielectric resonance.**

According to Equation 7.28, the magnitude of G_P and hence the energy loss is determined by ε_r''. In engineering applications of dielectrics in capacitors, we would like to minimize ε_r'' for a given ε_r'. We define the relative magnitude of ε_r'' with respect to ε_r' through a quantity, $\tan\delta$, called the **loss tangent** (or **loss factor**), as

$$\tan\delta = \frac{\varepsilon_r''}{\varepsilon_r'} \qquad [7.31]$$

Loss tangent

which is frequency dependent and peaks just beyond $\omega = 1/\tau$. The actual value of $1/\tau$ depends on the material, but typically for liquid and solid media it is in the gigahertz range, that is, microwave frequencies. We can easily find the energy per unit time—power—dissipated as dielectric loss in the medium. The resistance R_P represents the dielectric loss, so

$$W_{vol} = \frac{\text{Power loss}}{\text{Volume}} = \frac{V^2}{R_P} \times \frac{1}{dA} = \frac{V^2}{\dfrac{d}{\omega A\varepsilon_o\varepsilon_r''}} \times \frac{1}{dA} = \frac{V^2}{d^2}\omega\varepsilon_o\varepsilon_r''$$

Using Equation 7.31 and $\mathcal{E} = V/d$, we obtain

$$W_{vol} = \omega\mathcal{E}^2\varepsilon_o\varepsilon_r'\tan\delta \qquad [7.32]$$

Dielectric loss per unit volume

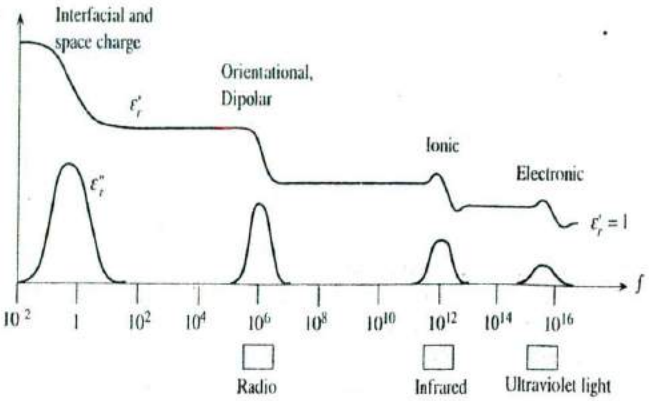


Figure 7.15 The frequency dependence of the real and imaginary parts of the dielectric constant in the presence of interfacial, orientational, ionic, and electronic polarization mechanisms.

Equation 7.32 represents the power dissipated per unit volume in the polarization mechanism: energy lost per unit time to random molecular collisions as heat. It is clear that dielectric loss is influenced by three factors: ω, $\mathcal{E}$, and $\tan\delta$.

Although we considered only orientational polarization, in general a dielectric medium will also exhibit other polarization mechanisms and certainly electronic polarization since there will always be electron clouds around individual atoms, or electrons in covalent bonds. If we were to consider the ionic polarizability in ionic solids, we would also find α_I to be frequency dependent and a complex number. In this case, lattice vibrations in the crystal, typically at frequencies ω_I in the infrared region of the electromagnetic spectrum, will dissipate the energy stored in the induced dipole moments just as energy was dissipated by molecular collisions in the gaseous dipolar medium. Thus, the energy loss will be greatest when the frequency of the polarizing field is the same as the lattice vibration frequency, $\omega = \omega_I$, which tries to randomize the polarization.

We can represent the general features of the frequency dependence of the real and imaginary parts of the dielectric constant as in Figure 7.15. Although the figure shows distinctive peaks in ε_r'' and transition features in ε_r', in reality these peaks and various features are broader. First, there is no single well-defined lattice vibration frequency but instead an allowed range of frequencies just as in solids where there is an allowed range of energies for the electron. Moreover, the polarization effects depend on the crystal orientation. In the case of polycrystalline materials, various peaks in different directions overlap to exhibit a broadened overall peak. At low frequencies the interfacial or space charge polarization features are even broader because there can be a number of conduction mechanisms (different species of charge carriers and different carrier mobilities) for the charges to accumulate at interfaces, each having its own speed. Orientational polarization, especially in many liquid dielectrics at room temperature, typically takes place at radio to microwave frequencies. In some polymeric materials, this type of polarization involves a limited rotation of dipolar side groups

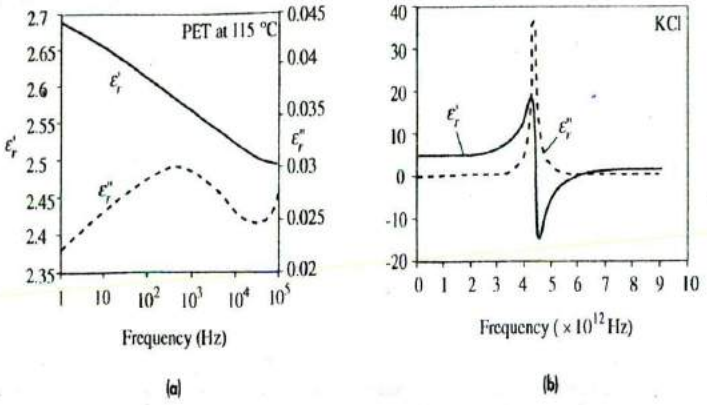


Figure 7.16 Real and imaginary parts of the dielectric constant, ε_r' and ε_r'', versus frequency for (a) a polymer, PET, at 115 °C and (b) an ionic crystal, KCl, at room temperature.

Both exhibit relaxation peaks but for different reasons.

SOURCE: Data for (a) from author's own experiments using a dielectric analyzer (DEA), (b) from C. Smart, G. R. Wilkinson, A. M. Karo, and J. R. Hardy, International Conference on Lattice Dynamics, Copenhagen, 1963, as quoted by D. H. Martin, "The Study of the Vibration of Crystal Lattices by Far Infra-Red Spectroscopy," *Advances in Physics*, 14, no. 53–56, 1965, pp. 39–100.

attached to the polymeric chain and can occur at much lower frequencies depending on the temperature. Figure 7.16 shows two typical examples of dielectric behavior, ε_r' and ε_r'' as a function of frequency, for a polymer (PET) and an ionic crystal (KCl). Both exhibit loss peaks, peaks in ε_r'' versus frequency, but for different reasons. The particular polymer, PET (a polyester), exhibits orientational polarization due to dipolar side groups, whereas KCl exhibits ionic polarization due to the displacement of K^+ and Cl^- ions. The frequency of the loss peak in the case of orientational polarization is highly temperature dependent. For the PET example in Figure 7.16 at 115 °C, the peak occurs at around 400 Hz, even below typical radio frequencies.

DIELECTRIC LOSS PER UNIT CAPACITANCE AND THE LOSS ANGLE δ Obtain the dielectric loss per unit capacitance in a capacitor in terms of the loss tangent. Obtain the phase difference between the current through the capacitor and that through R_P. What is the significance of δ?

EXAMPLE 7.5

SOLUTION

We consider the equivalent circuit in Figure 7.14. The power loss in the capacitor is due to R_P. If V is the rms value of the voltage across the capacitor, then the power dissipated per unit capacitance W_{cap} is

$$W_{cap} = \frac{V^2}{R_P} \times \frac{1}{C} = V^2 \frac{\omega \varepsilon_0 \varepsilon_r'' A}{d} \times \frac{d}{\varepsilon_0 \varepsilon_r' A} = V^2 \frac{\omega \varepsilon_r''}{\varepsilon_r'}$$

or

$$W_{cap} = V^2 \omega \tan \delta$$

Table 7.3 Dielectric properties of three insulators

	$f = 60$ Hz			$f = 1$ MHz		
Material	ε_r'	$\tan\delta$	$\omega \tan\delta$	ε_r'	$\tan\delta$	$\omega \tan\delta$
Polycarbonate	3.17	9×10^{-4}	0.34	2.96	1×10^{-2}	6.2×10^{4}
Silicone rubber	3.7	2.25×10^{-2}	8.48	3.4	4×10^{-3}	2.5×10^{4}
Epoxy with mineral filler	5	4.7×10^{-2}	17.7	3.4	3×10^{-2}	18×10^{4}

As $\tan\delta$ is frequency dependent and peaks at some frequency, so does the power dissipated per unit capacitance. A clear design objective would be to keep W_{cap} as small as possible. Further, for a given voltage, W_{cap} does not depend on the dielectric geometry. For a given voltage and capacitance, we therefore cannot reduce the power dissipation by simply changing the dimensions of the dielectric.

Consider the rms currents through R_P and C, I_{loss} and I_{cap} respectively, and their ratio,[7]

$$\frac{I_{loss}}{I_{cap}} = \frac{V}{R_P} \times \frac{1}{\frac{j\omega C}{V}} = \frac{\omega \varepsilon_o \varepsilon_r'' A}{d} \times \frac{d}{j\omega \varepsilon_o \varepsilon_r' A} = -j \tan\delta$$

As expected, the two are 90° out of phase ($-j$) and the loss current (through R_P) is a factor, $\tan\delta$, of the capacitive current (through C). The ratio of I_{cap} and the total current, $I_{total} = I_{cap} + I_{loss}$, is

$$\frac{I_{cap}}{I_{total}} = \frac{I_{cap}}{I_{cap} + I_{loss}} = \frac{1}{1 + \frac{I_{loss}}{I_{cap}}} = \frac{1}{1 - j \tan\delta}$$

The phase angle between I_{cap} and I_{total} is determined by the negative of the phase of the denominator term ($1 - j \tan\delta$). Thus the phase angle between I_{cap} and I_{total} is δ, where I_{cap} leads I_{total} by δ. δ is also called the **loss angle.** When the loss angle is zero, I_{cap} and I_{total} are equal and there is no loss in the dielectric.

EXAMPLE 7.6 **DIELECTRIC LOSS PER UNIT CAPACITANCE** Consider the three dielectric materials listed in Table 7.3 with their dielectric constant ε_r' (usually simply stated as ε_r) and loss factors $\tan\delta$. At a given voltage, which dielectric will have the lowest power dissipation per unit capacitance at 60 Hz? Is this also true at 1 MHz?

SOLUTION

The power dissipated at a given voltage per unit capacitance depends only on $\omega \tan\delta$, so we do not need to use ε_r'. Calculating $\omega \tan\delta$ or $(2\pi f) \tan\delta$, we find the values listed in the table at 60 Hz and 1 MHz. At 60 Hz, polycarbonate has the lowest power dissipation per unit capacitance, but at 1 MHz it is silicone rubber.

[7] These currents are phasors, each with a rms magnitude and phase angle.

Table 7.4 Dielectric loss per unit volume for two insulators (κ is the thermal conductivity)

	$f = 60$ Hz			$f = 1$ MHz			
Material	ε_r'	$\tan\delta$	Loss (mW cm^{-3})	ε_r'	$\tan\delta$	Loss (W cm^{-3})	κ (W cm^{-1} K^{-1})
XLPE	2.3	3×10^{-4}	0.230	2.3	4×10^{-4}	5.12	0.005
Alumina	8.5	1×10^{-3}	2.84	8.5	1×10^{-3}	47.3	0.33

EXAMPLE 7.7

DIELECTRIC LOSS AND FREQUENCY Calculate the heat generated per second due to dielectric loss per cm^3 of cross-linked polyethylene, XLPE (typical power cable insulator), and alumina, Al_2O_3 (typical substrate in thin- and thick-film electronics), at 60 Hz and 1 MHz at a field of 100 kV cm^{-1}. Their properties are given in Table 7.4. What is your conclusion?

SOLUTION

The power dissipated per unit volume is

$$W_{vol} = (2\pi f)E^2\varepsilon_o\varepsilon_r' \tan\delta$$

We can calculate W_{vol} by substituting the properties of individual dielectrics at the given frequency f. For example, for XLPE at 60 Hz,

$$W_{vol} = (2\pi\, 60\text{ Hz})(100 \times 10^3 \times 10^2\text{ V m}^{-1})^2(8.85 \times 10^{-12}\text{ F m}^{-1})(2.3)(3 \times 10^{-4})$$

$$= 230\text{ W m}^{-3}$$

We can convert this into per cm^3 by

$$W_{vol}' = \frac{W_{vol}}{10^6} = 0.230\text{ mW cm}^{-3}$$

which is shown in Table 7.4.

From similar calculations we can obtain the heat generated per second per cm^3 as shown in Table 7.4. The heats at 60 Hz are small. The thermal conductivity of the insulation and its connecting electrodes can remove the heat without substantially increasing the temperature of the insulation. At 1 MHz, the heats generated are not trivial. One has to remove 5.12 W of heat from 1 cm^3 of XLPE and 47.3 W from 1 cm^3 of alumina. The thermal conductivity κ of XLPE is about 0.005 W cm^{-1} K^{-1}, whereas that of alumina is almost 100 times larger, 0.33 W cm^{-1} K^{-1}. The poor thermal conductivity of polyethylene means that 5.12 W of heat cannot be conducted away easily and it will raise the temperature of the insulation until dielectric breakdown ensues. In the case of alumina, 47.3 W of heat will substantially increase the temperature. *Dielectric loss is the mechanism by which microwave ovens heat food.* Dielectric heating at high frequencies is used in industrial applications such as heating plastics and drying wood.

7.4.2 DEBYE EQUATIONS, COLE–COLE PLOTS, AND EQUIVALENT SERIES CIRCUIT

Consider a dipolar dielectric in which there are both orientational and electronic polarizations, α_d and α_e, respectively, contributing to the overall polarizability. Electronic polarization α_e will be independent of frequency over the typical frequency range of

operation of a dipolar dielectric, well below optical frequencies. At high frequencies, orientational polarization will be too sluggish too respond, $\alpha_d = 0$, and the ε_r will be $\varepsilon_{r\infty}$. (The subscript "infinity" simply means high frequencies where orientational polarization is negligible.) The dielectric constant and polarizabilities are generally related through[8]

Dielectric constant of a dipolar material

$$\varepsilon_r = 1 + \frac{N}{\varepsilon_0}\alpha_e + \frac{N}{\varepsilon_0}\alpha_d(\omega) = \varepsilon_{r\infty} + \frac{N}{\varepsilon_0}\alpha_d(\omega)$$

where we have combined 1 and α_e terms to represent the high frequency ε_r as $\varepsilon_{r\infty}$. Further $N\alpha_d(0)/\varepsilon_0$ determines the contribution of orientational polarization to the ...tic dielectric constant ε_{rdc}, so that $N\alpha_d(0)/\varepsilon_0$ is simply $(\varepsilon_{rdc} - \varepsilon_{r\infty})$. Substituting for the frequency dependence of $\alpha_d(\omega)$ from Equation 7.26, and writing ε_r in terms of real and imaginary parts,

Dipolar dielectric constant

$$\varepsilon_r' - j\varepsilon_r'' = \varepsilon_{r\infty} + \frac{N}{\varepsilon_0}\frac{\alpha_d(0)}{1 + j\omega\tau} = \varepsilon_{r\infty} + \frac{(\varepsilon_{rdc} - \varepsilon_{r\infty})}{1 + j\omega\tau} \qquad [7.33]$$

Debye equations for real and imaginary parts

We can eliminate the complex denominator by multiplying both the denominator and numerator of the right-hand side by $1 - j\omega\tau$ and equate real and imaginary parts to obtain what are known as **Debye equations:**

$$\varepsilon_r' = \varepsilon_{r\infty} + \frac{\varepsilon_{rdc} - \varepsilon_{r\infty}}{1 + (\omega\tau)^2} \qquad [7.34a]$$

Debye equations for real and imaginary parts

and

$$\varepsilon_r'' = \frac{(\varepsilon_{rdc} - \varepsilon_{r\infty})(\omega\tau)}{1 + (\omega\tau)^2} \qquad [7.34b]$$

Equations 7.34a and b reflect the behavior of ε_r' and ε_r'' as a function of frequency shown in Figure 7.13b. The imaginary part ε_r'' that represents the dielectric loss exhibits a peak at $\omega = 1/\tau$ which is called a **Debye loss peak.** Many dipolar gases and some liquids with dipolar molecules exhibit this type of behavior. In the case of solids the peak is typically much broader because we cannot represent the losses in terms of just one single well-defined relaxation time τ; the relaxation in the solid is usually represented by a distribution of relaxation times. Further, the simple relaxation process that is described in Equation 7.25 assumes that the dipoles do not influence each other either through their electric fields or through their interactions with the lattice; that is, they are not coupled. In solids, the dipoles can also couple, which complicates the relaxation process. Nonetheless, there are also many solids whose dielectric relaxation can be approximated by a nearly Debye relaxation or by slightly modifying Equation 7.33.

In dielectric studies of materials it is quite common to find a plot of the imaginary part (ε_r'') versus the real part (ε_r') as a function of frequency ω. Such plots are called **Cole–Cole plots** after their originators. The Debye equations 7.34a and b obviously

[8] This simple relationship is used because the Lorentz local field equation does not apply in dipolar dielectrics and the local field problem is particularly complicated in these dielectrics.

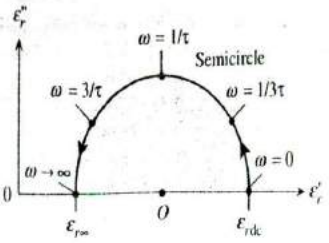


Figure 7.17 Cole–Cole plot is a plot of ε_r'' versus ε_r' as a function of frequency ω. As the frequency is changed from low to high, the plot traces out a semicircle.

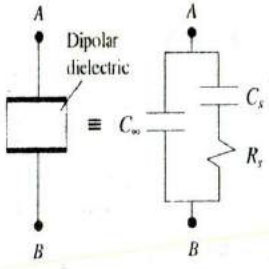


Figure 7.18 A capacitor with a dipolar dielectric and its equivalent circuit in terms of an ideal Debye relaxation.

provide the necessary values for ε_r' and ε_r'' to be plotted for the present simple dipolar relaxation mechanism that has only a single relaxation time τ. In fact, by simply putting in $\tau = 1$ second, we can calculate and plot ε_r'' versus ε_r' for $\omega = 0$ (dc) to $\omega \to \infty$ as shown in Figure 7.17. The result is a *semicircle*. While for certain substances, such as gases and some liquids, the Cole–Cole plots do indeed generate a semicircle, for many dielectrics, the curve is typically flattened and asymmetric, and not a semicircle.[9]

The Debye equations lead to a particular *RC* circuit representation of a dielectric material that is quite useful. Suppose that we have a resistance R_s in series with a capacitor C_s, both of which are in parallel with the capacitor C_∞ as in Figure 7.18. If we were to write down the equivalent admittance of this circuit, we would find that it corresponds to Equation 7.33, that is, the Debye equation. (The circuit mathematics is straightforward and is not reproduced here.) The reader may wonder why this circuit is different than the general model shown in Figure 7.14. Any series R_s and C_s circuit can be transformed to be equivalent to a parallel R_p and C_p (or G_p and C in Figure 7.14) circuit as is well known in circuit theory; the relationships between the elements depend on the frequency. Many electrolytic capacitors are frequently represented by an equivalent series R_s and C_s circuit as in Figure 7.18. If A is the area and d is the thickness of a parallel plate capacitor with a dipolar dielectric, then

Equivalent circuit of a Debye dielectric

$$C_\infty = \frac{\varepsilon_0 \varepsilon_{r\infty} A}{d} \qquad C_s = \frac{\varepsilon_0(\varepsilon_{rdc} - \varepsilon_{r\infty})A}{d} \qquad \text{and} \qquad R_s = \frac{\tau}{C_s} \qquad [7.35]$$

Notice that in this circuit model, R_s, C_s, and C_∞ do not depend on the frequency, which is only true for an ideal Debye dielectric, that with a single relaxation time τ.

[9] The departure is simply due to the fact that a simple relaxation process with a single relaxation time cannot describe the dielectric behavior accurately. [A good overview of non-Debye relaxations is given by Andrew Jonscher in *J. Phys D*, **32**, R57, 1999.]

EXAMPLE 7.8 **NEARLY DEBYE RELAXATION** There are some dielectric solids that exhibit nearly Debye relaxation. One example is the $La_{0.7}Sr_{0.3}MnO_3$ ceramic whose relaxation peak and Cole-Cole plots are similar to those shown in Figures 7.13b and 7.17,[10] especially in the high-frequency range past the resonance peak. $La_{0.7}Sr_{0.3}MnO_3$'s low frequency (ε_{rdc}) and high frequency ($\varepsilon_{r\infty}$) dielectric constants are 3.6 and 2.58, respectively, where *low* and *high* refer, respectively, to frequencies far below and above the Debye relaxation peak, *i.e.*, ε_{rdc} and $\varepsilon_{r\infty}$. The Debye loss peak occurs at 6 kHz. Calculate ε_r' and the dielectric loss factor $\tan\delta$ at 29 kHz.

SOLUTION

The loss peak occurs when $\omega_o = 1/\tau$, so that $\tau = 1/\omega_o = 1/(2\pi 6000) = 26.5\ \mu s$. We can now calculate the real and imaginary parts of ε_r at 29 kHz,

$$\varepsilon_r' = \varepsilon_{r\infty} + \frac{\varepsilon_{rdc} - \varepsilon_{r\infty}}{1 + (\omega\tau)^2} = 2.58 + \frac{3.6 - 2.58}{1 + [(2\pi)(29 \times 10^3)(26.5 \times 10^{-6})]^2} = 2.62$$

$$\varepsilon_r'' = \frac{(\varepsilon_{rdc} - \varepsilon_{r\infty})(\omega\tau)}{1 + (\omega\tau)^2} = \frac{(3.6 - 2.58)[(2\pi)(29 \times 10^3)(26.5 \times 10^{-6})]}{1 + [(2\pi)(29 \times 10^3)(26.5 \times 10^{-6})]^2} = 0.202$$

and hence

$$\tan\delta = \frac{\varepsilon_r''}{\varepsilon_r'} = \frac{0.202}{2.62} = 0.077$$

which is close to the experimental value of 0.084.

This example was a special case of nearly Debye relaxation. Debye equations have been modified over the years to account for the broad relaxation peaks that have been observed, particularly in polymeric dielectric, by writing the complex ε_r as

Non-Debye relaxation

$$\varepsilon_r = \varepsilon_{r\infty} + \frac{\varepsilon_{rdc} - \varepsilon_{r\infty}}{[1 + (j\omega\tau)^\alpha]^\beta} \qquad [7.36]$$

where α and β are constants, typically less than unity (setting $\alpha = \beta = 1$ generates the Debye equations). Such equations are useful in engineering for predicting ε_r at any frequency from a few known values at various frequencies, as highlighted in this simple nearly Debye example. Further, if τ dependence on the temperature T is known (often τ is thermally activated), then we can predict ε_r at any ω and T.

7.5 GAUSS'S LAW AND BOUNDARY CONDITIONS

An important fundamental theorem in electrostatics is Gauss's law, which relates the integration of the electric field over a surface to the total charge enclosed. It can be derived from Coulomb's law, or the latter can be derived from Gauss's law. Suppose $\mathcal{E}_n$ is the electric field normal to a small surface area dA on a closed surface, as shown in Figure 7.19; then summing $\mathcal{E}_n\,dA$ products over the whole surface gives total net charge Q_{total} inside it,

Gauss's law

$$\oint_{Surface} \mathcal{E}_n\,dA = \frac{Q_{total}}{\varepsilon_o} \qquad [7.37]$$

[10] Z. C. Kia *et al.*, *J. Phys. Cond. Matter*, **13**, 4359, 2001. The origin of the dipolar activity in this ceramic is quite complex and involves an electron hopping (jumping) from a Mn^{3+} to Mn^{4+} ion; we do not need the physical details in the example.

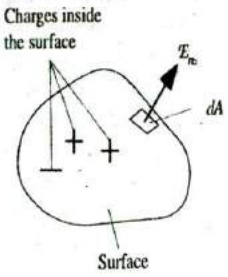


Figure 7.19 Gauss's law.
The surface integral of the electric field normal to the surface is the total charge enclosed. The field is positive if it is coming out, negative if it is going into the surface.

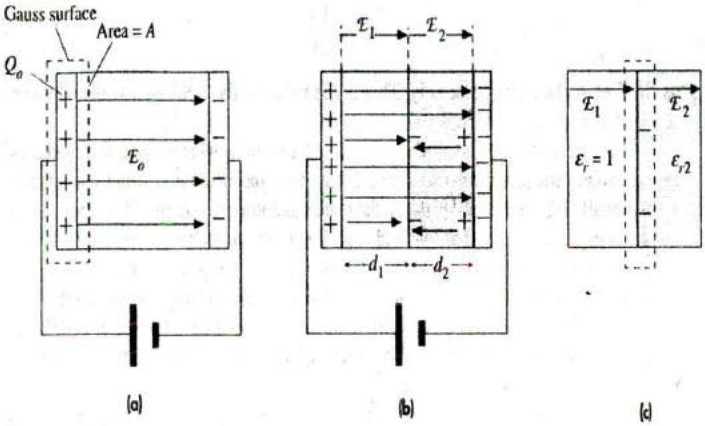


Figure 7.20
(a) The Gauss surface is a very thin rectangular surface just surrounding the positive electrode and enclosing the positive charges Q_o. The field cuts only the face just inside the capacitor.
(b) A solid dielectric occupies part of the distance between the plates. The vacuum (air)–dielectric boundary is parallel to the plates and normal to the fields $\mathcal{E}_1$ and $\mathcal{E}_2$.
(c) A thin rectangular Gauss surface at the boundary encloses the negative polarization charges.

where the circle on the integral sign represents integrating over the whole surface (any shape) enclosing the charges constituting Q_{total} as shown in Figure 7.19. The total charge Q_{total} includes *all charges*, both free charges and bound polarization charges. Gauss's law is one of the most useful laws for calculating electric fields in electrostatics, more so than the Coulomb law with which the reader is probably more familiar. The surface can be of any shape as long as it contains the charges. We generally choose convenient surfaces to simplify the integral in Equation 7.37, and these convenient surfaces are called Gauss surfaces. It should be noted from Figure 7.19 that the field $\mathcal{E}_n$ is coming *out* from the surface.

As an example, we can consider the field in the parallel plate capacitor in Figure 7.20a with no dielectric medium. We draw a thin rectangular Gauss surface (a hypothetical surface) just enclosing the positive electrode that contains the free charges $+Q_o$ on the plate. The field $\mathcal{E}_o$ is normal to the inner face (area A) of the Gauss surface.

Further, we can assume that $\mathcal{E}_o$ is uniform across the plate surface, which means that the integral of $\mathcal{E}_n\, dA$ in Equation 7.37 over the surface is simply $\mathcal{E}_o A$. There is no field on the other faces of this rectangular Gauss surface. Then from Equation 7.37,

$$\mathcal{E}_o A = \frac{Q_o}{\varepsilon_o}$$

which gives

$$\mathcal{E}_o = \frac{\sigma_o}{\varepsilon_o} \qquad [7.38]$$

where

$$\sigma_o = \frac{Q_o}{A}$$

is the free surface charge density. This is the same as the field we calculated using $\mathcal{E}_o = V/d$ and $Q_o = CV$.

An important application of Gauss's law is determining what happens at boundaries between dielectric materials. The simplest example is the insertion of a dielectric slab to only partially fill the distance between the plates, as shown in Figure 7.20b. The applied voltage remains the same, but the field is no longer uniform between the plates. There is an air–dielectric boundary. The field is different in the air and dielectric regions. Suppose that the field is $\mathcal{E}_1$ in the air region and $\mathcal{E}_2$ in the dielectric region. Both these fields are normal to the boundary by the choice of the dielectric shape (faces parallel to the plates). As a result of polarization, bound surface charges $+A\sigma_P$ and $-A\sigma_P$ appear on the surfaces of the dielectric slab, as shown in Figure 7.20b, where $\sigma_P = P$, the polarization in the dielectric. We draw a very narrow rectangular Gauss surface that encompasses the air–dielectric interface and hence the surface polarization charges $-A\sigma_P$ as shown in Figure 7.20c. The field coming *in* at the left face in air is $\mathcal{E}_1$ (taken as negative) and the field coming *out* at the right face in the dielectric is $\mathcal{E}_2$. The surface integral $\mathcal{E}_n\, dA$ and Gauss's law become

$$\mathcal{E}_2 A - \mathcal{E}_1 A = \frac{-(A\sigma_P)}{\varepsilon_o}$$

or

$$\mathcal{E}_1 = \mathcal{E}_2 + \frac{P}{\varepsilon_o}$$

The polarization P and the field $\mathcal{E}_2$ in the dielectric are related by

$$P = \varepsilon_o \chi_{e2} \mathcal{E}_2$$

or

$$P = \varepsilon_o(\varepsilon_{r2} - 1)\mathcal{E}_2$$

where χ_{e2} is the electrical susceptibility and ε_{r2} is the relative permittivity of the inserted dielectric. Then, substituting for P, we can relate $\mathcal{E}_1$ and $\mathcal{E}_2$,

$$\mathcal{E}_1 = \mathcal{E}_2 + (\varepsilon_{r2} - 1)\mathcal{E}_2$$

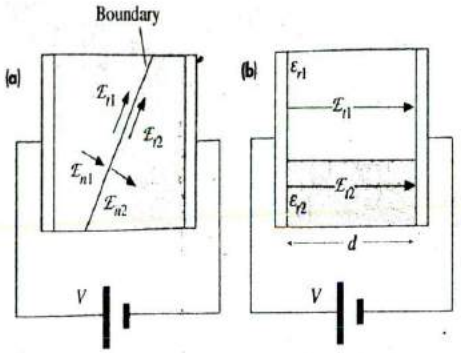


Figure 7.21
(a) Boundary conditions between dielectrics.
(b) The case for $\mathcal{E}_{t1} = \mathcal{E}_{t2}$

or

$$\mathcal{E}_1 = \varepsilon_{r2}\mathcal{E}_2$$

The field in the air part is $\mathcal{E}_1$ and the relative permittivity is 1. The example in Figure 7.20b involved a boundary between air (vacuum) and a dielectric solid, and the boundary was parallel to the plates and hence normal to the fields $\mathcal{E}_1$ and $\mathcal{E}_2$. A more general expression can be shown to relate the normal components of the electric field, shown as $\mathcal{E}_{n1}$ and $\mathcal{E}_{n2}$ in Figure 7.21a, on either side of a boundary by

General boundary condition

$$\varepsilon_{r1}\mathcal{E}_{n1} = \varepsilon_{r2}\mathcal{E}_{n2} \qquad [7.39]$$

There is a second boundary condition that relates the tangential components of the electric field, shown as $\mathcal{E}_{t1}$ and $\mathcal{E}_{t2}$ in Figure 7.21a, on either side of a boundary. These tangential fields must be equal.

General boundary condition

$$\mathcal{E}_{t1} = \mathcal{E}_{t2} \qquad [7.40]$$

We can readily appreciate this boundary condition by examining the fields in a parallel plate capacitor, which has two dielectrics longitudinally filling the space between the plates but with a boundary parallel to the field, as shown in Figure 7.21b. The field in each, $\mathcal{E}_{t1}$ and $\mathcal{E}_{t2}$, is parallel to the boundary. The voltage across each longitudinal dielectric slab is the same, and since $\mathcal{E} = dV/dx$, the field in each is the same, $\mathcal{E}_{t1} = \mathcal{E}_{t2} = V/d$.

The above boundary conditions are widely used in explaining dielectric behavior when boundaries are involved. For example, consider a small disk-shaped cavity within a solid dielectric between two electrodes, as depicted in Figure 7.22. The disk-shaped cavity has its face perpendicular to the electric field. Suppose that the dielectric length d is 1 cm and the cavity size is on the scale of micrometers. The average field within the dielectric will still be close to V/d because in integrating the field $\mathcal{E}(x)$ to find the voltage across the dielectric, the contribution from a tiny distance of a few microns will be negligible compared with contributions coming over the rest of the 1 cm. But the field within the cavity will not be the same as the average field $\mathcal{E}_1$ in the dielectric. If $\varepsilon_{r1} = 5$ for the dielectric medium and the cavity has air, then at the cavity face we have

$$\varepsilon_{r2}\mathcal{E}_2 = \varepsilon_{r1}\mathcal{E}_1$$

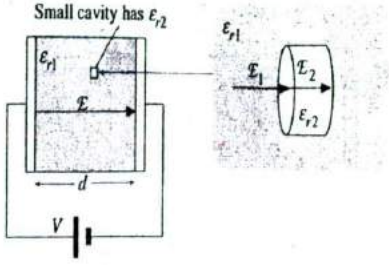


Figure 7.22 Field in the cavity is higher than the field in the solid.

which gives

$$\mathcal{E}_2 = 5\left(\frac{V}{d}\right)$$

Air insulation in a 100 micron (0.1 mm) thick cavity breaks down when $\mathcal{E}_2$ is typically 100 kV cm^{-1}. From $\mathcal{E}_2 = 5(V/d)$, a voltage of 20 kV will result in the breakdown of air in the cavity and hence a discharge current. This is called a **partial discharge** as only a partial breakdown of the insulation, that in the cavity, has occurred between the electrodes. Under an ac voltage, the discharge in the cavity can often be sustained by the capacitive current through the surrounding dielectric. Without this cavity, the dielectric would accept a greater voltage across it, which in this case is typically greater than 100 kV.

EXAMPLE 7.9 **FIELD INSIDE A THIN DIELECTRIC WITHIN A SECOND DIELECTRIC** When the dielectric fills the whole space between the plates of a capacitor, the net field within the dielectric is the same as before, $\mathcal{E} = V/d$. Explain what happens when a dielectric slab of thickness $t \ll d$ is inserted in the middle of the space between the plates, as shown in Figure 7.23. What is the field inside the dielectric?

SOLUTION

The problem is illustrated in Figure 7.23 and has symmetry in that the field in air on either side of the dielectric is the same and $\mathcal{E}_1$. The boundary conditions give

$$\varepsilon_{r1}\mathcal{E}_1 = \varepsilon_{r2}\mathcal{E}_2$$

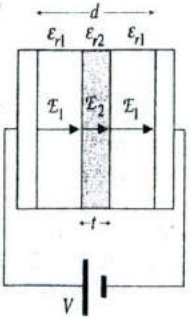


Figure 7.23 A thin slab of dielectric is placed in the middle of a parallel plate capacitor. The field inside the thin slab is $\mathcal{E}_2$.

Further, the integral of the field from one plate to the other must be V because $dV/dx = \mathcal{E}$. Examining Figure 7.23, we see that the integration is

$$\mathcal{E}_1(d - t) + \mathcal{E}_2 t = V$$

We now have to eliminate $\mathcal{E}_1$ between the previous two equations and obtain $\mathcal{E}_2$, which can be done by algebraic manipulation,

$$\mathcal{E}_2 = \frac{\varepsilon_{r1}}{\varepsilon_{r2} - \dfrac{t}{d}(\varepsilon_{r2} - \varepsilon_{r1})}\left(\frac{V}{d}\right) \qquad [7.41]$$

If $t \ll d$, then this approximates to

$$\mathcal{E}_2 = \frac{\varepsilon_{r1}}{\varepsilon_{r2}}\left(\frac{V}{d}\right) \quad \text{and} \quad \mathcal{E}_1 = \left(\frac{V}{d}\right) \qquad (t \ll d) \qquad [7.42]$$

Clearly $\mathcal{E}_1$ in the air space remains the same as the applied field V/d. Since $\varepsilon_{r1} = 1$ (air) and $\varepsilon_{r2} > 1$, $\mathcal{E}_2$ in the thin dielectric slab is smaller than the applied field V/d. On the other hand, if we have air space between two dielectric slabs, then the field in this air space will be greater than the field inside the two dielectric slabs. Indeed, if the applied voltage is sufficiently large, the field in the air gap can cause dielectric breakdown of this region.

GAUSS'S LAW WITHIN A DIELECTRIC AND FREE CHARGES Gauss's law in Equation 7.37 contains the total charge Q_{total}, enclosed within the surface. Generally, these enclosed charges are free charges Q_{free}, due to the free carriers on the electrode, and bound charges Q_P, due to polarization charges on the dielectric surface. Apply Gauss's law using a Gaussian rectangular surface enclosing the left electrode and the dielectric surface in Figure 7.24. Show that the electric field $\mathcal{E}$ in the dielectric can be expressed in terms of free charges only, Q_{free}, through

EXAMPLE 7.10

Free charges and field in a dielectric

$$\oint_{\text{Surface}} \mathcal{E}_n \, dA = \frac{Q_{\text{free}}}{\varepsilon_0 \varepsilon_r} \qquad [7.43]$$

where ε_r is the relative permittivity of the dielectric medium.

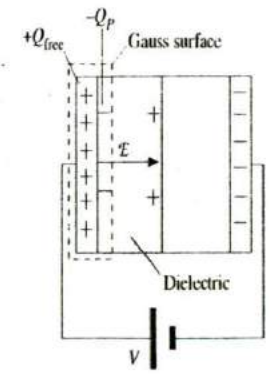


Figure 7.24 A convenient Gauss surface for calculating the field inside the dielectric is a very thin rectangular surface enclosing the surface of the dielectric.

The total charges enclosed are the free charges on the electrodes and the polarization charges on the surface of the dielectric.

SOLUTION

We apply Gauss's law to a hypothetical rectangular surface enclosing the left electrode and the dielectric surface. The field $\mathcal{E}$ in the dielectric is normal and outwards at the Gauss surface in Figure 7.24. Thus $\mathcal{E}_n = \mathcal{E}$ in the left-hand side of Equation 7.37.

$$\varepsilon_o A\mathcal{E} = Q_{\text{total}} = Q_{\text{free}} - Q_P = Q_{\text{free}} - AP = Q_{\text{free}} - A\varepsilon_o(\varepsilon_r - 1)\mathcal{E}$$

where we have used $P = \varepsilon_o(\varepsilon_r - 1)\mathcal{E}$. Rearranging,

$$\varepsilon_o\varepsilon_r A\mathcal{E} = Q_{\text{free}}$$

Since $A\mathcal{E}$ is effectively the surface integral of $\mathcal{E}_n$, the above corresponds to writing Gauss's law in a dielectric in terms of free charges as

$$\oint_{\text{Surface}} \mathcal{E}_n \, dA = \frac{Q_{\text{free}}}{\varepsilon_o\varepsilon_r}$$

The above equation assumes that polarization P and $\mathcal{E}$ are linearly related,

$$P = \varepsilon_o(\varepsilon_r - 1)\mathcal{E}$$

We note that if we only use free charges in Gauss's law, then we simply multiply ε_o by the dielectric constant of the medium. The above proof is by no means a rigorous derivation.

7.6 DIELECTRIC STRENGTH AND INSULATION BREAKDOWN

7.6.1 DIELECTRIC STRENGTH: DEFINITION

A defining property of a dielectric medium is not only its ability to increase capacitance but also, and equally important, its insulating behavior or low conductivity so that the charges are not simply conducted from one plate of the capacitor to the other through the dielectric. Dielectric materials are widely used as insulating media between conductors at different voltages to prevent the ionization of air and hence current flashovers between conductors. The voltage across a dielectric material and hence the field within it cannot, however, be increased without limit. Eventually a voltage is reached that causes a substantial current to flow between the electrodes, which appears as a short between the electrodes and leads to what is called **dielectric breakdown.** In gaseous and many liquid dielectrics, the breakdown does not generally permanently damage the material. This means that if the voltage causing breakdown is removed, then the dielectric can again sustain voltages until the voltage is sufficiently high to cause breakdown again. In solid dielectrics the breakdown process invariably leads to the formation of a permanent conducting channel and hence to permanent damage. The **dielectric strength** $\mathcal{E}_{\text{br}}$ is the maximum field that can be applied to an insulating medium without causing dielectric breakdown. Beyond $\mathcal{E}_{\text{br}}$, dielectric breakdown takes place. The dielectric strength of solids depends on a number of factors besides simply the molecular structure, such as the impurities in the material, microstructural defects (*e.g.*, microvoids), sample geometry, nature of the electrodes, temperature, and ambient conditions (*e.g.*, humidity), as well as the duration and frequency of the applied field. Dielectric strength

Table 7.5 Dielectric strength; typical values at room temperature and 1 atm

Dielectric Medium	Dielectric Strength	Comments
Atmosphere at 1 atm pressure	31.7 kV cm^{-1} at 60 Hz	1 cm gap. Breakdown by electron avalanche by impact ionization.
SF_6 gas	79.3 kV cm^{-1} at 60 Hz	Used in high-voltage circuit breakers to avoid discharges.
Polybutene	> 138 kV cm^{-1} at 60 Hz	Liquid dielectric used as oil filler and HV pipe cables.
Transformer oil	128 kV cm^{-1} at 60 Hz	
Amorphous silicon dioxide (SiO_2) in MOS technology	10 MV cm^{-1} dc	Very thin oxide films without defects. Intrinsic breakdown limit.
Borosilicate glass	10 MV cm^{-1} duration of 10 μs 6 MV cm^{-1} duration of 30 s	Intrinsic breakdown. Thermal breakdown.
Polypropylene	295–314 kV cm^{-1}	Likely to be thermal breakdown or electrical treeing.

is different under dc and ac conditions. There are also **aging effects** that slowly degrade the properties of the insulator and reduce the dielectric strength. For engineers involved in insulation, the dielectric strength of solids is therefore one of the most difficult parameters to interpret and use. For example, the breakdown field also depends on the thickness of the insulation because thicker insulators have more volume and hence a greater probability of containing a microstructural defect (*e.g.*, a microcavity) that can initiate a dielectric breakdown. Table 7.5 shows some typical dielectric strengths for various dielectrics used in electrical insulation. Unpressurized gases have lower breakdown strengths than liquids and solids.

7.6.2 DIELECTRIC BREAKDOWN AND PARTIAL DISCHARGES: GASES

Due to cosmic radiation, there are always a few free electrons in a gas. If the field is sufficiently large, then one of these electrons can be accelerated to sufficiently large kinetic energies to impact ionize a neutral gas molecule and produce an additional free electron and a positively charged gas ion. Both the first and liberated electrons are now available to accelerate in the field again and further impact ionize more neutral gas molecules, and so on. Thus, an avalanche of impact ionization processes creates many free electrons and positive gas ions in the gas, which give rise to a discharge current between the electrodes. The process is similar to avalanche breakdown in a reverse-biased *pn* junction. The breakdown in gases depends on the pressure. The concentration of gas molecules is greater at higher pressures. This means that the mean separation between molecules, and, hence, the mean free path of a free electron, is shorter. Shorter mean free paths inhibit the free electrons from accelerating to reach impact ionization energies unless the field is increased. Thus, generally, $\mathcal{E}_{br}$ increases with the gas pressure. The 60 Hz breakdown field for an air gap of 1 cm at room temperature and at atmospheric pressure is about 31.7 kV cm^{-1}. On the other hand, the gas sulfurhexafluoride, SF_6, has

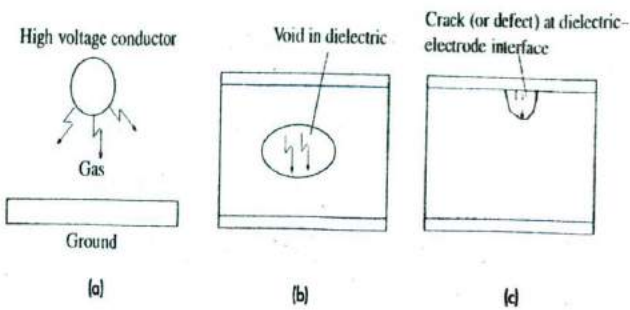


Figure 7.25

(a) The field is greatest on the surface of the cylindrical conductor facing the ground. If the voltage is sufficiently large, this field gives rise to a corona discharge.

(b) The field in a void within a solid can easily cause partial discharge.

(c) The field in the crack at the solid–metal interface can also lead to a partial discharge.

a dielectric strength of 79.3 kV cm^{-1} and an even higher strength when pressurized. SF_6 is therefore used instead of air in high-voltage circuit breakers.

A **partial discharge** occurs when only a local region of the dielectric is exhibiting discharge, so the discharge does not directly connect the two electrodes. For example, for the cylindrical conductor carrying a high voltage above a grounded plate, as in Figure 7.25a, the electric field is greatest on the surface of the conductor facing the ground. This field initiates discharge locally in this region because the field is sufficiently high to give rise to an electron avalanche effect. Away from the conductor, however, the field is not sufficiently strong to continue the electron avalanche discharge. This type of local discharge in high field regions is termed **corona discharge.** Voids and cracks occurring within solid dielectrics and discontinuities at the dielectric–electrode interface can also lead to partial discharges as the field in these voids is higher than the average field in the dielectric, and, further, the dielectric strength in the gas (*e.g.*, atmosphere) in the void is less than that of the continuous solid insulation. Figure 7.25b and c depict two examples of partial discharges occurring in voids, one inside the solid (perhaps an air or gas bubble introduced during the processing of the dielectric) and the other (perhaps in the form of a crack) at the solid–electrode interface. In practice, a variety of factors can lead to microvoids and microcavities inside solids as well as at interfaces. Partial discharges in these voids physically and chemically erode the surrounding dielectric region and lead to an overall deterioration of the dielectric strength. If uncontrolled, they can eventually give rise to a major breakdown.

7.6.3 DIELECTRIC BREAKDOWN: LIQUIDS

The processes that lead to the breakdown of insulation in liquids are not as clear as the electron avalanche effect in gases. In impure liquids with small conductive particles in suspension, it is believed that these impurities coalesce end to end to form a conducting bridge between the electrodes and thereby give rise to discharge. In some

liquids, the discharge initiates as partial discharges in gas bubbles entrapped in the liquid. These partial discharges can locally raise the temperature and vaporize more of the liquid and hence increase the size of the bubble. The eventual discharge can be a series of partial discharges in entrapped gas bubbles. Moisture absorption and absorption of gases from the ambient generally deteriorate the dielectric strength. Oxidation of certain liquids, such as oils, with time produces more acidic and hence higher conductivity inclusions or regions that eventually give discharge. In some liquids, the discharge involves the emission of a large number of electrons from the electrode into the liquid due to field emission at high fields. This is a discharge process by electrode injection.

7.6.4 DIELECTRIC BREAKDOWN: SOLIDS

There are various major mechanisms that can lead to dielectric breakdown in solids. The most likely mechanism depends on the dielectric material's condition and sometimes on extrinsic factors such as the ambient conditions, moisture absorption being a typical example.

Intrinsic Breakdown or Electronic Breakdown The most common type of electronic breakdown is an **electron avalanche breakdown.** A free electron in the conduction band (CB) of a dielectric in the presence of a large field can be accelerated to sufficiently large energies to collide with and ionize a host atom of the solid. The electron gains an energy $e\mathcal{E}_{br}\ell$ when it moves a distance ℓ under an applied field $\mathcal{E}_{br}$. If this energy is greater than the bandgap energy E_g, then the electron, as a result of a collision with the lattice vibrations, can excite an electron from the valence band to the conduction band, that is, break a bond. Both the primary and the released electron can further impact ionize other host atoms and thereby generate an electron avalanche effect that leads to a substantial current. The initial conduction electrons for the avalanche are either present in the CB or are injected from the metal into the CB as a result of field-assisted thermal emission from the Fermi energy in the metal to the CB in the dielectric. Taking typical values, $E_g \approx 5$ eV and ℓ to be of the order of the mean free path for lattice scattering, say ~50 nm, one finds $\mathcal{E}_{br} \approx 1$ MV cm^{-1}. Obviously, $\mathcal{E}_{br}$ depends on the choice of ℓ, but its order of magnitude indicates voltages that are quite large. This type of breakdown represents an upper theoretical limit that is probably approached by only certain dielectrics—those that have practically no defects. Usually, microstructural defects lead to a lower dielectric strength than the limit indicated by intrinsic breakdown. Silicon dioxide (SiO_2) films with practically no structural defects in present MOS (metal-oxide-semiconductor) capacitors (as in the gates of MOSFETs) probably exhibit an intrinsic breakdown.

If dielectric breakdown does not occur by an electron avalanche effect (perhaps due to short mean free paths in the insulator), then another insulation breakdown mechanism is the enormous increase in the injection of electrons from the metal electrode into the insulator at very high fields as a result of field-assisted emission.[11] It has

[11] The emission of electrons by tunneling from an electrode in the presence of a large field was treated in Chapter 4 as Fowler-Nordheim field emission.

been proposed that insulation breakdown under short durations in some thin polymer films is due to tunneling injection.

Thermal Breakdown Finite conductivity of the insulation means that there is Joule heat $\sigma \mathcal{E}^2$ being released within the solid. Further, at high frequencies, the dielectric loss, $V^2\omega \tan \delta$, becomes especially significant. For example, the work done by the external field in rotating the dipoles is transferred more frequently to random molecular collisions as heat as the frequency of the field increases. Both conduction and dielectric losses therefore generate heat within the dielectric. If this heat cannot be removed from the solid sufficiently quickly by thermal conduction (or by other means), then the temperature of the dielectric will increase. The increase in the temperature invariably increases the conductivity of an insulator. The increase in the conductivity then leads to more Joule heating and hence further rises in the temperature and so on. If the heat cannot be conducted away to limit the temperature, then the result is a thermal runaway condition in which the temperature and the current increase until a discharge occurs through various sections of the solid. As a consequence of sample inhomogeneities, frequently thermal runaway is severe in certain parts of the solid that become hot spots and suffer local melting and physical and chemical erosion. Hot spots are those local regions or inhomogeneities where σ or ε_r'' is larger or where the thermal conductivity is too poor to remove the heat generated. Local breakdown at various hot spots eventually leads to a conducting channel connecting the opposite electrodes and hence to a dielectric breakdown. Since it takes time to raise the temperature of the dielectric, due to the heat capacity, this breakdown process has a marked thermal lag. The time to achieve thermal breakdown depends on the heat generated, and hence on $\mathcal{E}^2$. Conversely, this means that the dielectric strength $\mathcal{E}_{br}$ depends on the duration of application of the field. For example, at 70 °C, pyrex has an $\mathcal{E}_{br}$ of typically 9 MV cm^{-1} if the applied field duration is kept short, not more than 1 ms or so. If the field is kept for 30 s, then the breakdown field is only 2.5 MV cm^{-1}. Dielectric breakdown in various ceramics and glasses at high frequencies has been attributed directly to thermal breakdown. A characteristic feature of thermal breakdown is not only the thermal lag, the time dependence, but also the temperature dependence. Thermal breakdown is facilitated by increasing the temperature of the dielectric, which means that $\mathcal{E}_{br}$ decreases with temperature.

Electromechanical Breakdown and Electrofracture A dielectric medium between oppositely charged electrodes experiences compressional forces because the opposite charges $+Q$ and $-Q$ on the plates attract each other, as depicted in Figure 7.26. As the voltage increases, so does the compressive load, and the dielectric becomes squeezed, or the thickness d gets smaller. At each stage, the increase in the compressive load is normally balanced by the elastic deformation of the insulation to a new smaller thickness. However, if the elastic modulus is sufficiently small, then compressive loads cannot be simply balanced by the elasticity of the solid, and there is a mechanical runaway for the following reasons. The decrease in d, due to the compressive load, leads to a higher field ($\mathcal{E} = V/d$) and also to more charges on the electrodes ($Q = CV$, $C = \varepsilon_o \varepsilon_r A/d$). This in turn leads to a greater compressive load, which further decreases d, and so on, until the shear stresses within the insulation cause the insulation to flow plastically (for example, by viscous deformation). Eventually, the insulation breaks down. In addition, the increase in $\mathcal{E}$ as d gets

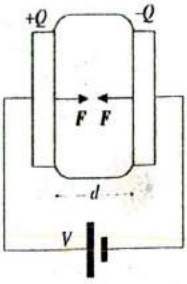
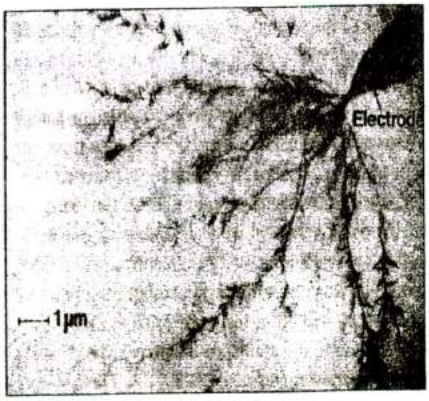


Figure 7.26 An exaggerated schematic illustration of a soft dielectric medium experiencing strong compressive forces due to the applied voltage.

smaller results in more Joule ($\sigma \mathcal{E}^2$) and dielectric-loss heating ($\omega \mathcal{E}^2 \tan \delta$) in the dielectric, which increases the temperature and hence lowers the elastic modulus and viscosity, thereby further deteriorating the mechanical stability. It is also possible for the field during the mechanical deformation of the dielectric to reach the thermal breakdown field, in which case the dielectric failure is not truly a mechanical breakdown mechanism though initiated by mechanical deformations. Another possibility is the initiation and growth of internal cracks (perhaps filamentary cracks) by internal stresses around inhomogeneous regions inside the dielectric. For example, an imperfection or a tiny cavity experiences shear stresses and also large local electric fields. Combined effects of both large shear stresses and large electric fields eventually lead to crack propagation and mechanical and, hence, dielectric failure. This type of process is sometimes called **electrofracture.** It is generally believed that certain thermoplastic polymers suffer from electromechanical dielectric breakdown, especially close to their softening temperatures. Polyethylene and polyisobutylene have been cited as examples.

Internal Discharges These are partial discharges that take place in microstructural voids, cracks, or pores within the dielectric where the gas atmosphere (usually air) has lower dielectric strength. A porous ceramic, for example, would experience partial discharges if the applied field is sufficiently large. The discharge current in a void,



Electrical breakdown by *treeing* (formation of discharge channels) in a low-density polyethylene insulation when a 50 Hz, 20 kV (rms) voltage is applied for 200 minutes to an electrode embedded in the insulation.

SOURCE: J. W. Billing and D. J. Groves, *Proceedings of the Institution of Electrical Engineers*, **212**, 1451, 1974.

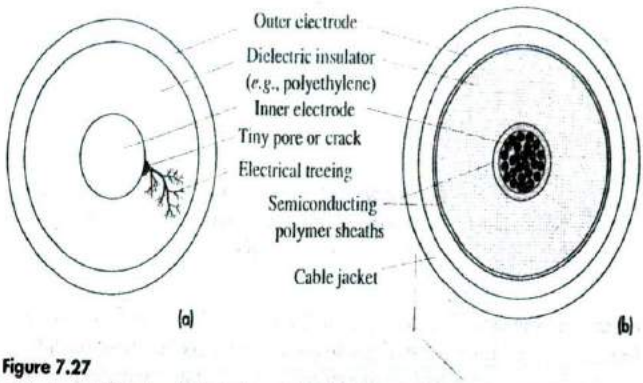


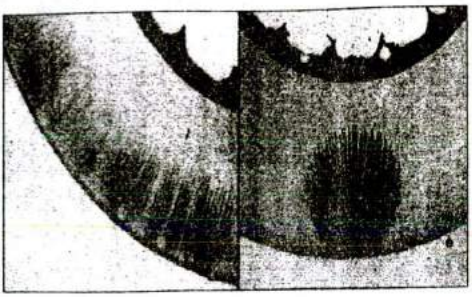
Figure 7.27

(a) A schematic illustration of electrical treeing breakdown in a high-voltage coaxial cable that was initiated by a partial discharge in the void at the inner conductor–dielectric interface.

(b) A schematic diagram of a typical high-voltage coaxial cable with semiconducting polymer layers around the inner conductor and around the outer surface of the dielectric.

such as those in Figure 7.25b and c, can be easily sustained under ac conditions, which accounts for the severity of this type of breakdown mechanism under ac conditions. Initially, the pore size (or the number of pores) may be small and the partial discharge insignificant, but with time the partial discharge erodes the internal surfaces of the void. Partial discharges can locally melt the insulator and can easily cause chemical transformations. Eventually, and usually, an **electrical tree** type of discharge develops from a partial discharge that has been eroding the dielectric, as depicted in Figure 7.27a for a high-voltage cable in which there is a tiny void at the interface between the dielectric and the inner conductor (generated perhaps by the differential thermal expansion of the electrode and polymeric insulation). The erosion of the dielectric by the partial discharge propagates like a branching tree. The "tree branches" are erosion channels—hollow filaments of various sizes—in which gaseous discharge takes place and forms a conducting channel during operation.

In the case of a coaxial high-voltage cable in Figure 7.27a, the dielectric is usually a polymer, polyethylene (PE) being one of the most popular. The electric field is maximum at the surface of the inner conductor, which is the reason for the initiation of most electrical trees near this surface. Electrical treeing is substantially controlled by having semiconductive polymer layers or sheaths surrounding the inner conductor and the outer surface of the insulator, as shown in Figure 7.27b. For flexibility, the inner conductor is frequently multicored, or stranded, rather than solid. Due to the extrusion process used to draw the insulation, the semiconductive polymer sheaths are bonded to the insulation. There are therefore practically no microvoids at the interfaces between the insulator and the semiconducting sheath. Further, these semiconducting polymer sheaths are sufficiently conductive to become "part of the electrodes." Both the conductor and the adjacent semiconductor are roughly at the same voltage, which means that there is no breakdown in the semiconductor–conductor interfaces. There is normally an outer jacket (*e.g.*, PVC) to protect the cable.



Some typical water trees found in field aged cables. Left: Trees in a cable with tape and graphite insulation. Right: Trees in a cable with strippable insulation.

SOURCE: P. Werelius, P. Tharning, R. Eriksson, B. Holmgren, J. Gafvert, "Dielectric Spectroscopy for Diagnosis of Water Tree Deterioration in XLPE Cables," *IEEE Transactions on Dielectrics and Electrical Insulation*, vol. 8, February 2001, p. 34, figure 10 (© IEEE, 2001).

Insulation Aging It is well recognized that during service, the properties of an insulating material become degraded and eventually dielectric breakdown occurs at a field below that predicted by experiments on fresh forms of the insulation. **Aging** is a term used to describe, in a general sense, the deterioration in the properties of the insulation. Aging therefore determines the useful life of the insulation. There are many factors that either directly or indirectly affect the properties and performance of an insulator in service. Even in the absence of an electric field, the insulation will experience physical and chemical aging whereby its physical and chemical properties change considerably. An insulation that is subjected to temperature and mechanical stress variations can develop structural defects, such as microcracks, which are quite damaging to the dielectric strength, as mentioned above. Irradiation by ionizing radiation such as X-rays, exposure to severe ambient conditions such as excessive humidity, ozone, and many other external conditions, through various chemical processes, deteriorate the chemical structure and properties of an insulator. This is generally much more severe for polymers than ceramics, but it is not practical to use a solid ceramic insulation in a coaxial power cable. Oxidation of a polymeric insulation with time is another form of chemical aging and is well-known to degrade the insulation performance. This is the reason for adding various antioxidants into semicrystalline polymers for use in insulation. The chemical aging processes are generally accelerated with temperature. In service, the insulation also experiences electrical aging as a result of the effects of the field on the properties of the insulation. For example, dc fields can disassociate and transport various ions in the structure and thereby slowly change the structure and properties of the insulation. Electrical trees develop as a result of electrical aging because, in service, the ac field gives rise to continual partial discharges in an internal or surface microcavity, which then erodes the region around it and slowly grows like a branching tree. In well-manufactured insulation systems, electrical treeing has been substantially reduced or eliminated from microvoids. A form of electrical aging that is currently in vogue is **water treeing,** which eventually leads to electrical treeing. The definition of a water tree, as viewed under an optical microscope, is a diffused bushy (or broccoli) type growth that consists of millions of microscopic voids (per mm^3) containing water or aqueous electrolyte. They invariably occur in moist environments and are relatively nonconducting, which means that they do not themselves lead to a direct discharge.

External Discharges There are many examples where the surface of the insulation becomes contaminated by ambient conditions such as excessive moisture, deposition

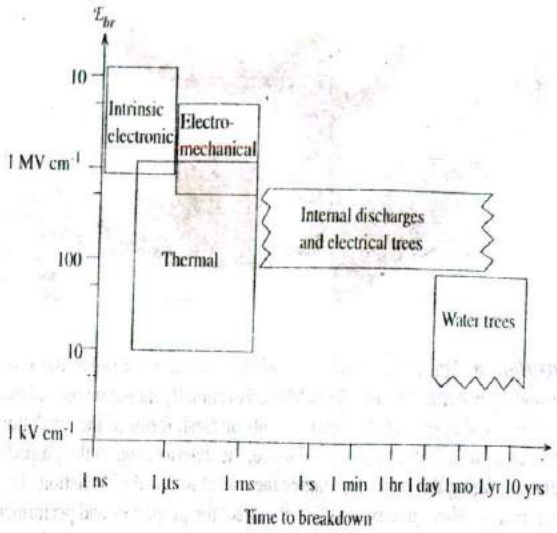


Figure 7.28 Time to breakdown and the field at breakdown $\mathcal{E}_{br}$ are interrelated and depend on the mechanism that causes the insulation breakdown.

External discharges have been excluded.

SOURCE: Based on L. A. Dissado and J. C. Fothergill, *Electrical Degradation and Breakdown in Polymers*, United Kingdom: Peter Peregrinus Ltd. for IEE, 1992, p. 63.

of pollutants, dirt, dust, and salt spraying. Eventually the contaminated surface develops sufficient conductance to allow discharge between the electrodes at a field below the normal breakdown strength of the insulator. This type of dielectric breakdown over the surface of the insulation is termed **surface tracking.**

It is apparent that there are a number of dielectric breakdown mechanisms and the one that causes eventual breakdown depends not only on the properties and quality of the material but also on the operating conditions, environmental factors being no less important. Figure 7.28 provides an illustrative diagram showing the relationship between the breakdown field and the time to breakdown. An insulation that can withstand large fields for a very short duration will break down at a lower field if the duration of the field increases. The breakdown mechanism is also likely to change from being intrinsic to being, perhaps, thermal. When insulation breakdown occurs in times beyond a few days, it is generally attributed to the degradation of the insulation, which eventually leads to a breakdown through, most probably, electrical treeing. It is also apparent that it is not possible to clearly identify a specific dielectric breakdown mechanism for a given material.

EXAMPLE 7.11 **DIELECTRIC BREAKDOWN IN A COAXIAL CABLE** Consider the coaxial cable in Figure 7.29 with a and b defining the radii of the inner and outer conductors.

a. Using Gauss's law, find the capacitance of the coaxial cable.

b. What is the electric field at r from the center of the cable ($r > a$)? Where is the field maximum?

c. Consider two candidate materials for the dielectric insulation: cross-linked polyethylene, (XLPE) and silicone rubber. Suppose that the inner conductor diameter is 5 mm and the insulation thickness is also 5 mm. What is the voltage that will cause dielectric breakdown in each insulator?

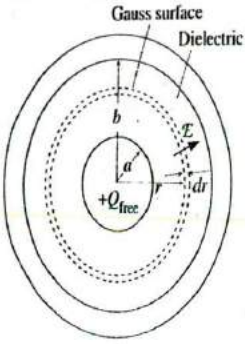


Figure 7.29 A schematic diagram for the calculation of the capacitance of a coaxial cable and the field at point r from the axis.

Consider an infinitesimally thin cylindrical shell of radius r and thickness dr in the dielectric and concentrically around the inner conductor. This surface is chosen as the Gauss surface. The voltage across the dielectric thickness dr is dV. The field $\mathcal{E} = -dV/dr$.

d. What typical voltage will initiate a partial discharge in a small air pore (perhaps formed during mechanical and thermal stressing) at the inner conductor–insulator interface? Assume that the breakdown field for air at 1 atm and gap spacing around 0.1 mm is about 100 kV cm^{-1}.

SOLUTION

Consider a cylindrical shell of thickness dr of the dielectric as shown in Figure 7.29. Suppose that the voltage across the shell thickness is dV. Then the field $\mathcal{E}$ at r is $-dV/dr$ (this is the definition of $\mathcal{E}$). Suppose that Q_{free} is the free charge on the inner conductor. We take a Gauss surface that is a cylinder of radius r and concentric with the inner conductor as depicted in Figure 7.29. The surface area A of this cylinder is $2\pi rL$ where L is the length of the cable. The field at the surface, at distance r, is $\mathcal{E}$, which is normal to A and coming out of A. Then from Equation 7.43

$$\mathcal{E}(2\pi rL) = \frac{Q_{free}}{\varepsilon_o \varepsilon_r} \qquad [7.44]$$

Thus

$$-\frac{dV}{dr} = \frac{Q_{free}}{\varepsilon_o \varepsilon_r 2\pi rL}$$

This can be integrated from $r = a$, where the voltage is V, to b, where $V = 0$. Then

$$V = -\frac{Q_{free}}{\varepsilon_o \varepsilon_r 2\pi L} \ln\left(\frac{b}{a}\right) \qquad [7.45]$$

We can obtain the capacitance of the coaxial cable from $C_{coax} = Q_{free}/V$, which is

$$C_{coax} = \frac{\varepsilon_o \varepsilon_r 2\pi L}{\ln\left(\frac{b}{a}\right)} \qquad [7.46]$$

Capacitance of a coaxial cable

The capacitance per unit length can be calculated using $a = 2.5$ mm and

$$b = a + \text{Thickness} = 7.5 \text{ mm}$$

and the appropriate dielectric constants, $\varepsilon_r = 2.3$ for XLPE and 3.7 for silicone rubber. The values are around 100–200 pF per meter, as listed in the fourth column in Table 7.6.

Table 7.6 Dielectric insulation candidates for a coaxial cable

Dielectric	ε_r (60 Hz)	Strength (60 Hz) (kV cm^{-1})	C (60 Hz) (pF m^{-1})	Breakdown Voltage (kV)	Voltage for Partial Discharge in a Microvoid (kV)
XLPE	2.3	217	116	59.6	11.9
Silicone rubber	3.7	158	187	43.4	7.4

The electric field $\mathcal{E}$ follows directly when we substitute for Q_{free} from Equation 7.45 into Equation 7.44,

Field in a coaxial cable

$$\mathcal{E} = \frac{V}{r \ln\left(\frac{b}{a}\right)} \qquad [7.47]$$

Equation 7.47 is valid for r from a to b (there is no field within the conductors). The field is maximum where $r = a$,

Maximum field in a coaxial cable

$$\mathcal{E}_{max} = \frac{V}{a \ln\left(\frac{b}{a}\right)} \qquad [7.48]$$

The breakdown voltage V_{br} is reached when this maximum field $\mathcal{E}_{max}$ reaches the dielectric strength or the breakdown field $\mathcal{E}_{br}$

Breakdown voltage

$$V_{br} = \mathcal{E}_{br} a \ln\left(\frac{b}{a}\right) \qquad [7.49]$$

The breakdown voltages calculated from Equation 7.49 are listed in the fifth column in Table 7.6. Although the values are high, it must be remembered that, due to a number of other factors such as insulation aging, one cannot expect the cable to withstand these voltages forever.

If there is an air cavity or bubble at the inner conductor to dielectric surface, then the field in this gaseous space will be $\mathcal{E}_{air} \approx \varepsilon_r \mathcal{E}_{max}$, where $\mathcal{E}_{max}$ is the field at $r = a$. Air breakdown occurs when

$$\mathcal{E}_{air} = \mathcal{E}_{air\text{-}br} = 100 \text{ kV cm}^{-1}$$

at 1 atm and 25 °C for a 0.1 mm gap. Then $\mathcal{E}_{max} \approx \mathcal{E}_{air\text{-}br}/\varepsilon_r$. The corresponding voltage from Equation 7.48 is

$$V_{air\text{-}br} \approx \frac{\mathcal{E}_{air\text{-}br}}{\varepsilon_r} a \ln\left(\frac{b}{a}\right)$$

The voltages for partial discharges for the two coaxial cables are shown in the sixth column of Table 7.6. It should be noted that these voltages will only give partial discharges contained within microvoids and will not normally lead to the immediate breakdown of the insulation. The partial discharges erode the cavities and also release vapor from the polymer that accumulates in the cavities. Thus, gaseous content and pressure in a cavity will change as the partial discharge continues. For example, the pressure buildup will increase the breakdown field and elevate the voltage for partial breakdown. Eventual degradation is likely to lead to electrical treeing.

We should also note that the actual field in the air cavity depends on the shape of the cavity, and the above treatment is only valid for a thin disk-like cavity lying perpendicular to the field (see Section 7.9, Additional Topics).

7.7 CAPACITOR DIELECTRIC MATERIALS

7.7.1 TYPICAL CAPACITOR CONSTRUCTIONS

The selection criteria of dielectric materials for capacitors depend on the capacitance value, frequency of application, maximum tolerable loss, and maximum working voltage, with size and cost being additional external constraints. Requirements for high-voltage power capacitors are distinctly different than those used in small integrated circuits. Large capacitance values are more easily obtained at low frequencies because low-frequency polarization mechanisms such as interfacial and dipolar polarization make a substantial contribution to the dielectric constant. At high frequencies, it becomes more difficult to achieve large capacitances and at the same time maintain acceptable low dielectric loss, inasmuch as the dielectric loss per unit volume is $\varepsilon_o \varepsilon'_r \omega \mathcal{E}^2 \tan \delta$.

The bar-chart diagrams in Figures 7.30 and 7.31 provide some typical examples of dielectrics for a range of capacitance values and for a range of usable frequencies. For example, electrolytic dielectrics characteristically provide capacitances between one to thousands of microfarads, but their frequency response is typically limited to below

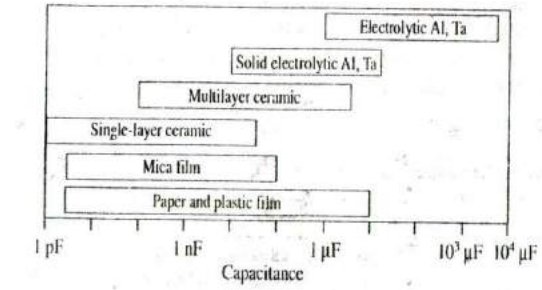


Figure 7.30 Examples of dielectrics that can be used for various capacitance values.

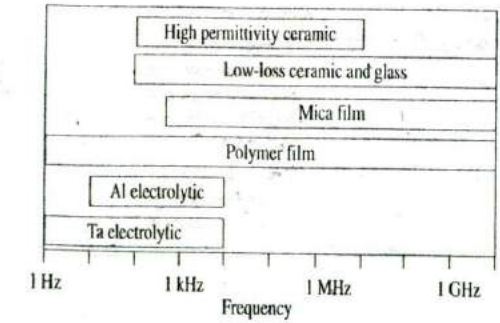


Figure 7.31 Examples of dielectrics that can be used in various frequency ranges.

known. From the ultrasonic velocity one can determine the elastic constants (Young's modulus) of the solid. Furthermore, if there are internal imperfections such as cracks in the solid, then they reflect or scatter the ultrasonic waves. These reflections can lead to echoes that can be detected by suitably located transducers. Such ultrasonic testing methods are widely used for nondestructive evaluations of solids in mechanical engineering.

It is clear that an important engineering factor in the use of piezoelectric transducers is the electromechanical coupling between electrical and mechanical energies. The **electromechanical coupling factor** k is defined in terms of k^2 by

Electromechanical coupling factor

$$k^2 = \frac{\text{Electrical energy converted to mechanical energy}}{\text{Input of electrical energy}} \qquad [7.56a]$$

or equivalently by

Electromechanical coupling factor

$$k^2 = \frac{\text{Mechanical energy converted to electrical energy}}{\text{Input of mechanical energy}} \qquad [7.56b]$$

Table 7.8 summarizes some typical piezoelectric materials with some applications. The so-called PZT ceramics are widely used in many piezoelectric applications. PZT stands for lead zirconate titanate and the ceramic is a solid solution of lead zirconate, $PbZrO_3$, and lead titanate, $PbTiO_3$, so its composition is $PbTi_{1-x}Zr_xO_3$ where x is determined by the extent of the solid solution but typically is around 0.5. PZT piezoelectric components are manufactured by sintering, which is a characteristic ceramic manufacturing process in which PZT powders are placed in a mold and subjected to a pressure at high temperatures. During sintering the ceramic powders are fused through interdiffusion. The final properties depend not only on the composition of the solid solution but also on the manufacturing process, which controls the average grain size or polycrystallinity. Electrodes are deposited onto the final ceramic component, which is then poled by the application of a temporary electric field to induce it to become

Table 7.8 Piezoelectric materials and some typical values for d and k

Crystal	d (m V^{-1})	k	Comment
Quartz (crystal SiO_2)	2.3×10^{-12}	0.1	Crystal oscillators, ultrasonic transducers, delay lines, filters
Rochelle salt ($NaKC_4H_4O_6 \cdot 4H_2O$)	350×10^{-12}	0.78	
Barium titanate ($BaTiO_3$)	190×10^{-12}	0.49	Accelerometers
PZT, lead zirconate titanate ($PbTi_{1-x}Zr_xO_3$)	480×10^{-12}	0.72	Wide range of applications including earphones, microphones, spark generators (gas lighters, car ignition), displacement transducers, accelerometers
Polyvinylidene fluoride (PVDF)	18×10^{-12}	—	Must be poled; heated, put in an electric field and then cooled. Large area and inexpensive

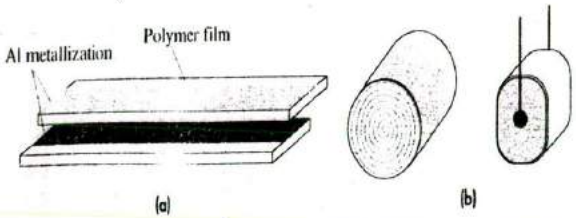


Figure 7.33 Two polymer tapes in (a), each with a metallized film electrode on the surface (offset from each other), can be rolled together (like a Swiss roll) to obtain a polymer film capacitor as in (b).
As the two separate metal films are lined at opposite edges, electroding is done over the whole side surface.

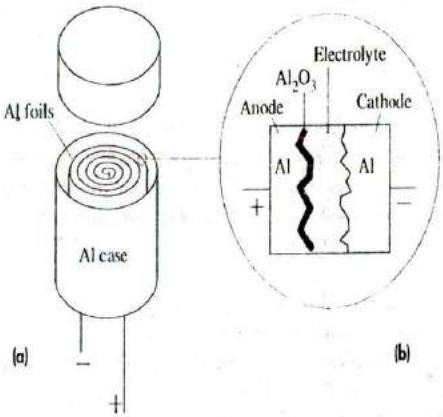


Figure 7.34 Aluminum electrolytic capacitor.

tapes have metallized electrodes (typically vacuum deposited Al) on one surface, leaving a margin on one side. These metal film electrodes have been offset in opposite directions so that they line up with the opposite sides of the tapes. The two tapes together are rolled up (like a Swiss-roll cake) and the opposite sides are electroded using suitable conducting glues or other means. The concept is therefore similar to the multilayer ceramic capacitor except that the layers are rolled up to form a circular cross section. It is also possible to cut and stack the layers as in the multilayer ceramic construction.

Electrolytic capacitors provide large values of capacitance while maintaining a tolerable size. There are various types of electrolytic capacitors. In aluminum electrolytic capacitors, the metal electrodes are two Al foils, typically 50–100 μm thick, that are separated by a porous paper medium soaked with a liquid electrolyte. The two foils together are wound into a cylindrical form and held within a cylindrical case, as shown in Figure 7.34a. Contrary to intuition, the paper-soaked electrolyte is not the dielectric. The dielectric medium is the thin alumina Al_2O_3 layer grown on the roughened surface of one of the foils, as shown in Figure 7.34b. This foil is then called the anode (+ terminal). Both Al foils are etched to obtain rough surfaces, which increases the surface area

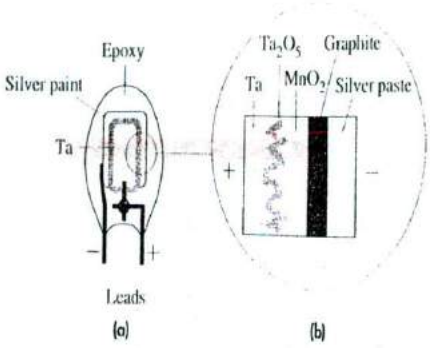


Figure 7.35 Solid electrolyte tantalum capacitor.
(a) A cross section without fine detail.
(b) An enlarged section through the Ta capacitor.

compared with smooth surfaces. The capacitor is called electrolytic because the Al_2O_3 layer is grown electrolytically on one of the foils and is typically 0.1 μm in thickness. This small thickness and the large surface area are responsible for the large capacitance. The electrolyte is conducting and serves to heal local minor breakdowns in the Al_2O_2 by an electrolytic reaction, provided that the anode has been positively biased. The capacitive behavior is due to the Al/(Al_2O_3)/electrolyte structure. Furthermore, Al/Al_2O_3 contact is like a metal to p-type semiconductor contact and has rectifying properties. It must be reverse-biased to prevent charge injection into the Al_2O_3 and hence conduction through the capacitor. Thus the Al must be connected to the positive terminal, which makes it the anode. When the electrolytic Al capacitor in Figure 7.34b is oppositely biased, it becomes conducting.

Electrolytic capacitors using liquid electrolytes tend to dry up over a long period, which is a disadvantage. **Solid electrolyte tantalum capacitors** overcome the drying-up problem by using a solid electrolyte. The structure of a typical solid Ta capacitor is shown in Figure 7.35a and b. The anode (+ electrode) is a porous (sintered) Ta pellet that has the surface anodized to obtain a thin surface layer of tantalum pentoxide, Ta_2O_5, which is the dielectric medium (with $\varepsilon_r' = 28$). The Ta pellet with Ta_2O_5 is then coated with a thick solid electrolyte, in this case MnO_2. Subsequently, graphite and silver paste layers are applied. Leads are then attached and the whole construction is molded into a resin chip. Solid tantalum capacitors are widely used in numerous electronics applications due to their small size, temperature and time stability, and high reliability.

7.7.2 DIELECTRICS: COMPARISON

The **capacitance per unit volume** C_{vol}, which characterizes the **volume efficiency** of a dielectric, can be obtained by dividing C by Ad,

Capacitance per unit volume

$$C_{vol} = \frac{\varepsilon_0 \varepsilon_r}{d^2} \qquad [7.51]$$

It is clear that large capacitances require high dielectric constants and thin dielectrics. We should note that d appears as d^2, so the importance of d cannot be understated.

Table 7.7 Comparison of dielectrics for capacitor applications

	Capacitor Name					
	Polypropylene	**Polyester**	**Mica**	**Aluminum, Electrolytic**	**Tantalum, Electrolytic, Solid**	**High-*K* Ceramic**
Dielectric	Polymer film	Polymer film	Mica	Anodized Al_2O_3 film	Anodized Ta_2O_5 film	X7R $BaTiO_3$ base
ε_r'	2.2–2.3	3.2–3.3	6.9	8.5	27	2000
$\tan\delta$	4×10^{-4}	4×10^{-3}	2×10^{-4}	0.05–0.1	0.01	0.01
$\mathcal{E}_{br}$ (kV mm^{-1}) dc	100–350	100–300	50–300	400–1000	300–600	10
d (typical minimum) (μm)	3–4	1	2–3	0.1	0.1	10
C_{vol} (μF cm^{-3})	2	30	15	7500*	24,000*	180
$R_p = 1/G_p$(kΩ) for $C = 1$ μF, $f = 1$ kHz	400	40	800	1.5–3	16	16
E_{vol} (mJ cm^{-3})†	10	15	8	1000	1200	100
Polarization	Electronic	Electronic and dipolar	Ionic	Ionic	Ionic	Large ionic displacement

* Proper volumetric calculations must also consider the volumes of electrodes and the electrolyte necessary for these dielectrics to work; hence the number would have to be decreased.

† E_{vol} depends very sensitively on $\mathcal{E}_{br}$ and the choice of η; hence it can vary substantially.

NOTES: Values are typical. Assume $\eta = 3$. The table is for comparison purposes only. Breakdown fields are typical dc values and can vary substantially, by at least an order of magnitude; $\mathcal{E}_{br}$ depends on the thickness, material quality, and the duration of the applied voltage. Polyester is PET, or polyethylene terephthalate. Mica is potassium aluminosilicate, a muscovite crystal. X7R is the name of a particular $BaTiO_3$-based ceramic solid solution.

Although mica has a higher ε_r than polymer films, the latter can be made quite thin, a few microns, which leads to a greater capacitance per unit volume. The reason that electrolytic aluminum capacitors can achieve large capacitance per unit volume is that d can be made very thin over a large surface area by using the liquid electrolyte to heal minor local dielectric breakdowns. Table 7.7 shows a selection of dielectric materials for capacitor applications and compares the "volume efficiency" C_{vol} based on a typical minimum thickness that a convenient process can handle. It is apparent that, compared with polymeric films, ceramics have substantial volume efficiency as a result of large dielectric constants (high-K ceramics) in some cases and as a consequence of a thin dielectric thickness in other cases (Al_2O_3).

Another engineering consideration in selecting a dielectric is the working voltage. Although d can be decreased to obtain large capacitances per unit volume, this also decreases the working voltage. The maximum voltage that can be applied to a capacitor depends on the breakdown field of the dielectric medium $\mathcal{E}_{br}$, which itself is a highly variable quantity. A safe working voltage must be some safety factor η less than the breakdown voltage $\mathcal{E}_{br}d$. Thus, if V_m is the maximum safe working voltage, then the maximum energy that can be stored per unit volume is given by

Maximum energy per unit volume

$$E_{vol} = \frac{1}{2}CV_m^2 \times \frac{1}{Ad} = \frac{\varepsilon_0\varepsilon_r'}{2\eta^2}\mathcal{E}_{br}^2 \qquad [7.52]$$

It is clear that both ε_r' and $\mathcal{E}_{br}$ of the dielectric are significant in determining the energy storage ability of the capacitor. Moreover, at the maximum working voltage, the rate of dielectric loss per unit volume in the capacitor becomes

Dielectric loss per unit volume

$$W_{vol} = \frac{\mathcal{E}_{br}^2}{\eta^2}\omega\varepsilon_0\varepsilon_r' \tan\delta \qquad [7.53]$$

Those materials that have relatively higher tan δ exhibit greater dielectric losses. Although dielectric losses may be small at low frequencies, at high frequencies they become quite significant. Table 7.7 compares the energy storage efficiency E_{vol} and tan δ for various dielectrics. It seems that ceramics have a better energy storage efficiency than polymers. High-K ceramics tend to have large tan δ values and suffer from greater dielectric loss. Polystyrene and polypropylene have particularly low tan δ as the polarization mechanism is due to electronic polarization and the dielectric losses are the least. Indeed, polystyrene and polypropylene capacitors have found applications in high-quality audio electronics. Equations 7.52 and 7.53 should be used with care, because the breakdown field $\mathcal{E}_{br}$ can depend on the thickness d, among many other factors, including the quality of the dielectric material. For example, for polypropylene insulation, $\mathcal{E}_{br}$ is typically quoted as roughly 50 kV mm^{-1} (500 kV cm^{-1}), whereas for thin films (*e.g.*, 25 μm), over short durations, $\mathcal{E}_{br}$ can be as high as 200 kV mm^{-1}. Further, in some cases, $\mathcal{E}_{br}$ is more suitably defined in terms of the maximum allowable leakage current, that is, a field at which the dielectric is sufficiently conducting.

The temperature stability of a capacitor is determined by the temperature dependences of ε_r' and tan δ, which are controlled by the dominant polarization mechanism. For example, polar polymers have permanent dipole groups attached to the polymer chains as in polyethyleneterephthalate (PET). In the absence of an applied field, these dipoles are randomly oriented and also restricted in their rotations by neighboring chains, as depicted in Figure 7.36a. In the presence of an applied dc field, as in Figure 7.36b, some very limited rotation enables partial dipolar (orientational) polarization to take place. Typically, at room temperature, dipolar contribution to ε_r under ac conditions, however, is small because restricted and hindered rotation prevents the dipoles to closely follow the ac field. Close to the softening temperature of the polymer, the molecular motions become easier and, further, there is more volume between chains for the dipoles to rotate. The dipolar side groups and polarized chains become capable of responding to the field. They can align with the field and also follow the field variations, as shown in Figure 7.36c. Dipolar contribution to ε_r is substantial even at high frequencies. Both ε_r' and tan δ therefore increase with temperature. Thus, polar polymers exhibit temperature dependent ε_r and tan δ, which reflect in the properties of the capacitor.

On the other hand, in nonpolar polymers such as polystyrene and polypropylene, the polarization is due to electronic polarization and ε_r and tan δ remain relatively constant. Thus polystyrene and polypropylene capacitors are more stable compared with PET (polyester) capacitors. The change in the capacitance with temperature is measured by the **temperature coefficient of capacitance** (TCC), which is defined as the fractional (or percentage) change in the capacitance per unit temperature change. The temperature controls not only ε_r but also the linear expansion of the dielectric,

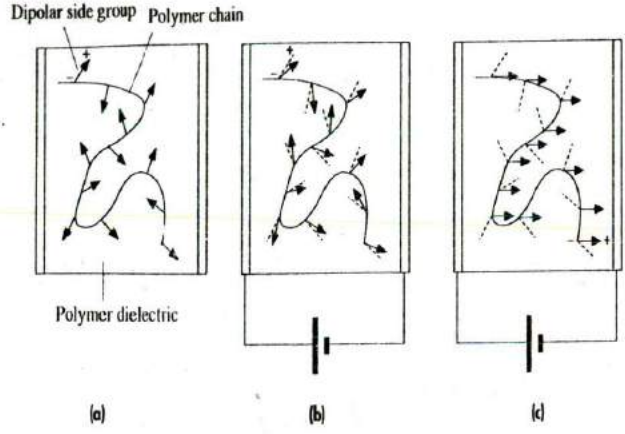


Figure 7.36

(a) A polymer dielectric that has dipolar side groups attached to the polymer chains. With no applied field, the dipoles are randomly oriented.

(b) In the presence of an applied field, some very limited rotation enables dipolar polarization to take place.

(c) Near the softening temperature of the polymer, the molecular motions are rapid and there is also sufficient volume between chains for the dipoles to align with the field. The dipolar contribution to ε_r is substantial, even at high frequencies.

which changes the dimensions A and d. For example, polystyrene, polycarbonate, and mica capacitors are particularly stable with small TCC values. Plastic capacitors are typically limited to operations well below their melting temperatures, which is one of their main drawbacks. The specified operating temperature, for example, from −55 °C to 125 °C, for many of the ceramic capacitors is often a limitation of the epoxy coating of the capacitor rather than the actual limitation of the ceramic material. In many capacitors, the working voltage has to be derated for operation at high temperatures and high frequencies because $\mathcal{E}_{bt}$ decreases with ambient temperature and the frequency of the applied field. For example, a 1000 V dc polypropylene capacitor will have a substantially lower ac working voltage, *e.g.*, 100 V at 10 kHz.

EXAMPLE 7.12

DIELECTRIC LOSS AND EQUIVALENT CIRCUIT OF A POLYESTER CAPACITOR AT 1 kHz Figure 7.37 shows the temperature dependence of ε_r' and $\tan\delta$ for a polyester film. Calculate the equivalent circuit at 25 °C at 1 kHz for a 560 pF PET capacitor that uses a 0.5 micron thick polyester film. What happens to these values at 100 °C?

SOLUTION

From Figure 7.37 at 25 °C, $\varepsilon_r' = 2.60$ and $\tan\delta \approx 0.002$. The capacitance C at 25 °C is given as 560 pF. The equivalent parallel conductance G_P, representing the dielectric loss, is given by

$$G_P = \frac{\omega A \varepsilon_0 \varepsilon_r' \tan\delta}{d} = \omega C \tan\delta$$

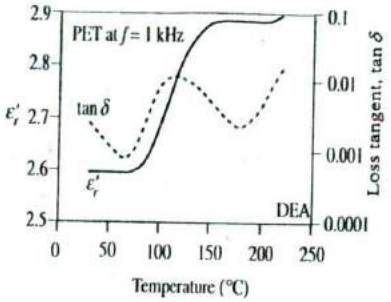


Figure 7.37 Real part of the dielectric constant ε_r' and loss tangent, tan δ, at 1 kHz versus temperature for PET.
SOURCE: Data obtained by Kasap and Maeda [1995] using a dielectric analyzer (DEA).

Substituting

$$\omega = 2\pi f = 2000\pi$$

and tan $\delta = 0.002$, we get

$$G_P = (2000\pi)(560 \times 10^{-12})(0.002) = 7.04 \times 10^{-9} \frac{1}{\Omega}$$

This is equivalent to a resistance of 142 MΩ. The equivalent circuit is an ideal (lossless) capacitor of 560 pF in parallel with a 142 MΩ resistance (this resistance value decreases with the frequency).

At 100 °C, $\varepsilon_r' = 2.69$ and tan $\delta \approx 0.01$, so the new capacitance is

$$C_{100\,^\circ C} = C_{25\,^\circ C}\frac{\varepsilon_r(100\,^\circ C)}{\varepsilon_r(25\,^\circ C)} = (560\text{ pF})\frac{2.69}{2.60} = 579\text{ pF}$$

The equivalent parallel conductance at 100 °C is

$$G_P = (2000\pi)(579 \times 10^{-12})(0.01) = 3.64 \times 10^{-8} \frac{1}{\Omega}$$

This is equivalent to a resistance of 27.5 MΩ. The equivalent circuit is an ideal (lossless) capacitor of 579 pF in parallel with a 27.5 MΩ resistance.

7.8 PIEZOELECTRICITY, FERROELECTRICITY, AND PYROELECTRICITY

7.8.1 PIEZOELECTRICITY

Certain crystals, for example, quartz (crystalline SiO_2) and $BaTiO_3$, become polarized when they are mechanically stressed. Charges appear on the surfaces of the crystal, as depicted in Figure 7.38a and b. Appearance of surface charges leads to a voltage difference between the two surfaces of the crystal. The same crystals also exhibit mechanical strain or distortion when they experience an electric field, as shown in Figure 7.38c and d. The direction of mechanical deformation (*e.g.*, extension or compression) depends on

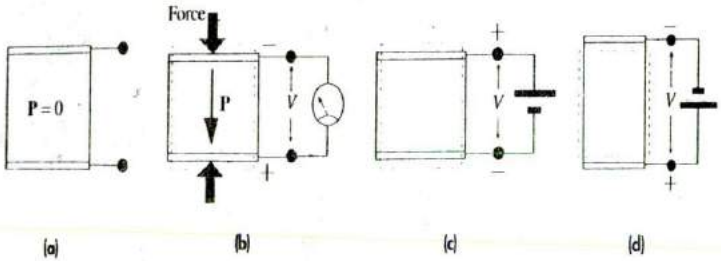


Figure 7.38 The piezoelectric effect.
(a) A piezoelectric crystal with no applied stress or field.
(b) The crystal is strained by an applied force that induces polarization in the crystal and generates surface charges.
(c) An applied field causes the crystal to become strained. In this case the field compresses the crystal.
(d) The strain changes direction with the applied field and now the crystal is extended.

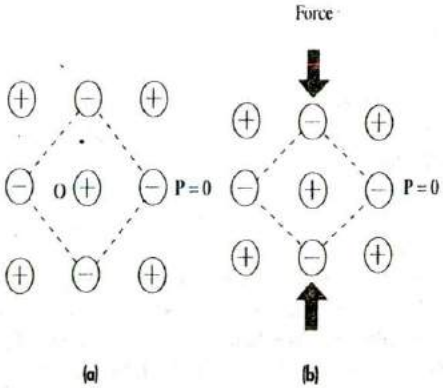


Figure 7.39 A NaCl-type cubic unit cell has a center of symmetry.
(a) In the absence of an applied force, the centers of mass for positive and negative ions coincide.
(b) This situation does not change when the crystal is strained by an applied force.

the direction of the applied field, or the polarity of the applied voltage. The two effects are complementary and define **piezoelectricity.**

Only certain crystals can exhibit piezoelectricity because the phenomenon requires a special crystal structure—that which has no center of symmetry. Consider a NaCl-type cubic unit cell in Figure 7.39a. We can describe the whole crystal behavior by examining the properties of the unit cell. This unit cell has a **center of symmetry** at O because if we draw a vector from O to any charge and then draw the reverse vector, we will find the same type of charge. Indeed, any point on any charge is a center of symmetry. Many similar cubic crystals (not all) possess a center of symmetry. When unstressed, the center of mass of the negative charges at the corners of the unit cell coincides with the positive charge at the center, as shown in Figure 7.39a. There is therefore no net polarization in the unit cell and $\mathbf{P} = 0$. Under stress, the unit cell becomes strained, as shown in Figure 7.39b, but the center of mass of the negative charges still coincides with the positive charge and the net polarization is still zero. Thus, the strained crystal

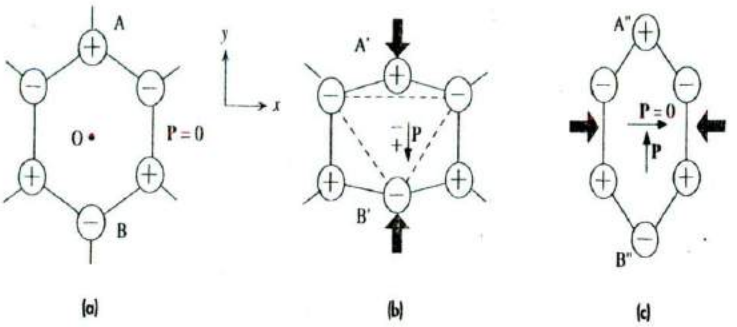


Figure 7.40 A hexagonal unit cell has no center of symmetry.
(a) In the absence of an applied force, the centers of mass for positive and negative ions coincide.
(b) Under an applied force in the y direction, the centers of mass for positive and negative ions are shifted, which results in a net dipole moment, **P**, along *y*.
(c) When the force is along a different direction, along *x*, there may not be a resulting net dipole moment in that direction though there may be a net **P** along a different direction (*y*).

still has $\mathbf{P} = 0$. This result is generally true for all crystals that have a center of symmetry. The centers of mass of negative and positive charges in the unit cell remain coincident when the crystal is strained.

Piezoelectric crystals have no center of symmetry. For example, the hexagonal unit cell shown in Figure 7.40a exhibits no center of symmetry. If we draw a vector from point O to any charge and then reverse the vector, we will find an opposite charge. The unit cell is said to be **noncentrosymmetric.** When unstressed, as shown in Figure 7.40a, the center of mass of the negative charges coincides with the center of mass of the positive charges, both at O. However, when the unit cell is stressed, as shown in Figure 7.40b, the positive charge at A and the negative charge at B both become displaced inwards to A′ and B′, respectively. The two centers of mass therefore become shifted and there is now a net polarization **P**. Thus, an applied stress produces a net polarization **P** in the unit cell, and in this case **P** appears to be in the same direction as the applied stress, along *y*.

The direction of the induced polarization depends on the direction of the applied stress. When the same unit cell in Figure 7.40a is stressed along *x*, as illustrated in Figure 7.40c, there is no induced dipole moment along this direction because there is no net displacement of the centers of mass in the *x* direction. However, the stress causes the atoms A and B to be displaced outwards to A″ and B″, respectively, and this results in the shift of the centers of mass away from each other along *y*. In this case, an applied stress along *x* results in an induced polarization along *y*. Generally, an applied stress in one direction can give rise to induced polarization in other crystal directions. Suppose that T_j is the applied mechanical stress along some *j* direction and P_i is the induced polarization along some *i* direction; then the two are linearly related by

Piezoelectric effect

$$P_i = d_{ij} T_j \qquad [7.54]$$

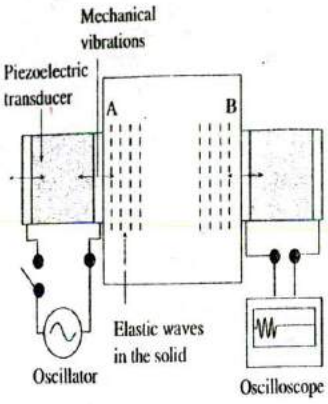


Figure 7.41 Piezoelectric transducers are widely used to generate ultrasonic waves in solids and also to detect such mechanical waves. The transducer on the left is excited from an ac source and vibrates mechanically. These vibrations are coupled to the solid and generate elastic waves. When the waves reach the other end, they mechanically vibrate the transducer on the right, which converts the vibrations to an electrical signal.

where d_{ij} are called the **piezoelectric coefficients.** Reversing the stress reverses the polarization. Although we did not specifically consider shear stresses in Figure 7.40, they, as well as tensile stresses, can also induce a net polarization, which means that T in Equation 7.54 can also represent shear stresses. The converse piezoelectric effect is that between an induced strain S_j along j and an applied electric field $\mathcal{E}_i$ along i,

Converse piezoelectric effect

$$S_j = d_{ij}\mathcal{E}_i \qquad [7.55]$$

The coefficients d_{ij} in Equations 7.54 and 7.55 are the same.[12]

As apparent from the foregoing discussions and Figure 7.38, piezoelectric crystals are essentially electromechanical transducers because they convert an electrical signal, an electric field, to a mechanical signal, strain, and vice versa. They are used in many engineering applications that involve electromechanical conversions, as in ultrasonic transducers, microphones, accelerometers, and so forth. Piezoelectric transducers are widely used to generate ultrasonic waves in solids and also to detect such mechanical waves, as illustrated in Figure 7.41. The transducer is simply a piezoelectric crystal, for example, quartz, that is appropriately cut and electroded to generate the desired types of mechanical vibrations (*e.g.*, longitudinal or transverse vibrations). The transducer on the left is attached to the surface A of the solid under examination, as shown in Figure 7.41. It is excited from an ac source, which means that it mechanically vibrates. These vibrations are coupled to the solid by a proper coupling medium (typically grease) and generate mechanical waves or elastic waves that propagate away from A. They are called **ultrasonic waves** as their frequencies are typically above the audible range. When the waves reach the other end, B, they mechanically vibrate the transducer attached to B, which converts the vibrations to an electrical signal that can readily be displayed on an oscilloscope. In this trivial example, one can easily measure the time it takes for elastic waves to travel in the solid from A to B and hence determine the ultrasonic velocity of the waves since the distance AB is

[12] The equivalence of the coefficients in Equations 7.54 and 7.55 can be shown by using thermodynamics and is not considered in this textbook. For rigorous piezoelectric definitions see IEEE Standard 176-1987 (*IEEE Trans. on Ultrasonics, Ferroelectrics and Frequency Control*, September 1996).

known. From the ultrasonic velocity one can determine the elastic constants (Young's modulus) of the solid. Furthermore, if there are internal imperfections such as cracks in the solid, then they reflect or scatter the ultrasonic waves. These reflections can lead to echoes that can be detected by suitably located transducers. Such ultrasonic testing methods are widely used for nondestructive evaluations of solids in mechanical engineering.

It is clear that an important engineering factor in the use of piezoelectric transducers is the electromechanical coupling between electrical and mechanical energies. The **electromechanical coupling factor** k is defined in terms of k^2 by

Electromechanical coupling factor

$$k^2 = \frac{\text{Electrical energy converted to mechanical energy}}{\text{Input of electrical energy}} \qquad [7.56a]$$

or equivalently by

Electromechanical coupling factor

$$k^2 = \frac{\text{Mechanical energy converted to electrical energy}}{\text{Input of mechanical energy}} \qquad [7.56b]$$

Table 7.8 summarizes some typical piezoelectric materials with some applications. The so-called PZT ceramics are widely used in many piezoelectric applications. PZT stands for lead zirconate titanate and the ceramic is a solid solution of lead zirconate, $PbZrO_3$, and lead titanate, $PbTiO_3$, so its composition is $PbTi_{1-x}Zr_xO_3$ where x is determined by the extent of the solid solution but typically is around 0.5. PZT piezoelectric components are manufactured by sintering, which is a characteristic ceramic manufacturing process in which PZT powders are placed in a mold and subjected to a pressure at high temperatures. During sintering the ceramic powders are fused through interdiffusion. The final properties depend not only on the composition of the solid solution but also on the manufacturing process, which controls the average grain size or polycrystallinity. Electrodes are deposited onto the final ceramic component, which is then poled by the application of a temporary electric field to induce it to become

Table 7.8 Piezoelectric materials and some typical values for d and k

Crystal	d (m V^{-1})	k	Comment
Quartz (crystal SiO_2)	2.3×10^{-12}	0.1	Crystal oscillators, ultrasonic transducers, delay lines, filters
Rochelle salt ($NaKC_4H_4O_6 \cdot 4H_2O$)	350×10^{-12}	0.78	
Barium titanate ($BaTiO_3$)	190×10^{-12}	0.49	Accelerometers
PZT, lead zirconate titanate ($PbTi_{1-x}Zr_xO_3$)	480×10^{-12}	0.72	Wide range of applications including earphones, microphones, spark generators (gas lighters, car ignition), displacement transducers, accelerometers
Polyvinylidene fluoride (PVDF)	18×10^{-12}	—	Must be poled; heated, put in an electric field and then cooled. Large area and inexpensive

piezoelectric. **Poling** refers to the application of a temporary electric field, generally at an elevated temperature, to align the polarizations of various grains and thereby develop piezoelectric behavior.

EXAMPLE 7.13

PIEZOELECTRIC SPARK GENERATOR The piezoelectric spark generator, as used in various applications such as lighters and car ignitions, operates by stressing a piezoelectric crystal to generate a high voltage which is discharged through a spark gap in air as schematically shown in Figure 7.42a. Consider a piezoelectric sample in the form of a cylinder as in Figure 7.42a. Suppose that the piezoelectric coefficient $d = 250 \times 10^{-12}$ m V^{-1} and $\varepsilon_r = 1000$. The piezoelectric cylinder has a length of 10 mm and a diameter of 3 mm. The spark gap is in air and has a breakdown voltage of about 3.5 kV. What is the force required to spark the gap? Is this a realistic force?

SOLUTION

We need to express the induced voltage in terms of the applied force. If the applied stress is T, then the induced polarization P is

$$P = dT = d\frac{F}{A}$$

Induced polarization P leads to induced surface polarization charges given by $Q = AP$. If C is the capacitance, then the induced voltage is

$$V = \frac{Q}{C} = \frac{AP}{\left(\frac{\varepsilon_o \varepsilon_r A}{L}\right)} = \frac{LP}{\varepsilon_o \varepsilon_r} = \frac{L\left(d\frac{F}{A}\right)}{\varepsilon_o \varepsilon_r} = \frac{dLF}{\varepsilon_o \varepsilon_r A}$$

Therefore, the required force is

$$F = \frac{\varepsilon_o \varepsilon_r AV}{dL} = \frac{(8.85 \times 10^{-12} \times 1000)\pi(1.5 \times 10^{-3})^2(3500)}{(250 \times 10^{-12})(10 \times 10^{-3})} = 87.6 \text{ N}$$

This force can be applied by squeezing by hand an appropriate lever arrangement; it is the weight of 9 kg. The force must be applied quickly because the piezoelectric charge generated

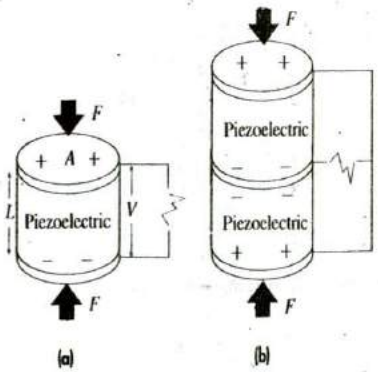


Figure 7.42 The piezoelectric spark generator.

will leak away (or become neutralized) if the charge is generated too slowly; many spark igniters use mechanical impact. The *energy* in the spark depends on the amount of charge generated. This can increase by using two piezoelectric crystals back to back as in Figure 7.42b, which is a more practical arrangement for a spark generator. The induced voltage per unit force V/F is proportional to $d/(\varepsilon_o\varepsilon_r)$ which is called the **piezoelectric voltage coefficient.** In general, if an applied stress $T = F/A$ induces a field $\mathcal{E} = V/L$ in a piezoelectric crystal, then the effect is related to the cause by the piezoelectric voltage coefficient g,

Piezoelectric voltage coefficient

$$\mathcal{E} = gT \qquad [7.57]$$

It is left as an exercise to show that $g = d/(\varepsilon_o\varepsilon_r)$.

7.8.2 PIEZOELECTRICITY: QUARTZ OSCILLATORS AND FILTERS

One of the most important applications of the piezoelectric quartz crystal in electronics is in the frequency control of oscillators and filters. Consider a suitably cut thin plate of a quartz crystal that has thin gold electrodes on the opposite faces. Suppose that we set up mechanical vibrations in the crystal by connecting the electrodes to an ac source, as in Figure 7.43a. It is possible to set up a mechanical resonance, or mechanical standing waves, in the crystal if the wavelength λ of the waves and the length ℓ along which the waves are traveling satisfy the condition for standing waves:

Mechanical standing waves

$$n\left(\frac{1}{2}\lambda\right) = \ell \qquad [7.58]$$

where n is an integer.

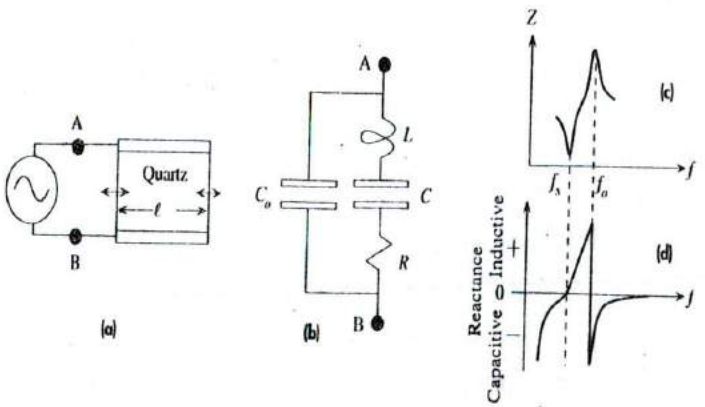


Figure 7.43 When a suitably cut quartz crystal with electrodes is excited by an ac voltage as in (a), it behaves as if it has the equivalent circuit in (b).

(c) and (d) The magnitude of the impedance Z and reactance (both between A and B) versus frequency, neglecting losses.

The frequency of these mechanical vibrations f_s is given by $f_s = v/\lambda$, where v is the velocity of the waves in the medium and λ is the wavelength. These mechanical vibrations in quartz experience very small losses and therefore have a high-quality factor Q, which means that resonance can only be set up if the frequency of the excitation, the electrical frequency, is close to f_s. Because of the coupling of energy between the electrical excitation and mechanical vibrations through the piezoelectric effect, mechanical vibrations appear like a series *LCR* circuit to the ac source, as shown in Figure 7.43b. This *LCR* series circuit has an impedance that is minimum at the **mechanical resonant frequency** f_s, given by

$$f_s = \frac{1}{2\pi\sqrt{LC}} \qquad [7.59]$$

Mechanical resonant frequency

In this series *LCR* circuit, L represents the mass of the transducer, C the stiffness, and R the losses or mechanical damping. Since the quartz crystal has electrodes at opposite faces, there is, in addition, the parallel plate capacitance C_o between the electrodes. Thus, the whole equivalent circuit is C_o in *parallel* with *LCR*, as in Figure 7.43b. As far as L is concerned, C_o and C are in series. There is a second higher resonant frequency f_a, called the **antiresonant frequency,** that is due to L resonating with C and C_o in series,

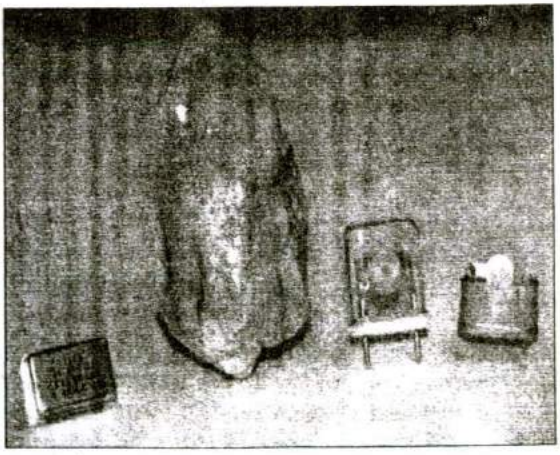
$$f_a = \frac{1}{2\pi\sqrt{LC'}} \qquad [7.60]$$

Antiresonant frequency

where

$$\frac{1}{C'} = \frac{1}{C_o} + \frac{1}{C}$$

The impedance between the terminals of the quartz crystal has the frequency dependence shown in Figure 7.43c. The two frequencies f_s and f_a are called the series and



Various quartz crystal "oscillators." Left to right: Raltron 40 MHz; a natural quartz crystal (South Dakota); Phillips 27 MHz; a cutaway view of a typical crystal oscillator.

parallel resonant frequencies, respectively. It is apparent that around f_a, the crystal behaves like a filter with a high Q value. If we were to examine the reactance of the crystal, whether it is behaving capacitively or inductively, we would find the behavior in Figure 7.43d, where positive reactance refers to an inductive and negative reactance to a capacitive behavior. Between f_s and f_a the crystal behaves inductively, and capacitively outside this range. Indeed, between f_s and f_a the response of the transducer is controlled by the mass of the crystal. This property has been utilized by electrical engineers in designing quartz oscillators.

In quartz oscillators, the crystal is invariably used in one of two modes. First, it can be used at f_s where it behaves as a resistance of R without any reactance. The circuit is designed so that oscillations can take place only when the crystal in the circuit exhibits no reactance or phase change—in other words, at f_s. Outside this frequency, the crystal introduces reactance or phase changes that do not lead to sustained oscillations. In a different mode of operation, the oscillator circuit is designed to make use of the **inductance** of the crystal just above f_s. Oscillations are maintained close to f_s because even very large changes in the inductance result in small changes in the frequency between f_s and f_a.

EXAMPLE 7.14 **THE QUARTZ CRYSTAL AND ITS EQUIVALENT CIRCUIT** From the following equivalent definition of the coupling coefficient,

$$k^2 = \frac{\text{Mechanical energy stored}}{\text{Total energy stored}}$$

show that

$$k^2 = 1 - \frac{f_s^2}{f_a^2}$$

Given that typically for an X-cut quartz crystal, $k = 0.1$, what is f_a for $f_s = 1$ MHz? What is your conclusion?

SOLUTION

C represents the mechanical mass where the mechanical energy is stored, whereas C_o is where the electrical energy is stored. If V is the applied voltage, then

$$k^2 = \frac{\text{Mechanical energy stored}}{\text{Total energy stored}} = \frac{\frac{1}{2}CV^2}{\frac{1}{2}CV^2 + \frac{1}{2}C_oV^2} = \frac{C}{C + C_o} = 1 - \frac{f_s^2}{f_a^2}$$

Rearranging this equation, we find

$$f_a = \frac{f_s}{\sqrt{1 - k^2}} \doteq \frac{1 \text{ MHz}}{\sqrt{1 - (0.1)^2}} = 1.005 \text{ MHz}$$

Thus, $f_a - f_s$ is only 5 kHz. The two frequencies f_s and f_a in Figure 7.43d are very close. An oscillator designed to oscillate at f_s, that is, at 1 MHz, therefore, cannot drift far (for example, a few kHz) because that would change the reactance enormously, which would upset the oscillation conditions.

EXAMPLE 7.15

QUARTZ CRYSTAL AND ITS INDUCTANCE A typical 1 MHz quartz crystal has the following properties:

$$f_s = 1 \text{ MHz} \qquad f_a = 1.0025 \text{ MHz} \qquad C_o = 5 \text{ pF} \qquad R = 20\ \Omega$$

What are C and L in the equivalent circuit of the crystal? What is the quality factor Q of the crystal, given that

$$Q = \frac{1}{2\pi f_s RC}$$

SOLUTION

The expression for f_s is

$$f_s = \frac{1}{2\pi\sqrt{LC}}$$

From the expression for f_a, we have

$$f_a = \frac{1}{2\pi\sqrt{LC'}} = \frac{1}{2\pi\sqrt{L\dfrac{CC_o}{C + C_o}}}$$

Dividing f_a by f_s eliminates L, and we get

$$\frac{f_a}{f_s} = \sqrt{\frac{C + C_o}{C_o}}$$

so that C is

$$C = C_o\left[\left(\frac{f_a}{f_s}\right)^2 - 1\right] = (5 \text{ pF})(1.0025^2 - 1) = 0.025 \text{ pF}$$

Thus

$$L = \frac{1}{C(2\pi f_s)^2} = \frac{1}{0.025 \times 10^{-12}(2\pi 10^6)^2} = 1.01 \text{ H}$$

This is a substantial inductance, and the enormous increase in the inductive reactance above f_s is intuitively apparent. The quality factor

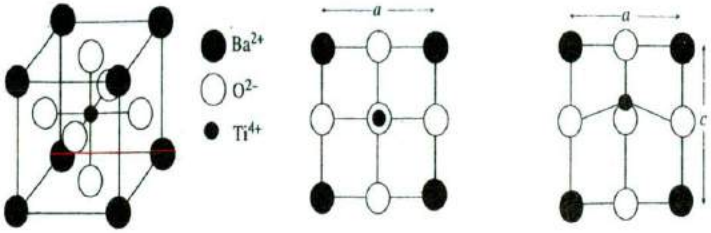
$$Q = \frac{1}{2\pi f_s RC} = 3.18 \times 10^5$$

is very large.

7.8.3 FERROELECTRIC AND PYROELECTRIC CRYSTALS

Certain crystals are permanently polarized even in the absence of an applied field. The crystal already possesses a finite polarization vector due to the separation of positive and negative charges in the crystal. These crystals are called **ferroelectric.**[13] Barium

[13] In analogy with the ferromagnetic crystals that already possess magnetization.



(a) $BaTiO_3$ cubic crystal structure above 130 °C

(b) $BaTiO_3$ cubic structure above 130 °C

(c) $BaTiO_3$ tetragonal structure below 130 °C

Figure 7.44 $BaTiO_3$ has different crystal structures above and below 130 °C that lead to different dielectric properties.

titanate ($BaTiO_3$) is probably the best cited example. Above approximately 130 °C, the crystal structure of $BaTiO_3$ has a cubic unit cell, as shown in Figure 7.44a. The centers of mass of the negative charges (O^{2-}) and the positive charges, Ba^{2+} and Ti^{4+}, coincide at the Ti^{4+} ion, as shown in Figure 7.44b. There is therefore no net polarization and $\mathbf{P} = 0$. Above 130 °C, therefore, the barium titanate crystal exhibits no permanent polarization and is not ferroelectric. However, below 130 °C, the structure of barium titanate is tetragonal, as shown in Figure 7.44c, in which the Ti^{4+} atom is not located at the center of mass of the negative charges. The crystal is therefore polarized by the separation of the centers of mass of the negative and positive charges. The crystal possesses a finite polarization vector **P** and is ferroelectric. The critical temperature above which ferroelectric property is lost, in this case 130 °C, is called the **Curie temperature** (T_C). Below the Curie temperature, the whole crystal becomes spontaneously polarized. The onset of spontaneous polarization is accompanied by the distortion of the crystal structure, as shown by the change from Figure 7.44b to Figure 7.44c. The spontaneous displacement of the Ti^{4+} ion below the Curie temperature elongates the cubic structure, which becomes tetragonal. It is important to emphasize that we have only described an observation and not the reasons for the spontaneous polarization of the whole crystal. The development of the permanent dipole moment below the Curie temperature involves long-range interactions between the ions outside the simple unit cell pictured in Figure 7.44. The energy of the crystal is lower when the Ti^{4+} ion in each unit cell is slightly displaced along the c direction, as in Figure 7.44c, which generates a dipole moment in each unit cell. The interaction energy of these dipoles when all are aligned in the same direction lowers the energy of the whole crystal. It should be mentioned that the distortion of the crystal that takes place when spontaneous polarization occurs just below T_C is very small relative to the dimensions of the unit cell. For $BaTiO_3$, for example, c/a is 1.01 and the displacement of the Ti^{4+} ion from the center is only 0.012 nm, compared with $a = 0.4$ nm.

An important and technologically useful characteristic of a ferroelectric crystal is its ability to be poled. Above 130 °C there is no permanent polarization in the crystal. If we apply a temporary field $\mathcal{E}$ and let the crystal cool to below 130 °C, we can induce the spontaneous polarization **P** to develop along the field direction. In other words, we would define the c axis by imposing a temporary external field. This process is called poling. The c axis is the polar axis along which **P** develops. It is also called the

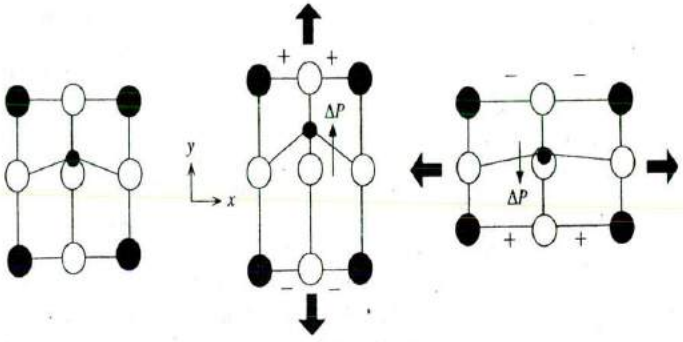


Figure 7.45 Piezoelectric properties of $BaTiO_3$ below its Curie temperature.

ferroelectric axis. Since below the Curie temperature the ferroelectric crystal already has a permanent polarization, it is not possible to use the expression

$$P = \varepsilon_o(\varepsilon_r - 1)\mathcal{E}$$

to define a relative permittivity. Suppose that we use a ferroelectric crystal as a dielectric medium between two parallel plates. Since any change ΔP normal to the plates changes the stored charge, what is of significance to the observer is the change in the polarization. We can appreciate this by noting that $C = Q/V$ is not a good definition of capacitance if there are already charges on the plates, even in the absence of voltage.[14] We then prefer a definition of C based on $\Delta Q/\Delta V$ where ΔQ is the change in stored charge due to a change ΔV in the voltage. Similarly, we define the relative permittivity ε_r in this case in terms of the change ΔP in P induced by $\Delta\mathcal{E}$ in the field $\mathcal{E}$,

$$\Delta P = \varepsilon_o(\varepsilon_r - 1)\,\Delta\mathcal{E}$$

An applied field along the a axis can displace the Ti^{4+} ion more easily than that along the c axis, and experiments show that $\varepsilon_r \approx 4100$ along a is much greater than $\varepsilon_r \approx 160$ along c. Because of their large dielectric constants, ferroelectric ceramics are used as high-K dielectrics in capacitors.

All ferroelectric crystals are also piezoelectric, but the reverse is not true: not all piezoelectric crystals are ferroelectric. When a stress along y is applied to the $BaTiO_3$ crystal in Figure 7.45a, the crystal is stretched along y, as a result of which the Ti^{4+} atom becomes displaced, as shown in Figure 7.45b. There is, however, no shift in the center of mass of the negative charges, which means that there is a change ΔP in the polarization vector along y. Thus, the applied stress induces a change in the polarization, which is a piezoelectric effect. If the stress is along x, as illustrated in Figure 7.45c, then the change in the polarization is along y. In both cases, ΔP is proportional to the stress, which is a characteristic of the piezoelectric effect.

[14] A finite Q on the plates of a capacitor when $V = 0$ implies an infinite capacitance, $C = \infty$. However, $C = dQ/dV$ definition avoids this infinity.

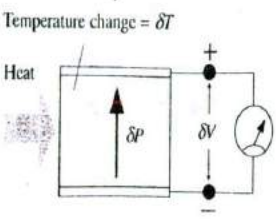


Figure 7.46 The heat absorbed by the crystal increases the temperature by δT, which induces a change δP in the polarization.
This is the pyroelectric effect. The change δP gives rise to a change δV in the voltage that can be measured.

The barium titanate crystal in Figure 7.44 is also said to be pyroelectric because when the temperature increases, the crystal expands and the relative distances of ions change. The Ti^{4+} ion becomes shifted, which results in a change in the polarization. Thus, a temperature change δT induces a change δP in the polarization of the crystal. This is called **pyroelectricity,** which is illustrated in Figure 7.46. The magnitude of this effect is quantized by the **pyroelectric coefficient** p, which is defined by

Pyroelectric coefficient

$$p = \frac{dP}{dT} \qquad [7.61]$$

A few typical pyroelectric crystals and their pyroelectric coefficients are listed in Table 7.9. Very small temperature changes, even in thousandths of degrees, in the material can develop voltages that can be readily measured. For example, for a PZT-type pyroelectric ceramic in Table 7.9, taking $\delta T = 10^{-3}$ K and $p \approx 380 \times 10^{-6}$, we find $\delta P = 3.8 \times 10^{-7}$ C m^{-2}. From

$$\delta P = \varepsilon_0(\varepsilon_r - 1)\,\delta\mathcal{E}$$

with $\varepsilon_r = 290$, we find

$$\delta\mathcal{E} = 148 \text{ V m}^{-1}$$

If the distance between the faces of the ceramic where the charges are developed is 0.1 mm, then

$$\delta V = 0.0148 \text{ V} \quad \text{or} \quad 15 \text{ mV}$$

Table 7.9 Some pyroelectric (and also ferroelectric) crystals and typical properties

Material	ε_r'	tan δ	Pyroelectric Coefficient ($\times 10^{-6}$ C m^{-2} K^{-1})	Curie Temperature (°C)
$BaTiO_3$	4100 ⊥ polar axis; 160 // polar axis	7×10^{-3}	20	130
$LiTaO_3$	47	5×10^{-3}	230	610
PZT modified for pyroelectric	290	2.7×10^{-3}	380	230
PVDF, polymer	12	0.01	27	80

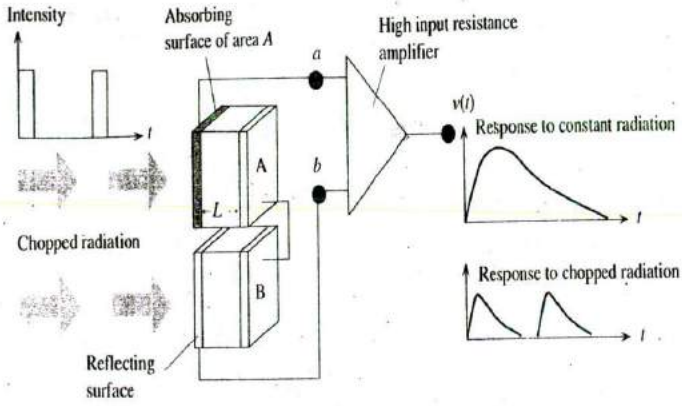


Figure 7.47 The pyroelectric detector.

Radiation is absorbed in the detecting element, A, which generates a pyroelectric voltage that is measured by the amplifier. The second element, B, has a reflecting electrode and does not absorb the radiation. It is a dummy element that compensates for the piezoelectric effects. Piezoelectric effects generate equal voltages in both A and B, which cancel each other across *a* and *b*, the input of the amplifier.

which can be readily measured. Pyroelectric crystals are widely used as infrared detectors. Any infrared radiation that can raise the temperature of the crystal even by a thousandth of a degree can be detected. For example, many intruder alarms use pyroelectric detectors because as the human or animal intruder passes by the view of detector, the infrared radiation from the warm body raises the temperature of the pyroelectric detector, which generates a voltage that actuates an alarm.

Figure 7.47 shows a simplified schematic circuit for a pyroelectric radiation detector. The detecting element, labeled A, is actually a thin crystal or ceramic (or even a polymer) of a pyroelectric material that has electrodes on opposite faces. Pyroelectric materials are also piezoelectric and therefore also sensitive to stresses. Thus, pressure fluctuations, for example, vibrations from the detector mount or sound waves, interfere with the response of the detector to radiation alone. These can be compensated for by having a second dummy detector B that has a reflecting coating and is subjected to the same vibrations (air and mount), as depicted in Figure 7.47. Thus, there are two elements in the detector, one with an absorbing surface, detecting element A, and the other with a reflecting surface, compensating element B. Stress fluctuations give rise to the same piezoelectric voltage in both, which then cancel each other between *a* and *b* at the input of the amplifier. When radiation is incident, then only the detecting element absorbs the radiation, becomes warmer, and hence generates a pyroelectric voltage. This voltage appears directly across *a* and *b*. As the incident radiation warms the detecting element and increases its temperature, the pyroelectric voltage increases with time. Eventually the temperature reaches a steady-state value determined by heat losses from the element. We therefore expect the pyroelectric voltage to reach a constant value as well. However, the problem is that a constant pyroelectric voltage cannot be sustained because the surface charges slowly become neutralized or leak away.

The constant radiation is therefore normally chopped to subject the detector to periodic bursts of radiation, as shown in Figure 7.47. The pyroelectric voltage is then a changing function of time, which is readily measured and related to the power in the incident radiation.

Many pyroelectric applications refer to a pyroelectric current that is generated by the temperature rise. There is another way to look at the pyroelectric phenomenon instead of considering the induced pyroelectric voltage that is created across the crystal (Figure 7.46). The induced polarization δP in a small time interval δt, due to the change δT in the temperature, generates an induced polarization charge density δP on the crystal's surfaces. This charge density δP flows in a time interval δt, and hence generates an *induced polarization current density* J_p to flow, *i.e.*,

Pyroelectric current density

$$J_p = \frac{dP}{dt} = p\frac{dT}{dt} \qquad [7.62]$$

J_p in Equation 7.62 is called the **pyroelectric current density** and depends on the rate of change of the temperature dT/dt brought about by the absorption of radiation.

Most pyroelectric detectors are characterized by their **current responsivity** $\mathcal{R}_J$ defined as the pyroelectric current generated per unit input radiation power,

Pyroelectric current responsivity

$$\mathcal{R}_J = \frac{\text{Pyroelectric current generated}}{\text{Input radiation power}} = \frac{J_p}{I} \qquad [7.63]$$

where I is the radiation intensity (W m^{-2}); $\mathcal{R}_J$ is quoted in A W^{-1}. If the pyroelectric current generated by the crystal flows into the self-capacitance of the crystal itself (no external resistors or capacitors connected, and the voltmeter is an ideal meter), it charges the self-capacitance to generate the observed voltage δV in Figure 7.46. The **pyroelectric voltage responsivity** $\mathcal{R}_V$ is defined similarly to Equation 7.63 but considers the voltage that is developed upon receiving the input radiation:

Pyroelectric voltage responsivity

$$\mathcal{R}_V = \frac{\text{Pyroelectric output voltage generated}}{\text{Input radiation power}} \qquad [7.64]$$

The output voltage that is generated depends not only on the pyroelectric crystal's dielectric properties, but also on the input impedance of the amplifier, and can be quite complicated. A typical commercial $LiTaO_3$ pyroelectric detector has a current responsivity of 0.1–1 μA/W.

EXAMPLE 7.16 **A PYROELECTRIC RADIATION DETECTOR** Consider the radiation detector in Figure 7.47 but with a single element A. Suppose that the radiation is chopped so that the radiation is passed to the detector for a time Δt seconds every τ seconds, where $\Delta t \ll \tau$. If Δt is sufficiently small, then the temperature rise ΔT is small and hence the heat losses are negligible during Δt. Using the heat capacity to find the temperature change during Δt, relate the magnitude of the voltage ΔV to the incident radiation intensity I. What is your conclusion?

Consider a PZT-type pyroelectric material with a density of about 7 g cm^{-3} and a specific heat capacity of about 380 J K^{-1} kg^{-1}. If $\Delta t = 0.2$ s and the minimum voltage that can be detected above the background noise is 1 mV, what is the minimum radiation intensity that can be measured?

SOLUTION

Suppose that the radiation of intensity I is received during a time interval Δt and delivers an amount of energy ΔH to the pyroelectric detector. This energy ΔH, in the absence of any heat losses, increases the temperature by ΔT. If c is the specific heat capacity (heat capacity per unit mass) and ρ is the density,

$$\Delta H = (AL\rho)c\,\Delta T$$

where A is the surface area and L the thickness of the detector. The change in the polarization ΔP is

$$\Delta P = p\,\Delta T = \frac{p\,\Delta H}{AL\rho c}$$

The change in the surface charge ΔQ is

$$\Delta Q = A\,\Delta P = \frac{p\,\Delta H}{L\rho c}$$

This change in the surface charge gives a voltage change ΔV across the electrodes of the detector. If $C = \varepsilon_o\varepsilon_r A/L$ is the capacitance of the pyroelectric crystal,

$$\Delta V = \frac{\Delta Q}{C} = \frac{p\,\Delta H}{L\rho c} \times \frac{L}{\varepsilon_o\varepsilon_r A} = \frac{p\,\Delta H}{A\rho c\varepsilon_o\varepsilon_r}$$

The absorbed energy (heat) ΔH during Δt depends on the intensity of incident radiation. Incident intensity I is the energy arriving per unit area per unit time. In time Δt, I delivers an energy $\Delta H = IA\,\Delta t$. Substituting for ΔH in the expression for ΔV, we find

Pyroelectric detector output voltage

$$\Delta V = \frac{pI\,\Delta t}{\rho c\varepsilon_r\varepsilon_o} = \left(\frac{p}{\rho c\varepsilon_r\varepsilon_o}\right) I\,\Delta t \qquad [7.65]$$

The parameters in the parentheses are material properties and reflect the "goodness" of the pyroelectric material for the application. We should emphasize that in deriving Equation 7.65 we did not consider any heat losses that will prevent the rise of the temperature indefinitely. If Δt is short, then the temperature change will be small and heat losses negligible.

For a PZT-type pyroelectric, we can take $p = 380 \times 10^{-6}\ \mathrm{C\,m^{-2}\,K^{-1}}$, $\varepsilon_r = 290$, $c = 380\ \mathrm{J\,K^{-1}\,kg^{-1}}$, and $\rho = 7 \times 10^3\ \mathrm{kg\,m^{-3}}$, and then from Equation 7.65 with $\Delta V = 0.001$ V and $\Delta t = 0.2$ s, we have

$$I = \left(\frac{p}{\rho c\varepsilon_o\varepsilon_r}\right)^{-1} \frac{\Delta V}{\Delta t} = \left(\frac{380 \times 10^{-6}}{(7000)(380)(290)(8.85 \times 10^{-12})}\right)^{-1} \frac{0.001}{0.2}$$

$$= 0.090\ \mathrm{W\,m^{-2}} \quad \text{or} \quad 9\,\mu\mathrm{W\,cm^{-2}}$$

We have assumed that all the incident radiation I is absorbed by the pyroelectric crystal. In practice, only a fraction η (called the *emissivity* of the surface), that is, ηI, will be absorbed instead of I. We also assumed that the output voltage ΔV is developed totally across the pyroelectric element capacitance; that is, the amplifier's input impedance (parallel combination of its input capacitance and resistance) is negligible compared with that of the pyroelectric crystal. As stated, we also neglected all heat losses from the pyroelectric crystal so that the absorbed radiation simply increases the crystal's temperature. These simplifying assumptions lead to the maximum signal ΔV that can be generated from a given input radiation signal I as stated in Equation 7.65. It is left as an exercise to show that Equation 7.65 can also be easily derived by starting from Equation 7.62 for the pyroelectric current density J_p, and have J_p charge up the capacitance $C = \varepsilon_o\varepsilon_r A/L$ of the crystal.

ADDITIONAL TOPICS

7.9 ELECTRIC DISPLACEMENT AND DEPOLARIZATION FIELD

Electric Displacement (D) and Free Charges Consider a parallel plate capacitor with free space between the plates, as shown in Figure 7.48a, which has been charged to a voltage V_o by connecting it to a battery of voltage V_o. The battery has been suddenly removed, which has left the free positive and negative charges Q_{free} on the plates. These charges are free in the sense that they can be conducted away. An ideal electrometer (with no leakage current) measures the total charge on the positive plate (or voltage of the positive plate with respect to the negative plate). The voltage across the plates is V_o and the capacitance is C_o. The field in the free space between the plates is

Electric field without dielectric

$$\mathcal{E}_o = \frac{Q_{free}}{\varepsilon_o A} = \frac{V_o}{d} \qquad [7.66]$$

where d is the separation of the plates.

When we insert a dielectric to fit between the plates, the field polarizes the dielectric and polarization charges $-Q_P$ and $+Q_P$ appear on the left and right surfaces of the dielectric, as shown in Figure 7.48b. As there is no battery to supply more free charges, the net charge on the left plate (positive plate) becomes $Q_{free} - Q_P$. Similarly the net negative charge on the right plate becomes $-Q_{free} + Q_P$. The field inside the dielectric is no longer $\mathcal{E}_o$ but less because induced polarization charges have the opposite polarity to the original free charges and the net charge on each plate has been reduced. The new field can be found by applying Gauss's law. Consider a Gauss surface just enclosing the left plate and the surface region of the dielectric with its negative polarization charges, as shown in Figure 7.49. Then Gauss's law gives

Gauss's law with dielectric

$$\oint_{\text{Surface}} \varepsilon_o \mathcal{E}\, dA = Q_{\text{total}} = Q_{free} - Q_P \qquad [7.67]$$

where A is the plate area (same as dielectric surface area) and we take the field $\mathcal{E}$ to be normal to the surface area dA, as indicated in Figure 7.49. If the polarization charge is

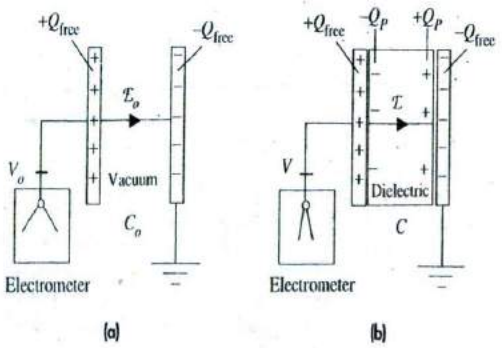


Figure 7.48
(a) Parallel plate capacitor with free space between plates that has been charged to a voltage V_o. There is no battery to maintain the voltage constant across the capacitor. The electrometer measures the voltage difference across the plates and, in principle, does not affect the measurement. (b) After the insertion of the dielectric, the voltage difference is V, less than V_o, and the field in the dielectric is $\mathcal{E}$ less than $\mathcal{E}_o$.

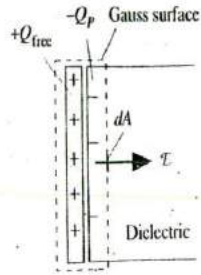


Figure 7.49 A Gauss surface just around the left plate and within the dielectric, encompassing both $+Q_{free}$ and $-Q_P$.

dQ_P over a small surface area dA of the dielectric, then the polarization charge density σ_P at this point is defined as

$$\sigma_P = \frac{dQ_P}{dA}$$

For uniform polarization, the charge distribution is Q_P/A, as we have used previously. Since $\sigma_P = P$, where P is the polarization vector, we can write

$$P = \frac{dQ_P}{dA}$$

and therefore express Q_P as

$$Q_P = \oint_{\text{Surface}} P\, dA \qquad [7.68]$$

We can now substitute for Q_P in Equation 7.67 and take this term to the left-hand side to add the two surface integrals. The right-hand side is left with only Q_{free}. Thus,

$$\oint_{\text{Surface}} (\varepsilon_o \mathcal{E} + P)\, dA = Q_{free} \qquad [7.69]$$

What is important here is that the surface integration of the quantity $\varepsilon_o \mathcal{E} + P$ is always equal to the total free charges on the surface. Whatever the dielectric material, this integral is always Q_{free}. It becomes convenient to define $\varepsilon_o \mathcal{E} + P$ as a usable quantity, called the **electric displacement** and denoted as D, that is,

Definition of electric displacement

$$D = \varepsilon_o \mathcal{E} + P \qquad [7.70]$$

Then, Gauss's law in terms of free charges alone in Equation 7.69 becomes

Gauss's law for free charges

$$\oint_{\text{Surface}} D\, dA = Q_{free} \qquad [7.71]$$

In Equation 7.71 we take D to be normal to the surface area dA as in the case of $\mathcal{E}$ in Gauss's law. Equation 7.71 provides a convenient way to calculate the electric displacement D, from which one should be able to determine the field. We should note that, in general, $\mathcal{E}$ is a vector and so is P, so the definition in Equation 7.70 is

strictly in terms of vectors. Inasmuch as the electric displacement depends only on free charges, as a vector it starts at negative free charges and finishes on positive free charges.

Equation 7.71 for D defines it in terms of $\mathcal{E}$ and P, but we can express D in terms of the field $\mathcal{E}$ in the dielectric alone. The polarization P and $\mathcal{E}$ are related by the definition of the relative permittivity ε_r,

$$P = \varepsilon_o(\varepsilon_r - 1)\mathcal{E}$$

Substituting for P in Equation 7.70 and rearranging, we find that D is simply given by

Electric displacement and the field

$$D = \varepsilon_o\varepsilon_r\mathcal{E} \qquad [7.72]$$

We should note that this simple equation applies in an isotropic medium where the field along one direction, for example, x, does not generate polarization along a different direction, for example, y. In those cases, Equation 7.72 takes a tensor form whose mathematics is beyond the scope of this book.

We can now apply Equation 7.71 for a Gauss surface surrounding the left plate,

$$D = \frac{Q_{\text{free}}}{A} = \varepsilon_o\mathcal{E}_o \qquad [7.73]$$

where we used Equation 7.66 to replace Q_{free}. Thus D does not change when we insert the dielectric because the same free charges are still on the plates (they cannot be conducted away anywhere). The new field $\mathcal{E}$ between the plates after the insertion of the dielectric is

$$\mathcal{E} = \frac{1}{\varepsilon_o\varepsilon_r}D = \frac{1}{\varepsilon_r}\mathcal{E}_o \qquad [7.74]$$

The original field is reduced by the polarization of the dielectric. We should recall that the field does *not* change in the case where the parallel plate capacitor is connected to a battery that keeps the voltage constant across the plates and supplies additional free charges (ΔQ_{free}) to make up for the induced opposite-polarity polarization charges.

Gauss's law in Equation 7.71 in terms of D and the enclosed free charges Q_{free} can also be written in terms of the field $\mathcal{E}$, but including the relative permittivity, because D and $\mathcal{E}$ are related by Equation 7.72. Using Equation 7.72, Equation 7.71 becomes

Gauss's law for free charges

$$\oint_{\text{Surface}} \varepsilon_o\varepsilon_r\mathcal{E}\,dA = Q_{\text{free}}$$

For an isotropic medium where ε_r is the same everywhere,

Gauss's law in an isotropic dielectric

$$\oint_{\text{Surface}} \mathcal{E}\,dA = \frac{Q_{\text{free}}}{\varepsilon_o\varepsilon_r} \qquad [7.75]$$

As before, $\mathcal{E}$ in the surface integral is taken as normal to dA everywhere. Equation 7.75 is a convenient way of evaluating the field from the free charges alone, given the dielectric constant of the medium.

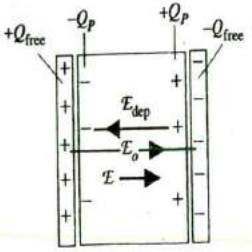


Figure 7.50 The field inside the dielectric can be considered to be the sum of the field due to the free charges (Q_{free}) and a field due to the polarization of the dielectric, called the depolarization field.

The Depolarizing Field We can view the field $\mathcal{E}$ as arising from two electric fields: that due to the free charges $\mathcal{E}_o$ and that due to the polarization charges, denoted as $\mathcal{E}_{dep}$. These two fields are indicated in Figure 7.50. $\mathcal{E}_o$ is called the **applied field** as it is due to the free charges that have been put on the plates. It starts and ends at free charges on the plates. The field due to polarization charges starts and ends at these bound charges and is in the *opposite* direction to the $\mathcal{E}_o$. Although $\mathcal{E}_o$ polarizes the molecules of the medium, $\mathcal{E}_{dep}$, being in the opposite direction, tries to depolarize the medium. It is called the **depolarizing field** (and hence the subscript). Thus the field inside the medium is

$$\mathcal{E} = \mathcal{E}_o - \mathcal{E}_{dep} \qquad [7.76]$$

The depolarizing field depends on the amount of polarization since it is determined by $+Q_P$ and $-Q_P$. For the dielectric plate in Figure 7.50, we know the field $\mathcal{E}$ is $\mathcal{E}_o/\varepsilon_r$, so we can eliminate $\mathcal{E}_o$ in Equation 7.76 and relate $\mathcal{E}_{dep}$ directly to $\mathcal{E}$,

$$\mathcal{E}_{dep} = \mathcal{E}(\varepsilon_r - 1)$$

However, the polarization P is related to the field $\mathcal{E}$ by

$$P = \varepsilon_o(\varepsilon_r - 1)\mathcal{E}$$

which means that the depolarization field is

$$\mathcal{E}_{dep} = \frac{1}{\varepsilon_o}P \qquad [7.77]$$

Depolarizing field in a dielectric plate

As we expected, the depolarizing field is proportional to the polarization P. We should emphasize that $\mathcal{E}_{dep}$ is in the *opposite direction* to $\mathcal{E}$ and P and Equation 7.77 is for magnitudes only. If we write it as a vector equation, then we must introduce a negative sign to give $\mathcal{E}_{dep}$ a direction opposite to that of P. Moreover, the relationship in Equation 7.77 is special to the dielectric plate geometry in Figure 7.50. In general, the depolarizing field is still proportional to the polarization, as in Equation 7.77, but it is given by

$$\mathcal{E}_{dep} = \frac{N_{dep}}{\varepsilon_o}P \qquad [7.78]$$

Depolarizing field in a dielectric

where N_{dep} is a numerical factor called the **depolarization factor.** It takes into account the shape of the dielectric and the variation in the polarization within the medium. For

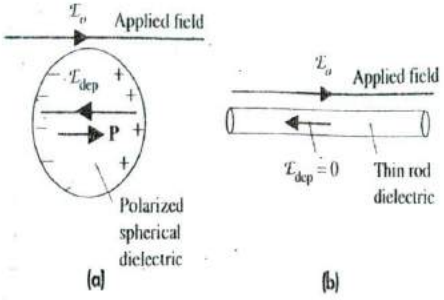


Figure 7.51
(a) Polarization and the depolarizing field in a spherical-shaped dielectric placed in an applied field.
(b) Depolarization field in a thin rod placed in an applied field is nearly zero.

a dielectric plate placed perpendicularly to an external field, $N_{dep} = 1$, as we found in Equation 7.77. For the spherical dielectric medium as in Figure 7.51a, $N_{dep} = \frac{1}{3}$. For a long thin dielectric rod placed with its axis along the applied field, as in Figure 7.51b, $N_{dep} \approx 0$ and becomes exactly zero as the diameter shrinks to zero. N_{dep} is always between 0 and 1. If we know N_{dep}, we can determine the field inside the dielectric, for example, in a small spherical cavity within an insulation given the external field.

7.10 LOCAL FIELD AND THE LORENTZ EQUATION

When a dielectric medium is placed in an electric field, it becomes polarized and there is a macroscopic, or an average, field $\mathcal{E}$ in the medium. The actual field at an atom, called the **local field** $\mathcal{E}_{loc}$, however, is not the same as the average field as illustrated in Figure 7.7.

Consider a dielectric plate polarized by placing it between the plates of a capacitor as shown in Figure 7.52a. The macroscopic field $\mathcal{E}$ in the dielectric is given by the applied field $\mathcal{E}_o$ due to the free charges Q_{free} on the plates, and the depolarization field due to P, or polarization charges on the dielectric plate surfaces A. Since we have a plate

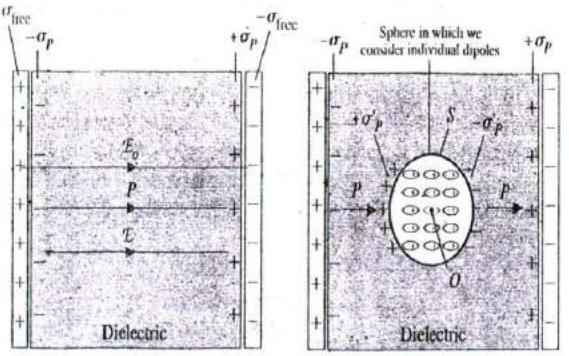


Figure 7.52
(a) The macroscopic field $\mathcal{E}$ is determined by the applied field $\mathcal{E}_o$ and the depolarization field due to P.
(b) Calculation of the local field involves making a hypothetical spherical cavity S inside the dielectric. This produces polarization surface charges on the inside surface S of the cavity. The effects of the dipoles inside the cavity are treated individually.

dielectric, the depolarization field is P/ε_o, so

$$\mathcal{E} = \mathcal{E}_o - \mathcal{E}_{\text{dep}} = \mathcal{E}_o - \frac{1}{\varepsilon_o} P$$

Consider the field at some atomic site, point O, but with the atom itself removed. We evaluate the field at O coming from all the charges except the atom at O itself since we are looking at the field experienced by this atom (the atom cannot become polarized by its own field). We then cut a (hypothetical) spherical cavity S centered at O and consider the atomic polarizations individually within the spherical cavity. In other words, the effects of the dipoles in the cavity are treated separately from the remaining dielectric medium which is now left with a spherical cavity. This remaining dielectric is considered as a continuous medium but with a spherical cavity. Its dielectric property is represented by its polarization vector P. Because of the cavity, we must now put polarization charges on the inner surface S of this cavity as illustrated in Figure 7.52b. This may seem surprising, but we should remember that we are treating the effects of the atomic dipoles within the cavity individually and separately by cutting out a spherical cavity from the medium and thereby introducing a surface S.

The field at O comes from four sources:

1. Free charges Q_{free} on the electrodes, represented by $\mathcal{E}_o$.
2. Polarization charges on the plate surfaces A, represented by $\mathcal{E}_{\text{dep}}$.
3. Polarization charges on the inner surface of the spherical cavity S, represented by $\mathcal{E}_S$.
4. Individual dipoles within the cavity, represented by $\mathcal{E}_{\text{dipoles}}$.

Thus,

$$\mathcal{E}_{\text{loc}} = \mathcal{E}_o + \mathcal{E}_{\text{dep}} + \mathcal{E}_S + \mathcal{E}_{\text{dipoles}}$$

Since the first two terms make up the macroscopic field, we can write this as

Local field in a crystal

$$\mathcal{E}_{\text{loc}} = \mathcal{E} + \mathcal{E}_S + \mathcal{E}_{\text{dipoles}}$$

The field from the individual dipoles surrounding O depends on the positions of these atomic dipoles which depend on the crystal structure. For cubic crystals, amorphous solids (*e.g.*, glasses), or liquids, effects of these dipoles around O cancel each other and $\mathcal{E}_{\text{dipoles}} = 0$. Thus,

Local field in a cubic crystal or a non-crystalline material

$$\mathcal{E}_{\text{loc}} = \mathcal{E} + \mathcal{E}_S \qquad [7.79]$$

We are then left with evaluating the field due to polarization charges on the inner surface S of the cavity. This field comes from polarization charges on the surface S. Consider a thin spherical shell on surface S as shown in Figure 7.53 which makes an angle θ with O. The radius of this shell is $a \sin\theta$, whereas its width (or thickness) is $a\,d\theta$. The surface area dS is then $(2\pi a \sin\theta)(a\,d\theta)$. The polarization charge dQ_P on this spherical shell surface is $P_n\,dS$ where P_n is the polarization vector normal to the surface dS. Thus,

$$dQ_P = P_n\,dS = (P\cos\theta)(2\pi a \sin\theta)(a\,d\theta)$$

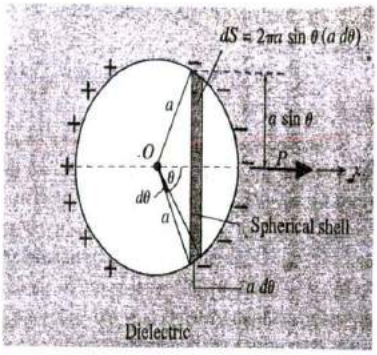


Figure 7.53 Calculation of the field due to polarization charges on the inner surface S of the spherical cavity. Consider a spherical shell of radius a. The surface area is $dS = 2\pi a \sin\theta\,(a\,d\theta)$.

But the field at O from dQ_P is given from electrostatics as

$$d\mathcal{E}_S = \frac{dQ_P}{4\pi\varepsilon_o a^2} = \frac{(P\cos\theta)(2\pi a\,\sin\theta)(a\,d\theta)}{4\pi\varepsilon_o a^2}$$

To find the total field coming from the whole surface S we have to integrate $d\mathcal{E}_S$ from $\theta = 0$ to $\theta = \pi$,

$$\mathcal{E}_S = \int_0^{\pi} \frac{(P\cos\theta)(\sin\theta)}{2\varepsilon_o}\,d\theta$$

which integrates to

Field in a spherical cavity

$$\mathcal{E}_S = \frac{1}{3\varepsilon_o}P \qquad [7.80]$$

Local field in a cubic crystal or noncrystalline material

The local field by Equation 7.79 is

$$\mathcal{E}_{\text{loc}} = \mathcal{E} + \frac{1}{3\varepsilon_o}P \qquad [7.81]$$

Equation 7.81 is the **Lorentz relation** for the local field in terms of the polarization P of the medium and is valid for cubic crystals and noncrystalline materials, such as glasses. It does *not* apply to dipolar dielectrics in which the local field can be quite complicated.

7.11 DIPOLAR POLARIZATION

Consider a medium with molecules that have permanent dipole moments. Each permanent dipole moment is p_o. In the presence of an electric field the dipoles try to align perfectly with the field, but random thermal collisions, *i.e.*, thermal agitation, act against this perfect alignment. A molecule that manages to rotate and align with the field finds itself later colliding with another molecule and losing its alignment. We are interested in the mean dipole moment in the presence of an applied field taking into

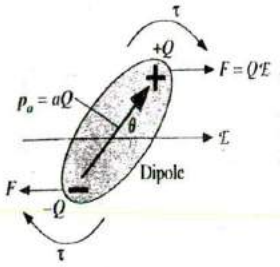


Figure 7.54 In the presence of an applied field a dipole tries to rotate to align with the field against thermal agitation.

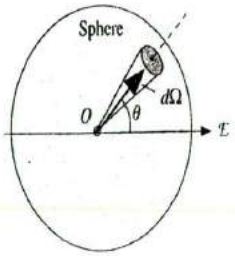


Figure 7.55 The dipole is pointing within a solid angle $d\Omega$.

account the thermal energies of the molecules and their random collisions. We will assume that the probability that a molecule has an energy E is given by the Boltzmann factor, $\exp(-E/kT)$.

Consider an arbitrary dipolar molecule in an electric field as in Figure 7.54 with its dipole moment p_o at an angle θ with the field $\mathcal{E}$. The torque experienced by the dipole is given by $\tau = (F \sin\theta)a$ or $\mathcal{E}p_o \sin\theta$ where $p_o = aQ$. The potential energy E at an angle θ is given by integrating $\tau d\theta$,

$$E = \int_0^{\theta} p_o \mathcal{E} \sin\theta \, d\theta = -p_o \mathcal{E} \cos\theta + p_o \mathcal{E}$$

Potential energy of a dipole at an angle θ

Inasmuch as the PE depends on the orientation, there is a certain probability of finding a dipole oriented at this angle as determined by the Boltzmann distribution. The fraction f of molecules oriented at θ is proportional to $\exp(-E/kT)$,

$$f \propto \exp\left(\frac{p_o \mathcal{E} \cos\theta}{kT}\right) \qquad [7.82]$$

Boltzmann distribution

The initial orientation of the dipole should be considered in three dimensions and not as in the two-dimensional illustration in Figure 7.54. In three dimensions we use solid angles, and the fraction f then represents the fraction of molecules pointing in a direction defined by a small solid angle $d\Omega$ as shown in Figure 7.55. The whole sphere around the dipole corresponds to a solid angle of 4π. Furthermore, we need to find the average dipole moment along $\mathcal{E}$ as this will be the induced net dipole moment by the field. The dipole moment along $\mathcal{E}$ is $p_o \cos\theta$. Then from the definition of the average

$$p_{av} = \frac{\int_0^{4\pi} (p_o \cos\theta) f \, d\Omega}{\int_0^{4\pi} f \, d\Omega} \qquad [7.83]$$

where f is the Boltzmann factor given in Equation 7.82 and depends on $\mathcal{E}$ and θ. The final result of this integration is a special function called the **Langevin function** which

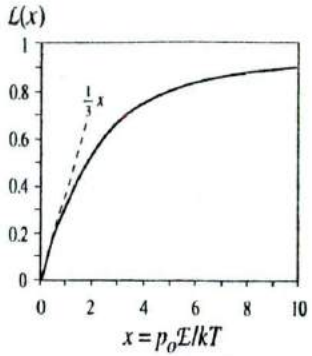


Figure 7.56 The Langevin function.

Average dipole moment and the Langevin function

is denoted as $\mathcal{L}(x)$ where x is the argument of the function (not the x coordinate). The integration of Equation 7.83 then gives

$$p_{av} = p_o\mathcal{L}(x) \qquad \text{and} \qquad x = \frac{\mathcal{E}}{kT} \tag{7.84}$$

The behavior of the Langevin function is shown in Figure 7.56. At the highest fields $\mathcal{L}(x)$ tends toward saturation at unity. Then, $p_{av} = p_o$, which corresponds to nearly all the dipoles aligning with the field, so increasing the field cannot increase p_{av} anymore. In the low field region, p_{av} increases linearly with the field. In practice, the applied fields are such that all dipolar polarizations fall into this linear behavior region where the Langevin function $\mathcal{L}(x) \approx \frac{1}{3}x$. Then Equation 7.84 becomes

Average induced dipole in orientational polarization

$$p_{av} = \frac{1}{3}\frac{p_o^2\mathcal{E}}{kT} \tag{7.85}$$

Dipolar or orientational polarizability

The **dipolar or orientational polarizability** is then simply

$$\alpha_d = \frac{1}{3}\frac{p_o^2}{kT} \tag{7.86}$$

7.12 IONIC POLARIZATION AND DIELECTRIC RESONANCE

In ionic polarization, as shown in Figure 7.9, the applied field displaces the positive and negative ions in opposite directions, which results in a net dipole moment per ion, called the *induced dipole moment* p_i per ion. We can calculate the ionic polarizability α_i and the ionic contribution to the relative permittivity as a function of frequency by applying an ac field of the form $\mathcal{E} = \mathcal{E}_o\exp(j\omega t)$.

Consider two oppositely charged neighboring ions, *e.g.*, Na^+ and Cl^-, which experience forces $Q\mathcal{E}$ in opposite directions where Q is the magnitude of the ionic charge of each ion as shown in Figure 7.57. The bond between the ions becomes

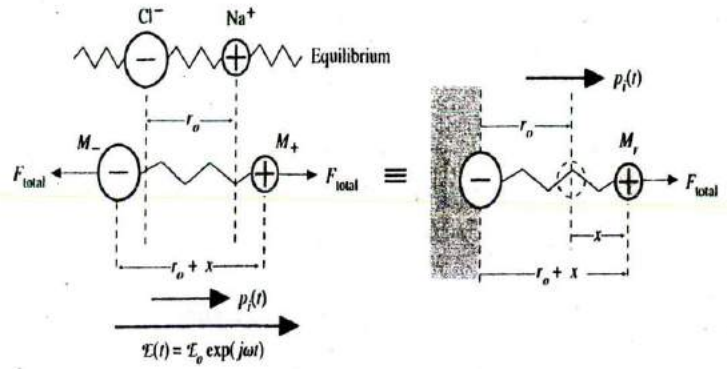


Figure 7.57 Consider a pair of oppositely charged ions. In the presence of an applied field $\mathcal{E}$ along x, the Na^+ and Cl^- ions are displaced from each other by a distance x. The net average (or induced) dipole moment is p_i.

stretched, and the two ions become displaced from the equilibrium separation r_o to a new separation $r_o + x$ as depicted in Figure 7.57. The force $F = Q\mathcal{E}$ of the applied field is the polarizing force, which causes the relative displacement. We take F to be along the x direction. The applied force is resisted by a **restoring force** F_r that is due to the stretching of the bond (Hooke's law) and is proportional to the amount of bond stretching, *i.e.*, $F_r = -\beta x$ where β is the **spring constant** associated with the ionic bond (easily calculated from the potential energy curve of the bond), and the negative sign ensures that F_r is directed in the opposite direction to the applied force. Thus, the net force acting on the ions is $Q\mathcal{E} - \beta x$. As the ions are oscillated by the applied force, they couple some of the energy in the applied field to lattice vibrations and this energy is then lost as heat (lattice vibrations) in the crystal. As in classical mechanics, this type of energy loss through a coupling mechanism can be represented as a **frictional force** (force associated with losses) F_{loss} that acts against the effect of the applied force. This frictional force is proportional to the velocity of the ions or dx/dt, so it is written as $F_{loss} = -\gamma(dx/dt)$ where γ is a proportionality constant that depends on the exact mechanism for the energy loss from the field, and the negative sign ensures that it is opposing the applied field. The total (net) force on the ions is

$$F_{total} = F + F_r + F_{loss} = Q\mathcal{E} - \beta x - \gamma \frac{dx}{dt} \qquad \textit{Total force}$$

Normally we would examine the equations of motion (Newton's second law) under forced oscillation for each ion separately, and then we would use the results to find the overall extension x. An equivalent procedure (as well known in mechanics) is to keep one ion stationary and allow the other one to oscillate with a reduced mass M_r, which is $M_r = (M_+M_-)/(M_+ + M_-)$ where M_+ and M_- are the masses of Na^+ and Cl^- ions, respectively. For example, we can simply examine the oscillations of the Na^+ -ion within the reference frame of the Cl^- -ion (kept "stationary") and attach

a reduced mass M_r to Na^+ as depicted in Figure 7.57. Then Newton's second law gives

Forced oscillations of Na^+–Cl^- ion pair

$$M_r \frac{d^2x}{dt^2} = Q\mathcal{E} - \beta x - \gamma \frac{dx}{dt} \qquad [7.87]$$

It is convenient to put M_r and β together into a new constant ω_I which represents the **resonant** or **natural angular frequency** of the ionic bond, or the natural oscillations when the applied force is removed. Defining $\omega_I = (\beta/M_r)^{1/2}$ and γ_I as γ per unit reduced mass, *i.e.*, $\gamma_I = \gamma/M_r$, we have

Forced dipole oscillator, ionic polarization

$$\frac{d^2x}{dt^2} + \gamma_I \frac{dx}{dt} + \omega_I^2 x = \frac{Q}{M_r}\mathcal{E}_o \exp(j\omega t) \qquad [7.88]$$

Equation 7.88 is a second-order differential equation for the induced displacement x of a pair of neighboring ions about the equilibrium separation as a result of an applied force $Q\mathcal{E}$. It is called the *forced oscillator* equation and is well known in mechanics. (The same equation would describe the damped motion of a ball attached to a spring in a viscous medium and oscillated by an applied force.) The solution to Equation 7.88 will give the displacement $x = x_o \exp(j\omega t)$, which will have the same time dependence as $\mathcal{E}$ but *phase shifted;* that is, x_o will be a complex number. The *relative* displacement of the ions from the equilibrium gives rise to a *net* or **induced polarization** $p_i = Qx$. Thus Equation 7.88 can be multiplied by Q to represent the forced oscillations of the induced dipole. Equation 7.88 is also called the **Lorentz dipole oscillator model.**

The induced dipole p_i will also be phase shifted with respect to the applied force $Q\mathcal{E}$. When we divide p_i by the applied field $\mathcal{E}$, we get the **ionic polarizability** α_i, given by

Ionic polarizability

$$\alpha_i = \frac{p_i}{\mathcal{E}} = \frac{Qx}{\mathcal{E}} = \frac{Q^2}{M_r\left(\omega_I^2 - \omega^2 + j\gamma_I\omega\right)} \qquad [7.89]$$

It can be seen that the polarizability is also a complex number as we expect; there is a phase shift between $\mathcal{E}$ and induced p_i. It therefore has real α_i' and imaginary α_i'' parts and can be written as $\alpha_i = \alpha_i' - j\alpha_i''$. We note that, by convention, the imaginary part is written with a minus sign to keep α_i'' as a positive quantity. Further, when $\omega = 0$, under dc conditions, the ionic polarizability $\alpha_i(0)$ from Equation 7.89 is

DC ionic polarizability

$$\alpha_i(0) = \frac{Q^2}{M_r\omega_I^2} \qquad [7.90]$$

The dc polarizability is a real quantity as there can be no phase shift under dc conditions. We can then write the ionic polarizability in Equation 7.89 in terms of the normalized frequency (ω/ω_I) as

AC ionic polarizability

$$\alpha_i(\omega) = \frac{\alpha_i(0)}{\left[1 - \left(\frac{\omega}{\omega_I}\right)^2 + j\left(\frac{\gamma_I}{\omega_I}\right)\left(\frac{\omega}{\omega_I}\right)\right]} \qquad [7.91]$$

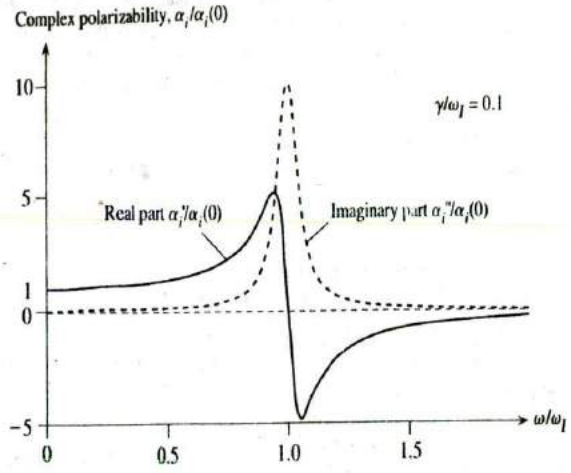


Figure 7.58 A schematic representation of the frequency dependence of the real and imaginary parts of normalized polarizability $\alpha_i/\alpha_i(0)$ versus ω/ω_I.

The dependences of the real and imaginary parts of α_i on the frequency of the field are shown in Figure 7.58 in terms of the normalized frequency (ω/ω_I) for one particular value of the loss factor, $\gamma_I = 0.1\omega_I$. Note that α_i'' peaks at a frequency very close to the ionic bond resonant frequency ω_I (it is exactly ω_I when $\gamma_I = 0$). The sharpness and magnitude of the α_i'' peak depends on the loss factor γ_I. The peak is sharper and higher for smaller γ_I. Notice that α_i' is nearly constant at frequencies lower than ω_I. Indeed, in a dc field, $\alpha_i' = \alpha_i(0)$. But, through ω_I, α_i' shows a rapid change from positive to negative values and then it tends toward zero for frequencies greater than ω_I.

Zero or negative α_i' should not be disconcerting since the actual magnitude of the polarizability is $|\alpha_i| = (\alpha_i'^2 + \alpha_i''^2)^{1/2}$, which is always positive through ω_I and maximum at ω_I. The phase of α_i however changes through ω_I. The phase of α_i, and hence the phase of the polarization with respect to the field, are zero at low frequencies $(\omega \ll \omega_I)$. As the frequency increases, the polarization lags behind the field and the phase of α_i becomes more negative. At $\omega = \omega_I$, the polarization lags behind the field by 90°. However, the rate of change of polarization is in phase with the field oscillations, which leads to a maximum energy transfer. At high frequencies, well above ω_I, the ions cannot respond to the rapidly changing field and the coupling between the field and the ions is negligible. The peak in the α_i'' versus ω behavior around $\omega = \omega_I$ is what is called the **dielectric resonance peak,** and in this particular case it is called the **ionic polarization relaxation peak** and is due to the strong coupling of the applied field with the natural vibrations of the ionic bond at $\omega = \omega_I$.

The resulting relative permittivity ε_r can be found from the Clausius–Mossotti equation. But we also have to consider the electronic polarizability α_e of the two types of ions since this type of polarization operates up to optical frequencies $(\omega \gg \omega_I)$, which means that

Dielectric constant of an ionic solid

$$\frac{\varepsilon_r(\omega) - 1}{\varepsilon_r(\omega) + 2} = \frac{N_i}{3\varepsilon_o}[\alpha_i + \alpha_{e+} + \alpha_{e-}] \qquad [7.92]$$

where N_i is the concentrations of negative and positive ion pairs (assuming an equal number of positive and negative ions), and α_{e+} and α_{e-} are the electronic polarizabilities of the negative and positive ion species, respectively. Inasmuch as α_i is a complex quantity, so is the relative permittivity $\varepsilon_r(\omega)$. We can express Equation 7.92 differently by noting that at very high frequencies, $\omega \gg \omega_l$, $\alpha_i = 0$, and the relative permittivity is then denoted as ε_{rop}. Equation 7.92 then becomes

Dispersion relation for ionic polarization

$$\frac{\varepsilon_r(\omega) - 1}{\varepsilon_r(\omega) + 2} - \frac{\varepsilon_{rop} - 1}{\varepsilon_{rop} + 2} = \frac{N_i \alpha_i}{3\varepsilon_o} = \frac{N_i Q^2}{3\varepsilon_o M_r \left(\omega_l^2 - \omega^2 + j\gamma_l \omega\right)} \qquad [7.93]$$

This is called the **dielectric dispersion relation** between the relative permittivity, due to ionic polarization, and the frequency of the electric field. Figure 7.16b shows the behavior of $\varepsilon_r(\omega)$ with frequency for KCl where ε_r'' peaks at $\omega = \omega_l = 2\pi(4.5 \times 10^{12})$ rad s^{-1} and ε_r' exhibits sharp changes around this frequency. It is clear that as ω gets close to ω_l, there are rapid changes in $\varepsilon_r(\omega)$. The resonant frequencies (ω_l) for ionic polarization relaxations are typically in the infrared frequency range, and the "applied" field in the crystal is then due to a propagating electromagnetic (EM) wave rather than an ac applied field between two external electrodes placed on the crystal.[15]

It should be mentioned that electronic polarization can also be described by the Lorentz oscillator model, and can also be represented by Equation 7.91 if we appropriately replace α_i by α_e and interpret ω_l and γ_l as the resonant frequency and loss factor involved in electronic polarization.

EXAMPLE 7.17 **IONIC POLARIZATION RESONANCE IN KCl** Consider a KCl crystal which has the FCC crystal structure and the following properties. The optical dielectric constant is 2.19, the dc dielectric constant is 4.84, and the lattice parameter a is 0.629 nm. Calculate the dc ionic polarizability $\alpha_i(0)$. Estimate the ionic resonance absorption frequency and compare the value with the experimentally observed resonance at 4.5×10^{12} Hz in Figure 7.16b. The atomic masses of K and Cl are 39.09 and 35.45 g mol^{-1}, respectively.

SOLUTION

At optical frequencies the dielectric constant ε_{rop} is determined by electronic polarization. At low frequencies and under dc conditions, the dielectric constant ε_{rdc} is determined by both electronic and ionic polarization. If N_i is the concentration of negative and positive ion pairs, then equation 7.93 becomes

$$\frac{\varepsilon_{rdc} - 1}{\varepsilon_{rdc} + 2} = \frac{\varepsilon_{rop} - 1}{\varepsilon_{rop} + 2} + \frac{1}{3\varepsilon_o} N_i \alpha_i(0)$$

There are four negative and positive ion pairs per unit cell, and the cell dimension is a. The concentration of negative and positive ion pairs N_i is

$$N_i = \frac{4}{a^3} = \frac{4}{(0.629 \times 10^{-9}\ \text{m})^3} = 1.61 \times 10^{28}\ \text{m}^{-3}$$

[15] More rigorous theories of ionic polarization would consider the interactions of a propagating electromagnetic wave with various phonon modes within the crystal, which is beyond the scope of this book.

Substituting $\varepsilon_{rdc} = 4.84$ and $\varepsilon_{rop} = 2.19$ and N_i in Equation 7.93

$$\alpha_i(0) = \frac{3\varepsilon_o}{N_i}\left[\frac{\varepsilon_{rdc}-1}{\varepsilon_{rdc}+2} - \frac{\varepsilon_{rop}-1}{\varepsilon_{rop}+2}\right] = \frac{3(8.85\times10^{-12})}{1.61\times10^{28}}\left[\frac{4.84-1}{4.84+2} - \frac{2.19-1}{2.19+2}\right]$$

we find

$$\alpha_i(0) = 4.58\times10^{-40}\ \text{F m}^2$$

The relationship between $\alpha_i(0)$ and the resonance absorption frequency involves the reduced mass M_r of the K^+–Cl^- ion pair,

$$M_r = \frac{M_+M_-}{M_+ + M_-} = \frac{(39.09)(35.45)(10^{-3})}{(39.09+35.45)(6.022\times10^{23})} = 3.09\times10^{-26}\ \text{kg}$$

At $\omega = 0$, the polarizability is given by Equation 7.90, so the resonance absorption frequency ω_I is

$$\omega_I = \left[\frac{Q^2}{M_r\alpha_i(0)}\right]^{1/2} = \left[\frac{(1.6\times10^{-19})^2}{(3.09\times10^{-26})(4.58\times10^{-40})}\right]^{1/2} = 4.26\times10^{13}\ \text{rad s}^{-1}$$

or

$$f_I = \frac{\omega_I}{2\pi} = 6.8\times10^{12}\ \text{Hz}$$

This is about a factor of 1.5 greater than the observed resonance absorption frequency of 4.5×10^{12} Hz. Typically one accounts for the difference by noting that the actual ionic charges may not be exactly $+e$ on K^+ and $-e$ on Cl^-, but Q is effectively $0.76e$. Taking $Q = 0.76e$ makes $f_I = 5.15\times10^{12}$ Hz, only 14 percent greater than the observed value. A closer agreement can be obtained by refining the simple theory and considering how many effective dipoles there are in the unit cell along the direction of the applied field.

7.13 DIELECTRIC MIXTURES AND HETEROGENEOUS MEDIA

Many dielectrics are composite materials; that is, they are mixtures of two or more different types of dielectric materials with different relative permittivities and loss factors. The simplest example is a porous dielectric which has small air pores randomly dispersed within the bulk of the material as shown in Figure 7.59a (analogous to a random raisin pudding). Another example would be a dielectric material composed of two distinctly different phases that are randomly mixed, as shown in Figure 7.59b, somewhat like a Swiss cheese that has air bubbles. We often need to find the overall or the **effective dielectric constant** ε_{reff} of the mixture, which is not a trivial problem.[16] This overall ε_{reff} can then be used to treat the mixture as if it were one dielectric substance with this particular dielectric constant; for example, the capacitance can be calculated

[16] The theories that try to represent a heterogeneous medium in terms of effective quantities are called effective medium theories (or approximations). The theory of finding an effective dielectric constant of a mixture has intrigued many famous scientists in the past. Over the years, many quite complicated mixture rules have been developed, and there is no shortage of formulas in this field. Many engineers however still tend to use simple empirical rules to model a composite dielectric. The primary reason is that many theoretical mixture rules depend on the exact knowledge of the geometrical shapes, sizes, and distributions of the mixed phases.

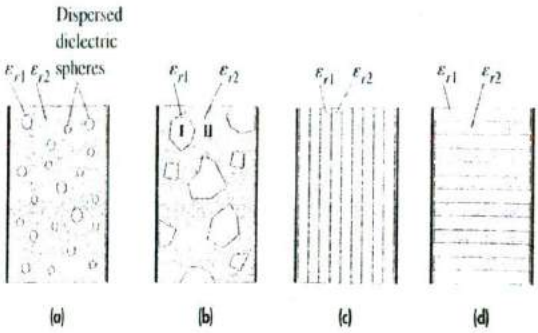


Figure 7.59 Heterogeneous dielectric media examples.
(a) Dispersed dielectric spheres in a dielectric matrix.
(b) A heterogeneous medium with two distinct phases I and II.
(c) Series mixture rule.
(d) Parallel mixture rule.

from $C = \varepsilon_o \varepsilon_{reff} A/d$ by simply using ε_{reff}. It should be emphasized that if mixing occurs at the atomic level so that the material is essentially a *solid solution*, then, in principle, the Clausius–Mossotti equation can be used in which we simply add the polarizabilities of each species of atoms or ions weighted by their concentration. (We did this for CsCl in Example 7.4.) The present problem examines **heterogeneous materials**, and hence excludes such solid solutions.

The theoretical treatment of mixtures can be quite complicated since one has to consider not only individual dielectric properties but also the geometrical shapes, sizes, and distributions of the two (or more) phases present in the composite material. In many cases, empirical rules that have been shown to work have been used to predict ε_{reff}. Consider a heterogeneous dielectric that has two mixed phases I and II with dielectric constants ε_{r1} and ε_{r2}, and volume fractions v_1 and v_2, respectively, $(v_1 + v_2 = 1)$ as in Figure 7.59b. One simple and useful mixture rule is

Generalized mixture rule

$$\varepsilon_{reff}^{n} = v_1 \varepsilon_{r1}^{n} + v_2 \varepsilon_{r2}^{n} \qquad [7.94]$$

where n is an index (a constant), usually determined empirically, that depends on the type of mixture. If we have a parallel stack of plates of I and II in alternating (or in random) sequence between the two electrodes, this would be like many series-connected dielectrics and n would be -1. If the phases are in parallel as plates of I and II stacked on top of each other, as shown in Figure 7.59d, then n is 1. As n approaches 0, Equation 7.94 can be shown to be equivalent to a **logarithmic mixture** rule:

Lichtenecker formula

$$\ln \varepsilon_{reff} = v_2 \ln \varepsilon_{r1} + v_2 \ln \varepsilon_{r2} \qquad [7.95]$$

which is known as the **Lichtenecker formula** (1926). Although its scientific basis is not strong, it has shown remarkable applicability to various heterogeneous media; perhaps due to the fact that it is a kind of compromise between the two extreme limits of series and parallel mixtures.

There is one particular mixture rule for dispersed dielectric spheres (with ε_{r1}), such as air pores, in a continuous dielectric matrix (with ε_{r2}), that works quite well for volume fractions up to about 20 percent, called the **Maxwell–Garnett formula**

$$\frac{\varepsilon_{r\text{eff}} - \varepsilon_{r2}}{\varepsilon_{r\text{eff}} + 2\varepsilon_{r2}} = v_1 \frac{\varepsilon_{r1} - \varepsilon_{r2}}{\varepsilon_{r1} + 2\varepsilon_{r2}} \qquad [7.96]$$

Maxwell–Garnett formula

The Maxwell-Garnett equation can predict the effective dielectric constant of many different types of dielectrics that have dispersed pores. There are other mixture rules, but the above are some of the common types.

EXAMPLE 7.16

LOW-κ POROUS DIELECTRICS FOR MICROELECTRONICS It was mentioned in Chapter 2 that today's high transistor density ICs have multilayers of metal interconnect lines that are separated by an **interlayer dielectric (ILD)**. The speed of the chip (as limited by the RC time constant) depends on the overall interconnect capacitance, which depends on the relative permittivity $\varepsilon_{r\text{ILD}}$ of the ILD. The traditional ILD material has been SiO_2 with $\varepsilon_r = 3.9$. There is much research interest in finding suitable low-κ (also called low-k) materials for such ILD applications, especially in ultralarge-scale integration (ULSI). What is the required porosity in SiO_2 if its effective relative permittivity is to be 2.5?

SOLUTION

The Maxwell–Garnett equation is particularly useful for such porous media calculations. Substituting $\varepsilon_{r2} = 3.9$, $\varepsilon_{r1} = 1$ (air pores), and setting $\varepsilon_{r\text{eff}} = 2.5$ in Equation 7.96 we have

$$\frac{2.5 - 3.9}{2.5 + 2(3.9)} = v_1 \frac{1 - 3.9}{1 + 2(3.9)}$$

and solving gives

$$v_1 = 0.412, \quad \text{or} \quad 41\% \text{ porosity}$$

Such porosity is achievable but it may have side effects such as poorer mechanical properties and lower breakdown voltage. Note that the Lichtenecker formula gives 32.6 percent porosity. As apparent from this example, there is a distinct advantage in starting with a dielectric that has a low initial ε_r, and then using porosity to lower ε_r further. For example, if we start with $\varepsilon_r = 3$, then the same 41 percent porosity will yield $\varepsilon_{r\text{eff}} = 2.05$. Many polymeric materials have ε_r values ~2.5 and have been candidate materials for low-κ ILD applications in microelectronics.

CD Selected Topics and Solved Problems

Selected Topics

Static Dielectric Constant of Materials
Piezoelectric Materials and Devices: Elementary Concepts
Real and Imaginary Dielectric Constant
Conduction in Solid Insulating Materials

Solved Problems

Static Electronic Polarizability
Relative Permittivity of an Ionic Crystal at Low and Optical Frequencies
Piezoelectric Coefficients
Piezoelectric Spark Generator
Electric Field in Coaxial Cables: Double Layer Insulation for Controlling the Maximum Field

DEFINING TERMS

Boundary conditions relate the normal and tangential components of the electric field next to the boundary. The tangential component must be continuous through the boundary. Suppose that $\mathcal{E}_{n1}$ is the normal component of the field in medium 1 at the boundary and ε_{r1} is the relative permittivity in medium 1. Using a similar notation for medium 2, then the boundary condition is $\varepsilon_{r1}\mathcal{E}_{n1} = \varepsilon_{r2}\mathcal{E}_{n2}$.

Clausius–Mossotti equation relates the dielectric constant (ε_r), a macroscopic property, to the polarizability (α), a microscopic property.

Complex relative permittivity ($\varepsilon_r' + j\varepsilon_r''$) has a real part ($\varepsilon_r'$) that determines the charge storage ability and an imaginary part (ε_r'') that determines the energy losses in the material as a result of the polarization mechanism. The real part determines the capacitance through $C = \varepsilon_o\varepsilon_r' A/d$ and the imaginary part determines the electric power dissipation per unit volume as heat by $\mathcal{E}^2\omega\varepsilon_o\varepsilon_r''$.

Corona discharge is a local discharge in a gaseous atmosphere where the field is sufficiently high to cause dielectric breakdown, for example, by avalanche ionization.

Curie temperature T_C is the temperature above which ferroelectricity disappears, that is, the spontaneous polarization of the crystal is lost.

Debye equations attempt to describe the frequency response of the complex relative permittivity $\varepsilon_r' + j\varepsilon_r''$ of a dipolar medium through the use of a single relaxation time τ to describe the sluggishness of the dipoles driven by the external ac field.

Dielectric is a material in which energy can be stored by the polarization of the molecules. It is a material that increases the capacitance or charge storage ability of a capacitor. Ideally, it is a nonconductor of electrical charge so that an applied field does not cause a flow of charge but instead relative displacement of opposite charges and hence polarization of the medium.

Dielectric loss is the electrical energy lost as heat in the polarization process in the presence of an applied ac field. The energy is absorbed from the ac voltage and converted to heat during the polarization of the molecules. It should not be confused with conduction loss $\sigma\mathcal{E}^2$ or V^2/R.

Dielectric strength is the maximum field ($\mathcal{E}_{ds}$) that can be sustained in a dielectric beyond which dielectric breakdown ensues; that is, there is a large conduction current through the dielectric shorting the plates.

Dipolar (orientational) polarization arises when randomly oriented polar molecules in a dielectric are rotated and aligned by the application of a field so as to give rise to a net average dipole moment per molecule. In the absence of the field, the dipoles (polar molecules) are randomly oriented and there is no average dipole moment per molecule. In the presence of the field, the dipoles are rotated, some partially and some fully, to align with the field and hence give rise to a net dipole moment per molecule.

Dipolar relaxation equation describes the time response of the induced dipole moment per molecule in a dipolar material in the presence of a time-dependent applied field. The response of the dipoles depends on their relaxation time, which is the mean time required to dissipate the stored electrostatic energy in the dipole alignment to heat through lattice vibrations or molecular collisions.

Dipole relaxation (dielectric resonance) occurs when the frequency of the applied ac field is such that there is maximum energy transfer from the ac voltage source to heat in the dielectric through the alternating polarization and depolarization of the molecules by the ac field. The stored electrostatic energy is dissipated through molecular collisions and lattice vibrations (in solids). The peak occurs when the angular frequency of the ac field is the reciprocal of the relaxation time.

Electric dipole moment exists when a positive charge $+Q$ is separated from a negative charge $-Q$. Even though the net charge is zero, there is nonetheless an electric dipole moment $\mathbf{p}$ given by $\mathbf{p} = Q\mathbf{x}$ where $\mathbf{x}$ is the distance vector from $-Q$ to $+Q$. Just as two charges exert a Coulombic force on each other, two dipoles also exert a force on each other that depends on the magnitudes of the dipoles, their separation, and orientation.

Electric susceptibility (χ_e) is a material quantity that measures the extent of polarization in the material per unit field. It relates the amount of polarization P at a point in the dielectric to the field $\mathcal{E}$ at that point via $P = \chi_e \varepsilon_o \mathcal{E}$. If ε_r is the relative permittivity, then $\chi_e = \varepsilon_r - 1$. Vacuum has no electric susceptibility.

Electromechanical breakdown and electrofracture are breakdown processes that directly or indirectly involve electric field–induced mechanical weakening, for example, crack propagation, or mechanical deformation that eventually lead to dielectric breakdown.

Electronic bond polarization is the displacement of valence electrons in the bonds in covalent solids (*e.g.*, Ge, Si). It is a collective displacement of the electrons in the bonds with respect to the positive nuclei.

Electronic polarization is the displacement of the electron cloud of an atom with respect to the positive nucleus. Its contribution to the relative permittivity of a solid is usually small.

External discharges are discharges or shorting currents over the surface of the insulator when the conductance of the surface increases as a result of surface contamination, for example, excessive moisture, deposition of pollutants, dirt, dust, and salt spraying. Eventually the contaminated surface develops sufficient conductance to allow discharge between the electrodes at a field below the normal breakdown strength of the insulator. Dielectric breakdown over the surface of an insulation is termed **surface tracking.**

Ferroelectricity is the occurrence of spontaneous polarization in certain crystals such as barium titanate ($BaTiO_3$). Ferroelectric crystals have a permanent polarization **P** as a result of spontaneous polarization. The direction of **P** can be defined by the application of an external field.

Gauss's law is a fundamental law of physics that relates the surface integral of the electric field over a closed (hypothetical) surface to the sum of all the charges enclosed within the surface. If $\mathcal{E}_n$ is the field normal to a small surface area dA and Q_{total} is the enclosed total charge, then over the whole closed surface $\varepsilon_o \oint \mathcal{E}_n \, dA = Q_{total}$.

Induced polarization is the polarization of a molecule as a result of its placement in an electric field. The induced polarization is along the direction of the field. If the molecule is already polar, then induced polarization is the additional polarization that arises due to the applied field alone and it is directed along the field.

Insulation aging is a term used to describe the physical and chemical deterioration in the properties of the insulation so that its dielectric breakdown characteristics worsen with time. Aging therefore determines the useful life of the insulation.

Interfacial polarization occurs whenever there is an accumulation of charge at an interface between two materials or between two regions within a material. Grain boundaries and electrodes are regions where charges generally accumulate and give rise to this type of polarization.

Internal discharges are partial discharges that take place in microstructural voids, cracks, or pores within the dielectric where the gas atmosphere (usually air) has lower dielectric strength. A porous ceramic, for example, would experience partial discharges if the field is sufficiently large. Initially, the pore size (or the number of pores) may be small and the partial discharge insignificant, but with time the partial discharge erodes the internal surfaces of the void. Eventually (and usually) an *electrical tree* type of discharge develops from a partial discharge that has been eroding the dielectric. The erosion of the dielectric by the partial discharge propagates like a branching tree. The "tree branches" are erosion channels, filaments of various sizes, in which gaseous discharge takes place and forms a conducting channel during operation.

Intrinsic breakdown or **electronic breakdown** commonly involves the avalanche multiplication of electrons (and holes in solids) by impact ionization in the presence of high electric fields. The large number of free carriers generated by the avalanche of impact ionizations leads to a runaway current between the electrodes and hence to insulation breakdown.

Ionic polarization is the relative displacement of oppositely charged ions in an ionic crystal that results in the polarization of the whole material. Typically, ionic polarization is important in ionic crystals below the infrared wavelengths.

Local field ($\mathcal{E}_{loc}$) is the true field experienced by a molecule in a dielectric that arises from the free charges on the plates and all the induced dipoles

surrounding the molecule. The true field at a molecule is not simply the applied field (V/d) because of the field of the neighboring induced dipoles.

Loss tangent or **tan** δ is the ratio of the dielectric constant's imaginary part to the real part, $\varepsilon_r''/\varepsilon_r'$. The angle δ is the phase angle between the capacitive current and the total current. If there is no dielectric loss, then the two currents are the same and $\delta = 0$.

Partial discharge occurs when only a local region of the dielectric is exhibiting discharge, so the discharge does not directly connect the two electrodes.

Piezoelectric material has a noncentrosymmetric crystal structure that leads to the generation of a polarization vector P, or charges on the crystal surfaces, upon the application of a mechanical stress. When strained, a piezoelectric crystal develops an internal field and therefore exhibits a voltage difference between two of its faces.

PLZT, lead lanthanum zirconate titanate, is a PZT-type material with lanthanum occupying the Pb site.

Polarizability (α) is the ability of an atom or molecule to become polarized in the presence of an electric field. It is induced polarization in the molecule per unit field along the field direction.

Polarization is the separation of positive and negative charges in a system so that there is a net electric dipole moment per unit volume.

Polarization vector (P) measures the extent of polarization in a unit volume of dielectric matter. It is the vector sum of dielectric dipoles per unit volume. If **p** is the average dipole moment per molecule and n is the number of molecules per unit volume, then $\mathbf{P} = n\mathbf{p}$. In a polarized dielectric matter (*e.g.*, in an electric field), the bound surface charge density σ_p due to polarization is equal to the normal component of **P** at that point, $\sigma_p = P_{\text{normal}}$.

Poling is the application of a temporary electric field to a piezoelectric (or ferroelectric) material, generally at an elevated temperature, to align the polarizations of various grains and thereby develop piezoelectric behavior.

Pyroelectric material is a polar dielectric (such as barium titanate) in which a temperature change ΔT induces a proportional change ΔP in the polarization, that is, $\Delta P = p\,\Delta T$, where p is the pyroelectric coefficient of the crystal.

PZT is a general acronym for the lead zirconate titanate ($PbZrO_3$-$PbTiO_3$ or $PbTi_{0.48}Zr_{0.52}O_3$) family of crystals.

Q-factor or **quality factor** for an impedance is the ratio of its reactance to its resistance. The Q-factor of a capacitor is X_t/R_p where $X_t = 1/\omega C$ and R_p is the equivalent parallel resistance that represents the dielectric and conduction losses. The Q-factor of a resonant circuit measures the circuit's peak response at the resonant frequency and also its bandwidth. The greater the Q, the higher the peak response and the narrower the bandwidth. For a series RLC resonant circuit,

$$Q = \frac{\omega_o L}{R} = \frac{1}{\omega_o C R}$$

where ω_o is the resonant angular frequency, $\omega_o = 1/\sqrt{LC}$. The width of the resonant response curve between half-power points is $\Delta\omega = \omega_o/Q$.

Relative permittivity (ε_r) or **dielectric constant** of a dielectric is the fractional increase in the stored charge per unit voltage on the capacitor plates due to the presence of the dielectric between the plates (the whole space between the plates is assumed to be filled). Alternatively, we can define it as the fractional increase in the capacitance of a capacitor when the insulation between the plates is changed from a vacuum to a dielectric material, keeping the geometry the same.

Relaxation time (τ) is a characteristic time that determines the sluggishness of the dipole response to an applied field. It is the mean time for the dipole to lose its alignment with the field due to its random interactions with the other molecules through molecular collisions, lattice vibrations, and so forth.

Surface tracking is an *external dielectric breakdown* that occurs over the surface of the insulation.

Temperature coefficient of capacitance (TCC) is the fractional change in the capacitance per unit temperature change.

Thermal breakdown is a breakdown process that involves thermal runaway, which leads to a runaway current or discharge between the electrodes. If the heat generated by dielectric loss, due to ε_r'', or Joule heating, due to finite σ, cannot be removed sufficiently rapidly, then the temperature of the dielectric rises, which increases the conductivity and the dielectric loss. The increases in ε_r'' and σ lead to more heat generation and a further rise in the temperature, so thermal runaway ensues, followed by either a large shorting current or local thermal decomposition of the insulation accompanied by a partial discharge in this region.

Transducer is a device that converts electrical energy into another form of usable energy or vice versa. For example, piezoelectric transducers convert electrical energy to mechanical energy and vice versa.

QUESTIONS AND PROBLEMS

7.1 Relative permittivity and polarizability

a. Show that the local field is given by

$$E_{\text{loc}} = \mathcal{E}\left(\frac{\varepsilon_r + 2}{3}\right) \qquad \textit{Local field}$$

b. Amorphous selenium (a-Se) is a high-resistivity semiconductor that has a density of approximately 4.3 g cm^{-3} and an atomic number and mass of 34 and 78.96, respectively. Its relative permittivity at 1 kHz has been measured to be 6.7. Calculate the relative magnitude of the local field in a-Se. Calculate the polarizability per Se atom in the structure. What type of polarization is this? How will ε_r depend on the frequency?

c. If the electronic polarizability of an isolated atom is given by

$$\alpha_e \approx 4\pi\varepsilon_0 r_a^3$$

where r_a is the radius of the atom, then calculate the electronic polarizability of an isolated Se atom, which has $r_a = 0.12$ nm, and compare your result with that for an atom in a-Se. Why is there a difference?

7.2 Electronic polarization and SF_6 Because of its high dielectric strength, SF_6 (sulfur hexafluoride) gas is widely used as an insulator and a dielectric in HV applications such as HV transformers, switches, circuit breakers, transmission lines, and even HV capacitors. The SF_6 gas at 1 atm and at room temperature has a dielectric constant of 1.0015. The number of SF_6 molecules per unit volume N can be found by the gas law, $P = (N/N_A)RT$. Calculate the electronic polarizability α_e of the SF_6 molecule. How does α_e compare with the α_e versus Z line in Figure 7.4? (Note: The SF_6 molecule has no net dipole. Assume that the overall polarizability of SF_6 is due to electronic polarization.)

7.3 Electronic polarization in liquid xenon Liquid xenon has been used in radiation detectors. The density of the liquid is 3.0 g cm^{-3}. What is the relative permittivity of liquid xenon given its electronic polarizability in Table 7.1? (The experimental ε_r is 1.96.)

7.4 Relative permittivity, bond strength, bandgap, and refractive index Diamond, silicon, and germanium are covalent solids with the same crystal structure. Their relative permittivities are shown in Table 7.10.

a. Explain why ε_r increases from diamond to germanium.

b. Calculate the polarizability per atom in each crystal and then plot polarizability against the elastic modulus Y (Young's modulus). Should there be a correlation?

c. Plot the polarizability from part (*b*) against the bandgap energy E_g. Is there a relationship?

d. Show that the refractive index n is $\sqrt{\varepsilon_r}$. When does this relationship hold and when does it fail?

e. Would your conclusions apply to ionic crystals such as NaCl?

Table 7.10 Properties of diamond, Si, and Ge

	ε_r	M_{at}	Density (g cm^{-3})	α_e	Y (GPa)	E_g (eV)	n
Diamond	5.8	12	3.52		827	5.5	2.42
Si	11.9	28.09	2.33		190	1.12	3.45
Ge	16	72.61	5.32		75.8	0.67	4.09

7.5 **Dipolar liquids** Given the static dielectric constant of water as 80, its high-frequency dielectric constant (due to electronic polarization) as 4, and its density as 1 g cm^{-3}, calculate the permanent dipole moment p_o per water molecule assuming that it is the orientational and electronic polarization of individual molecules that gives rise to the dielectric constant. Use both the simple relationship in Equation 7.14 where the local field is the same as the macroscopic field and also the Clausius–Mossotti equation and compare your results with the permanent dipole moment of the water molecule which is 6.1×10^{-30} C m. What is your conclusion? What is ε_r calculated from the Clausius–Mossotti equation taking the true p_o (6.1×10^{-30} C m) of a water molecule? (Note: Static dielectric constant is due to both orientational and electronic polarization. The Clausius–Mossotti equation does not apply to dipolar materials because the local field is not described by the Lorentz field.)

7.6 **Dielectric constant of water vapor or steam** The isolated water molecule has a permanent dipole p_o of 6.1×10^{-30} C m. The electronic polarizability α_e of the water molecule under dc conditions is about 4×10^{-40} C m. What is the dielectric constant of steam at a pressure of 10 atm (10×10^5 Pa) and at a temperature of 400 °C? [Note: The number of water molecules per unit volume N can be found from the simple gas law, $P = (N/N_A)RT$. The Clausius–Mossotti equation does not apply to orientational polarization. Since N is small, use Equation 7.14.]

7.7 **Dipole moment in a nonuniform electric field** Figure 7.60 shows an electric dipole moment p in a nonuniform electric field. Suppose the gradient of the field is $d\mathcal{E}/dx$ at the dipole p, and the dipole is oriented to be along the direction of increasing $\mathcal{E}$ as in Figure 7.60. Show that the *net force* acting on this dipole is given by

Net force on a dipole

$$F = p\frac{d\mathcal{E}}{dx}$$

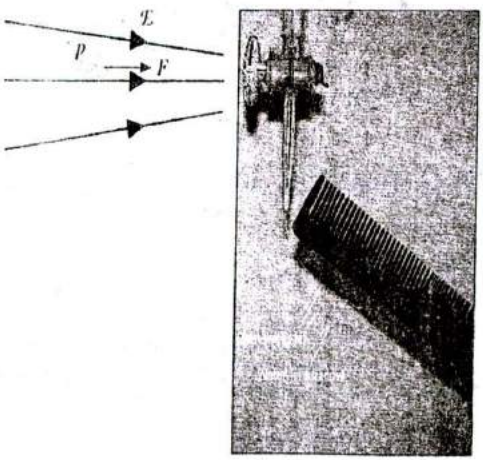


Figure 7.60
Left: A dipole moment in a nonuniform field experiences a net force F that depends on the dipole moment p and the field gradient $d\mathcal{E}/dx$.
Right: When a charged comb (by combing hair) is brought close to a water jet, the field from the comb polarizes the liquid by orientational polarization. The induced polarization vector P and hence the liquid is attracted to the comb where the field is higher.

Which direction is the force? What happens to this net force when the dipole moment is facing the direction of decreasing field? Given that a dipole normally also experiences a torque as described in Section 7.3.2, explain qualitatively what happens to a randomly placed dipole in a nonuniform electric field. Explain the experimental observation of bending a flow of water by a nonuniform field from a charged comb as shown in the photograph in Figure 7.60? (Remember that a dielectric medium placed in a field develops polarization P directed along the field.)

7.8 **Ionic and electronic polarization** Consider a CsBr crystal that has the CsCl unit cell crystal structure (one Cs^+–Br^- pair per unit cell) with a lattice parameter (a) of 0.430 nm. The electronic polarizability of Cs^+ and Br^- ions are 3.35×10^{-40} F m^2 and 4.5×10^{-40} F m^2, respectively, and the mean ionic polarizability per ion pair is 5.8×10^{-40} F m^2. What is the low-frequency dielectric constant and that at optical frequencies?

7.9 **Electronic and ionic polarization in KCl** KCl has the same crystal structure as NaCl. KCl's lattice parameter is 0.629 nm. The ionic polarizability per ion pair (per K^+–Cl^- ion) is 4.58×10^{-40} F m^2. The electronic polarizability of K^+ is 1.26×10^{-40} F m^2 and that of Cl^- is 3.41×10^{-40} F m^2. Calculate the dielectric constant under dc operation and at optical frequencies. Experimental values are 4.84 and 2.19.

7.10 **Debye relaxation** We will test the Debye equations for approximately calculating the real and imaginary parts of the dielectric constant of water just above the freezing point at 0.2 °C. Assume the following values in the Debye equations for water: $\varepsilon_{rdc} = 87.46$ (dc), $\varepsilon_{r\infty} = 4.87$ (at $f = 300$ GHz well beyond the relaxation peak), and $\tau = 1/\omega_0 = (2\pi 9.18\ \text{GHz})^{-1} = 0.017$ ns. Calculate the real and imaginary, ε_r' and ε_r'', parts of ε_r for water at frequencies in Table 7.11, and plot both the experimental values and your calculations on a linear–log plot (frequency on the log axis). What is your conclusion? (Note: It is possible to obtain a better agreement by using two relaxation times or using more sophisticated models.)

Table 7.11 Dielectric properties of water at 0.2 °C

	f (GHz)												
	0.3	**0.5**	**1**	**1.5**	**3**	**5**	**9.18**	**10**	**20**	**40**	**70**	**100**	**300**
ε_r'	87.46	87.25	86.61	85.34	76.20	68.19	46.13	42.35	19.69	10.16	7.20	6.14	4.87
ε_r''	2.60	4.50	8.85	13.18	24.28	34.53	40.55	40.24	30.23	17.68	11.15	8.31	3.68

SOURCE: Data extracted from R. Buchner et al., *Chem. Phys. Letts*, **306**, 57, 1999.

***7.11** **Debye and non-Debye relaxation and Cole–Cole plots** Consider the Debye equation

$$\varepsilon_r = \varepsilon_{r\infty} + \frac{\varepsilon_{rdc} - \varepsilon_{r\infty}}{1 + j\omega\tau} \qquad \textit{Debye relaxation}$$

and also the **generalized dielectric relaxation** equation, which "stretches" (broadens) the Debye function,

$$\varepsilon_r = \varepsilon_{r\infty} + \frac{\varepsilon_{rdc} - \varepsilon_{r\infty}}{[1 + (j\omega\tau)^\alpha]^\beta} \qquad \textit{Generalized dielectric relaxation}$$

Take $\tau = 1$, $\varepsilon_{rdc} = 5$, $\varepsilon_{r\infty} = 2$, and $\alpha = 0.8$, and $\beta = 1$. Plot the real and imaginary parts of ε_r versus frequency (on a log scale) for both functions from $\omega = 0, 0.1/\tau, 1/3\tau, 1/\tau, 3/\tau$, and 10τ. For the same ω values, plot ε_r'' versus ε_r' (Cole–Cole plot) for both functions using a graph in which the x and y axes have the same divisions. What is your conclusion?

7.12 **Equivalent circuit of a polyester capacitor** Consider a 1 nF polyester capacitor that has a polymer (PET) film thickness of 1 μm. Calculate the equivalent circuit of this capacitor at 50 °C and at 120 °C for operation at 1 kHz. What is your conclusion?

7.13 **Student microwaves mashed potatoes** A microwave oven uses electromagnetic waves at 2.48 GHz to heat food by dielectric loss, that is, making use of ε_r'' of the food material, which normally has substantial water content. An undergraduate student microwaves 10 cm^3 of mashed potatoes in 60 seconds. The microwave generates an rms field of $\mathcal{E}_{rms}$ of 200 V cm^{-1} in mashed potatoes. At 2.48 GHz, mashed potatoes have $\varepsilon_r'' = 21$. Calculate the average power dissipated per cm^3, and also the total energy dissipated heating the food. (Note: You can use $\mathcal{E}_{rms}$ instead of E in Equation 7.32.)

7.14 **Dielectric loss per unit capacitance** Consider the three dielectric materials listed in Table 7.12 with the real and imaginary dielectric constants ε_r' and ε_r''. At a given voltage, which dielectric will have the lowest power dissipation per unit capacitance at 1 kHz and at an operating temperature of 50 °C? Is this also true at 120 °C?

Table 7.12 Dielectric properties of three insulators at 1 kHz

	$T = 50$ °C		$T = 120$ °C	
Material	ε_r'	ε_r''	ε_r'	ε_r''
Polycarbonate	2.47	0.003	2.535	0.003
PET	2.58	0.003	2.75	0.027
PEEK	2.24	0.003	2.25	0.003

[1] SOURCE: Data taken using a DEA by Kasap and Nomura (1995).

7.15 **Parallel and series equivalent circuits** Figure 7.61 shows simplified parallel and series equivalent circuits for a capacitor. The elements R_p and C_p in the parallel circuit and the elements R_s and C_s in the series circuit are related. We can write down the impedance Z_{AB} between the terminals A and B for both the circuits, and then equate Z_{AB}(parallel) = Z_{AB}(series). Show that

Equivalent series resistance and capacitance

$$R_s = \frac{R_p}{1 + (\omega R_p C_p)^2} \quad \text{and} \quad C_s = C_p\left[1 + \frac{1}{(\omega R_p C_p)^2}\right]$$

and similarly by considering the admittance (1/impedance),

Equivalent series resistance and capacitance

$$R_p = R_s\left[1 + \frac{1}{(\omega R_s C_s)^2}\right] \quad \text{and} \quad C_p = \frac{C_s}{1 + (\omega R_s C_s)^2}$$

A 10 nF capacitor operating at 1 MHz has a parallel equivalent resistance of 100 kΩ. What are C_s and R_s?

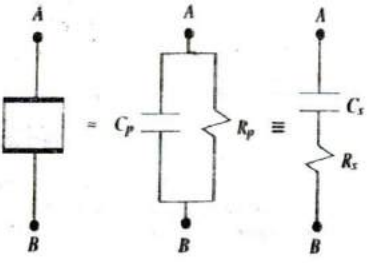


Figure 7.61 An equivalent parallel R_p and C_p circuit is equivalent to a series R_s and C_s circuit. The elements R_p and C_p in the parallel circuit are related to the elements R_s and C_s in the series circuit.

7.16 **Tantalum capacitors** Electrolytic capacitors tend to be modeled by a series $R_s + j\omega C_s$ equivalent circuit. A nominal 22 μF Ta capacitor (22 μF at low frequencies) has the following properties at 10 kHz:

$\varepsilon_r' \approx 20$ (at this frequency), $\tan\delta \approx 0.05$, dielectric thickness $d = 0.16\ \mu m$, effective area $A = 150\ cm^2$. Calculate C_p, R_p, C_s, and R_s.

7.17 Tantalum versus niobium oxide capacitors Niobium oxide (Nb_2O_5) is a competing dielectric to Ta_2O_5 (the dielectric in the tantalum capacitor). The dielectric constants are 41 for Nb_2O_5 and 27 for Ta_2O_5. For operation at the same voltage, the Ta_2O_5 thickness is 0.17 μm, and that of Nb_2O_5 is 0.25 μm. Explain why the niobium oxide capacitor is superior (or inferior) to the Ta capacitor. (Use a quantitative argument, such as the capacitance per unit volume.) What other factors would you consider if you were choosing between the two?

***7.18 TCC of a polyester capacitor** Consider the parallel plate capacitor equation

$$C = \frac{\varepsilon_0 \varepsilon_r x y}{z}$$

where ε_r is the relative permittivity (or ε_r'), x and y are the side lengths of the dielectric so that xy is the area A, and z is the thickness of the dielectric. The quantities ε_r, x, y, and z change with temperature. By differentiating this equation with respect to temperature, show that the **temperature coefficient of capacitance** (TCC) is

Temperature coefficient of capacitance

$$\text{TCC} = \frac{1}{C}\frac{dC}{dT} = \frac{1}{\varepsilon_r}\frac{d\varepsilon_r}{dT} + \lambda$$

where λ is the linear expansion coefficient defined by

$$\lambda = \frac{1}{L}\frac{dL}{dT}$$

where L stands for any length of the material (x, y, or z). Assume that the dielectric is isotropic and λ is the same in all directions. Using ε_r' versus T behavior in Figure 7.62 and taking $\lambda = 50 \times 10^{-6}\ K^{-1}$ as a typical value for polymers, predict the TCC at room temperature and at 10 kHz.

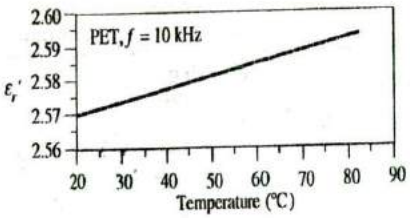


Figure 7.62 Temperature dependence of ε_r' at 10 kHz.
SOURCE: Data taken by Kasap and Maeda (1995).

7.19 Dielectric breakdown of gases and Paschen curves Dielectric breakdown in gases typically involves the avalanche ionization of the gas molecules by energetic electrons accelerated by the applied field. The mean free path between collisions must be sufficiently long to allow the electrons to gain sufficient energy from the field to impact ionize the gas molecules. The breakdown voltage V_b, between two electrodes depends on the distance d between the electrodes as well as the gas pressure P, as shown in Figure 7.63. V_b, versus Pd plots are called **Paschen curves.** We consider gaseous insulation, air and SF_6, in an HV switch.

a. What is the breakdown voltage between two electrodes of a switch separated by a 5 mm gap with air at 1 atm when the gaseous insulation is air and when it is SF_6?

b. What are the breakdown voltages in the two cases when the pressure is 10 times greater? What is your conclusion?

c. At what pressure is the breakdown voltage a minimum?

d. What air gap spacing d at 1 atm gives the minimum breakdown voltage?

e. What would be the reasons for preferring gaseous insulation over liquid or solid insulation?

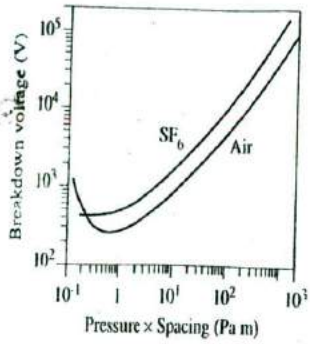


Figure 7.63 Breakdown voltage versus (pressure × electrode spacing) (Paschen curves).

***7.20 Capacitor design** Consider a nonpolarized 100 nF capacitor design at 60 Hz operation. Note that there are three candidate dielectrics, as listed in Table 7.13.

a. Calculate the volume of the 100 nF capacitor for each dielectric, given that they are to be used under low voltages and each dielectric has its minimum fabrication thickness. Which one has the smallest volume?

b. How is the volume affected if the capacitor is to be used at a 500 V application and the maximum field in the dielectric must be a factor of 2 less than the dielectric strength? Which one has the smallest volume?

c. At a 500 V application, what is the power dissipated in each capacitor at 60 Hz operation? Which one has the lowest dissipation?

Table 7.13 Comparison of dielectric properties at 60 Hz (typical values)

	Polymer Film PET	Ceramic TiO_2	High-κ Ceramic ($BaTiO_3$ based)
Name	Polyester	Polycrystalline titania	X7R
ε_r'	3.2	90	1800
$\tan\delta$	5×10^{-3}	4×10^{-4}	5×10^{-2}
E_{br} (kV cm^{-1})	150	50	100
Typical minimum thickness	1–2 μm	10 μm	10 μm

***7.21 Dielectric breakdown in a coaxial cable** Consider a coaxial undersea high-voltage cable as in Figure 7.64a. The current flowing through the inner conductor generates heat, which has to flow through the dielectric insulation to the outer conductor where it will be carried away by conduction and convection. We will assume that steady state has been reached and the inner conductor is carrying a dc current I. Heat generated per unit second $Q' = dQ/dt$ by Joule heating of the inner conductor is

Rate of heat generation

$$Q' = RI^2 = \frac{\rho L I^2}{\pi a^2} \qquad [7.97]$$

where ρ is the resistivity, a the radius of the conductor, and L the cable length.

This heat flows radially out from the inner conductor through the dielectric insulator to the outer conductor, then to the ambient. This heat flow is by thermal conduction through the dielectric. The rate of heat flow Q' depends on the temperature difference $T_i - T_o$ between the inner and outer conductors;

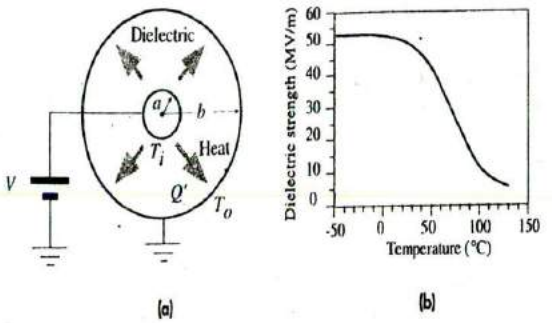


Figure 7.64
(a) The Joule heat generated in the core conductor flows outward radially through the dielectric material.
(b) Typical temperature dependence of the dielectric strength of a polyethylene-based polymeric insulation.

on the sample geometry (a, b, and L), and on the thermal conductivity κ of the dielectric. From elementary thermal conduction theory, this is given by

$$Q' = (T_i - T_o)\frac{2\pi\kappa L}{\ln\left(\frac{b}{a}\right)} \qquad [7.98]$$

Rate of heat conduction

The inner core temperature T_i rises until, in the steady state, the rate of Joule heat generation by the electric current in Equation 7.97 is just removed by the rate of thermal conduction through the dielectric insulation, given by Equation 7.98.

a. Show that the inner conductor temperature is

$$T_i = T_o + \frac{\rho I^2}{2\pi^2 a^2 \kappa}\ln\left(\frac{b}{a}\right) \qquad [7.99]$$

Steady-state inner conductor temperature

b. The breakdown occurs at the maximum field point, which is at $r = a$, just outside the inner conductor and is given by (see Example 7.11).

$$\mathcal{E}_{max} = \frac{V}{a\ln\left(\frac{b}{a}\right)} \qquad [7.100]$$

Maximum field in a coaxial cable

The dielectric breakdown occurs when $\mathcal{E}_{max}$ reaches the dielectric strength $\mathcal{E}_{br}$. However the dielectric strength $\mathcal{E}_{br}$ for many polymeric insulation materials depends on the temperature, and generally it decreases with temperature, as shown for a typical example in Figure 7.64b. If the load current I increases, then more heat Q' is generated per second and this leads to a higher inner core temperature T_i by virtue of Equation 7.99. The increase in T_i with I eventually lowers $\mathcal{E}_{br}$ so much that it becomes equal to $\mathcal{E}_{max}$ and the insulation breaks down (thermal breakdown). Suppose that a certain coaxial cable has an aluminum inner conductor of diameter 10 mm and resistivity 27 nΩ m. The insulation is 3 mm thick and is a polyethylene-based polymer whose long-term dc dielectric strength is shown in Figure 7.64b. Suppose that the cable is carrying a voltage of 40 kV and the outer shield temperature is the ambient temperature, 25 °C. Given that the thermal conductivity of the polymer is about 0.3 W K^{-1} m^{-1}, at what dc current will the cable fail?

c. Rederive T_i in Equation 7.99 by considering that ρ depends on the temperature as $\rho = \rho_o[1 + \alpha_o(T - T_o)]$ (Chapter 2). Recalculate the maximum current in *b* given that $\alpha_o = 3.9 \times 10^{-3}$ °C^{-1} at 25 °C.

7.22 **Piezoelectricity** Consider a quartz crystal and a PZT ceramic filter both designed for operation at $f_s = 1$ MHz. What is the bandwidth of each? Given Young's modulus (Y), density (ρ) for each, and that the filter is a disk with electrodes and is oscillating radially, what is the diameter of the disk for each material? For quartz, $Y = 80$ GPa and $\rho = 2.65$ g cm^{-3}. For PZT, $Y = 70$ GPa and $\rho = 7.7$ g m^{-3}.

Assume that the velocity of mechanical oscillations in the crystal is $v = \sqrt{Y/\rho}$ and the wavelength $\lambda = v/f_s$. Consider only the fundamental mode ($n = 1$).

7.23 Piezoelectric voltage coefficient The application of a stress T to a piezoelectric crystal leads to a polarization P and hence to an electric field $\mathcal{E}$ in the crystal such that

Piezoelectric voltage coefficient

$$\mathcal{E} = gT$$

where g is the *piezoelectric voltage coefficient*. If $\varepsilon_o\varepsilon_r$ is the permittivity of the crystal, show that

$$g = \frac{d}{\varepsilon_o\varepsilon_r}$$

A $BaTiO_3$ sample, along a certain direction (called 3), has $d = 190$ pC N^{-1}, and its $\varepsilon_r \approx 1900$ along this direction. What do you expect for its g coefficient for this direction and how does this compare with the measured value of approximately 0.013 m^2 C^{-1}?

7.24 Piezoelectricity and the piezoelectric bender

a. Consider using a piezoelectric material in an application as a mechanical positioner where the displacements are expected to be small (as in a scanning tunneling microscope). For the piezoelectric plate shown in Figure 7.65a, we will take $L = 20$ mm, $W = 10$ mm, and D (thickness) = 0.25 mm. Under an applied voltage of V, the plate changes length, width, and thickness according to the piezoelectric coefficients d_{ij}, relating the applied field along i to the resulting strain along j.

Suppose we define direction 3 along the thickness D and direction 1 along the length L, as shown in Figure 7.65a. Show that the changes in the thickness and length are

$$\delta D = d_{33}\,V$$

Piezoelectric effects

$$\delta L = \left(\frac{L}{D}\right) d_{31}\,V$$

Given $d_{33} \approx 500 \times 10^{-12}$ m V^{-1} and $d_{31} \approx -250 \times 10^{-12}$ m V^{-1}, calculate the changes in the length and thickness for an applied voltage of 100 V. What is your conclusion?

b. Consider two oppositely poled and joined ceramic plates, A and B, forming a bimorph, as shown in Figure 7.65b. This piezoelectric bimorph is mounted as a cantilever; one end is fixed and the other end is free to move. Oppositely poled means that the electric field elongates A and contracts B, and the two relative motions *bend* the plate. The displacement h of the tip of the cantilever is given by

Piezoelectric bending

$$h = \frac{3}{2} d_{31} \left(\frac{L}{D}\right)^2 V$$

What is the deflection of the cantilever for an applied voltage of 100 V? What is your conclusion?

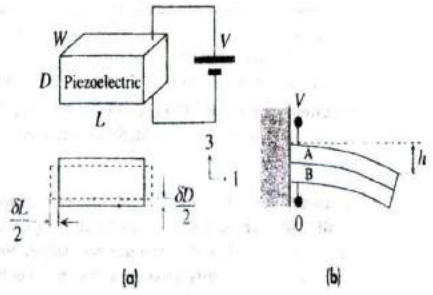


Figure 7.65
(a) A mechanical positioner using a piezoelectric plate under an applied voltage of V.
(b) A cantilever-type piezoelectric bender. An applied voltage bends the cantilever.

7.25 Piezoelectricity The wavelength λ of mechanical oscillations in a piezoelectric slab satisfies

$$n\left(\frac{1}{2}\lambda\right) = L$$

where n is an integer, L is the length of the slab along which mechanical oscillations are set up, and the wavelength λ is determined by the frequency f and velocity v of the waves. The ultrasonic wave velocity v depends on Young's modulus Y as

$$v = \left(\frac{Y}{\rho}\right)^{1/2}$$

where ρ is the density. For quartz, $Y = 80$ GPa and $\rho = 2.65$ g cm^{-3}. Considering the fundamental mode ($n = 1$), what are practical dimensions for crystal oscillators operating at 1 kHz and 1 MHz?

7.26 Pyroelectric detectors Consider two different radiation detectors using PZT and PVDF as pyroelectric materials whose properties are summarized in Table 7.14. The receiving area is 4 mm^2. The thicknesses of the PZT ceramic and the PVDF polymer film are 0.1 mm and 0.005 mm, respectively. In both cases the incident radiation is chopped periodically to allow the radiation to pass for a duration of 0.05 s.

a. Calculate the magnitude of the output voltage for each detector if both receive a radiation of intensity 10 μW cm^{-2}. What is the corresponding current in the circuit? In practice, what would limit the magnitude of the output voltage?

b. What is the minimum detectable radiation intensity if the minimum detectable signal voltage is 10 nV?

Table 7.14 Properties of PZT and PVDF

	ε_r'	Pyroelectric Coefficient ($\times 10^{-6}$ C m^{-2} K^{-1})	Density (g cm^{-3})	Heat Capacity (J K^{-1} g^{-1})
PZT	290	380	7.7	0.3
PVDF	12	27	1.76	1.3

7.27 $LiTaO_3$ pyroelectric detector $LiTaO_3$ (lithium tantalate) detectors are available commercially. $LiTaO_3$ has the following properties: pyroelectric coefficient $p \approx 200 \times 10^{-6}$ C m^{-2} K^{-1}, density $\rho = 7.5$ g cm^{-3}, specific heat capacity $c_v = 0.43$ J K^{-1} g^{-1}. A particular detector has a cylindrical crystal with a diameter of 10 mm and thickness of 0.2 mm. Suppose we chop the input radiation and allow the radiation to fall on the detector for short periods of time. Each input radiation pulse has a duration of $\Delta t = 10$ ms. (The time between the radiation pulses is long, so consider only the response of the detector to a single pulse of radiation.) Suppose that all the incident radiation is absorbed. If the input radiation has an intensity of 10 μW cm^{-2}, calculate the pyroelectric current, and the maximum possible output voltage that can be generated assuming that the input impedance of the amplifier is sufficiently large to be negligible. What is the current responsivity of this detector? What are the major assumptions in your calculation of the voltage signal?

***7.28 Pyroelectric detectors** Consider a typical pyroelectric radiation detector circuit as shown in Figure 7.66. The FET circuit acts as a voltage follower (source follower). The resistance R_1 represents the input resistance of the FET in parallel with a bias resistance that is usually inserted between the gate and source. C_1 is the overall input capacitance of the FET including any stray capacitance but excluding the capacitance of the pyroelectric detector. Suppose that the incident radiation intensity is constant and equal to I. Emissivity η of a surface characterizes what fraction of the incident radiation that is absorbed? ηI is the energy absorbed per unit area per unit time. Some of the absorbed energy will increase

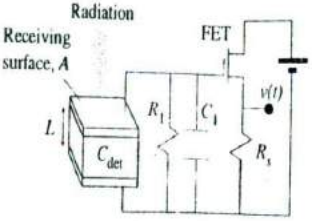


Figure 7.66 A pyroelectric detector with an FET voltage follower circuit.

the temperature of the detector and some of it will be lost to surroundings by thermal conduction and convection. Let the detector receiving area be A, thickness be L, density be ρ, and specific heat capacity (heat capacity per unit mass) be c. The heat losses will be proportional to the temperature difference between the detector temperature T and the ambient temperature T_a, as well as the surface area A (much greater than L). Energy balance requires that

Rate of increase in the internal energy (heat content) of the detector
= Rate of energy absorption − Rate of heat losses

that is,

$$(AL\rho)c\frac{dT}{dt} = A\eta I - KA(T - T_a)$$

where K is a constant of proportionality that represents the heat losses and hence depends on the thermal conductivity κ. If the heat loss involves pure thermal conduction from the detector surface to the detector base (detector mount), then $K = \kappa/L$. In practice, this is generally not the case and $K = \kappa/L$ is an oversimplification.

a. Show that the temperature of the detector rises exponentially as

Detector temperature

$$T = T_a + \frac{\eta I}{K}\left[1 - \exp\left(-\frac{t}{\tau_{th}}\right)\right]$$

where τ_{th} is a **thermal time constant** defined by $\tau_{th} = L\rho c/K$. Further show that for very small K, this equation simplifies to

$$T = T_a + \frac{\eta I}{L\rho c}t$$

b. Show that temperature change dT in time dt leads to a pyroelectric current i_p given by

Pyroelectric current

$$i_p = Ap\frac{dT}{dt} = \frac{Ap\eta I}{L\rho c}\exp\left(-\frac{t}{\tau_{th}}\right)$$

where p is the pyroelectric coefficient. What is the initial current?

c. The voltage across the FET and hence the output voltage $v(t)$ is given by

Pyroelectric detector output voltage

$$v(t) = V_o\left[\exp\left(-\frac{t}{\tau_{th}}\right) - \exp\left(-\frac{t}{\tau_{el}}\right)\right]$$

where V_o is a constant and τ_{el} is the **electrical time constant** given by R_1C_t, where C_t, total capacitance, is $(C_1 + C_{det})$, where C_{det} is the capacitance of the detector. Consider a particular PZT pyroelectric detector with an area of 1 mm^2 and a thickness of 0.05 mm. Suppose that this PZT has $\varepsilon_r = 250$, $\rho = 7.7$ g cm^{-3}, $c = 0.3$ J K^{-1} g^{-1}, and $\kappa = 1.5$ W K^{-1} m^{-1}. The detector is connected to an FET circuit that has $R_1 = 10$ MΩ and $C_1 = 3$ pF. Taking the thermal conduction loss constant K as κ/L, and $\eta = 1$, calculate τ_{th} and τ_{el}. Sketch schematically the output voltage. What is your conclusion?

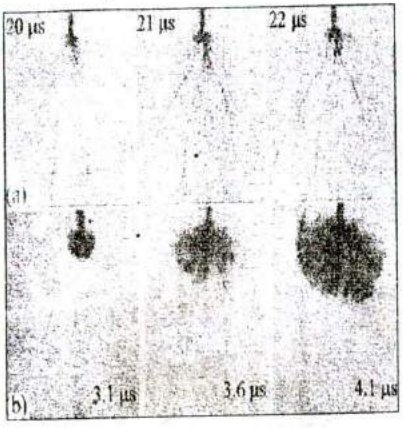
7.29 **Spark generator design** Design a PLZT piezoelectric spark generator using two back-to-back PLZT crystals that provide a 60 μJ spark in an air gap of 0.5 mm from a force of 50 N. At 1 atm in an air gap of 0.5 mm, the breakdown voltage is about 3000 V. The design will need to specify the dimensions of the crystal and the dielectric constant. Assume that the piezoelectric voltage coefficient is 0.023 V m N^{-1}.

7.30 **Ionic polarization resonance in CsCl** Consider a CsCl crystal which has the following properties. The optical dielectric constant is 2.62, the dc dielectric constant is 7.20, and the lattice parameter a is 0.412 nm. There is only one ion pair (Cs^+–Cl^-) in the cubic-type unit cell. Calculate (estimate) the ionic resonance absorption frequency and compare the value with the experimentally observed resonance at 3.1×10^{12} Hz. What effective value of Q would bring the calculated value to within 10 percent of the experimental value?

7.31 **Low-κ porous dielectrics for microelectronics** Interconnect technologies need lower ε_r interlayer dielectrics (ILDs) to minimize the interconnect capacitances. These materials are called **low-κ dielectrics.**

a. Consider fluorinated silicon dioxide, also known as fluorosilicate glass (FSG). Its ε_r is 3.2. What would be the effective dielectric constant if the ILD is 40 percent porous?

b. What should be the starting ε_r if we need an effective ε_r less than 2 and the porosity cannot exceed 40 percent?



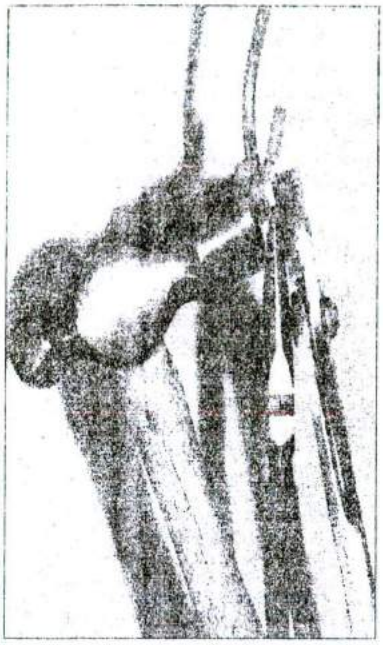
Tree and bush type electrical discharge structures. (a) Voltage V = 160 kV, gap spacing d = 0.06 m at various times. (b) Dense bush discharge structure, V = 300 kV, d = 0.06 m at various times.

SOURCE: V. Lopatin, M. D. Noskov, R. Badent, K. Kist, A. J. Swab, "Positive Discharge Development in Insulating Oil: Optical Observation and Simulation," *IEEE Trans. on Dielec. and Elec. Insulation*, vol. 5, no. 2, 1998, p. 251, figure 2. (© IEEE, 1998)

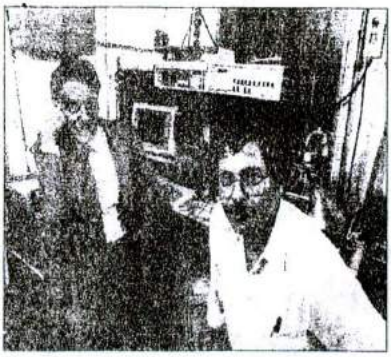


Coaxial cable connector with traces of corona discharge, electrical treeing

SOURCE: M. Mayer and G. H. Schröder, "Coaxial 30 kV Connectors for the RG220/U Cable: 20 Years of Operational Experience," *IEEE Electrical Insulation Magazine*, vol. **16**, March/April 2000, p. 11, figure 6. (© IEEE, 2000)



This small neodymium-iron-boron permanent magnet (diameter about the same as one-cent coin) is capable of lifting up to 10 pounds. Nd-Fe-B magnets typically have large $(BH)_{max}$ values (200–275 kJ m^{-3}).



In 1986 J. George Bednorz (right) and K. Alex Müller, at IBM Research Laboratories in Zurich, discovered that a copper oxide based ceramic-type compound (La-Ba-Cu-O) which normally has high resistivity becomes superconducting when cooled below 35 K. This Nobel prize winning discovery opened a new era of high-temperature-superconductivity research; now there are various ceramic compounds that are superconducting above the liquid nitrogen (an inexpensive cryogen) temperature (77 K).

SOURCE: IBM Zürich Research Laboratories.

CHAPTER

8

Magnetic Properties and Superconductivity

Many electrical engineering devices such as inductors, transformers, rotating machines, and ferrite antennas are based on utilizing the magnetic properties of materials. There are many instances where permanent magnets are also used either on their own or as part of a device such as a rotating machine or a loud speaker. The majority of engineering devices make use of the ferromagnetic and ferrimagnetic properties, which are therefore treated in much more detail than other magnetic properties such as diamagnetism and paramagnetism. Although superconductivity involves the vanishing of the resistivity of a conductor at low temperatures and is normally explained within quantum mechanics, we treat the subject in this chapter because all superconductors are perfect diamagnets and, further, they have present or potential uses that involve magnetic fields. The advent of high-T_c superconductivity, discovered in 1986 by George Bednorz and Alex Müller at IBM Research Laboratories in Zürich, is undoubtedly one of the most significant discoveries over the last 50 years, as popularized in various magazines. High-T_c superconductors are already finding applications in such devices as superconducting solenoids, sensitive magnetometers, and high-Q microwave filters.

8.1 MAGNETIZATION OF MATTER

8.1.1 Magnetic Dipole Moment

Magnetic properties of materials involve concepts based on the magnetic dipole moment. Consider a current loop, as shown in Figure 8.1, where the circulating current is I. This may, for example, be a coil carrying a current. For simplicity we will assume that the current loop lies within a single plane. The area enclosed by the current is A. Suppose that $\mathbf{u}_n$ is a unit vector coming out from the area A. The direction of $\mathbf{u}_n$ is such that

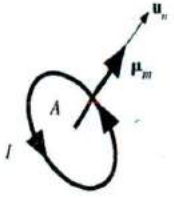


Figure 8.1 Definition of a magnetic dipole moment.

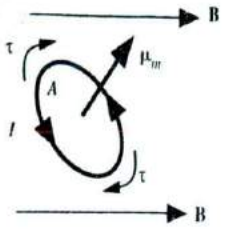


Figure 8.2 A magnetic dipole moment in an external field experiences a torque.

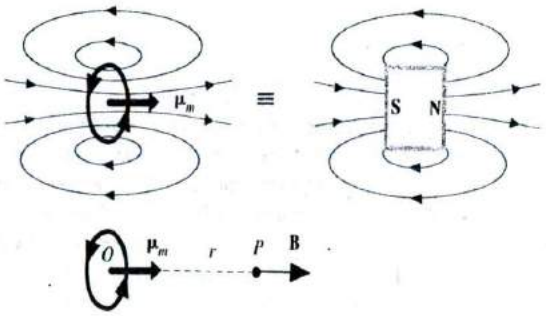


Figure 8.3 A magnetic dipole moment creates a magnetic field just like a bar magnet.
The field **B** depends on $\boldsymbol{\mu}_m$.

Definition of magnetic moment

looking along it, the current circulates clockwise. Then the **magnetic dipole moment,** or simply the **magnetic moment** $\boldsymbol{\mu}_m$, is defined by[1]

$$\boldsymbol{\mu}_m = IA\mathbf{u}_n \qquad [8.1]$$

When a magnetic moment is placed in a magnetic field, it experiences a torque that tries to rotate the magnetic moment to align its axis with the magnetic field, as depicted in Figure 8.2. Moreover, since a magnetic moment is a current loop, it gives rise to a magnetic field **B** around it, as shown in Figure 8.3, which is similar to the magnetic field around a bar magnet. We can find the field **B** from the current I and its geometry, which are treated in various physics textbooks. For example, the field **B** at a point P at a distance r along the axis of the coil from the center, as shown in Figure 8.3, is directly proportional to the magnitude of the magnetic moment but inversely proportional to r^3, that is, $B \propto \boldsymbol{\mu}_m / r^3$.

[1] The symbol $\boldsymbol{\mu}$ for the magnetic dipole moment should not be confused with the permeability. Absolute and relative permeabilities will be denoted by μ_0 and μ_r.

8.1.2 ATOMIC MAGNETIC MOMENTS

An orbiting electron in an atom behaves much like a current loop and has a magnetic dipole moment associated with it, called the **orbital magnetic moment** (μ_{orb}), as illustrated in Figure 8.4. If ω is the angular frequency of the electron, then the current I due to the orbiting electron is

$$I = \text{Charge flowing per unit time} = -\frac{e}{\text{Period}} = -\frac{e\omega}{2\pi}$$

If r is the radius of the orbit, then the magnetic dipole moment is

$$\mu_{orb} = I(\pi r^2) = -\frac{e\omega r^2}{2}$$

But the velocity v of the electron is ωr and its orbital angular momentum is

$$L = (m_e v)r = m_e \omega r^2$$

Using this in μ_{orb}, we get

$$\mu_{orb} = -\frac{e}{2m_e}L \qquad [8.2]$$

Orbital magnetic moment of the electron

We see that the magnetic moment is proportional to the orbital angular momentum through a factor that has the charge to mass ratio of the electron. The numerical factor, in this case $e/2m_e$, relating the angular momentum to the magnetic moment, is called the **gyromagnetic ratio.** The negative sign in Equation 8.2 indicates that μ_{orb} is in the opposite direction to L and is due to the negative charge of the electron.

The electron also has an intrinsic angular momentum S, that is, spin. The spin of the electron has a **spin magnetic moment,** denoted by μ_{spin}, but the relationship between μ_{spin} and S is not the same as that in Equation 8.2. The gyromagnetic ratio is a factor of 2 greater,

$$\mu_{spin} = -\frac{e}{m_e}S \qquad [8.3]$$

Spin magnetic moment of the electron

The overall magnetic moment of the electron consists of $\boldsymbol{\mu}_{orb}$ and $\boldsymbol{\mu}_{spin}$ appropriately added. We cannot simply add them numerically as they are vector quantities. Furthermore, the overall magnetic moment $\boldsymbol{\mu}_{atom}$ of the atom itself depends on the

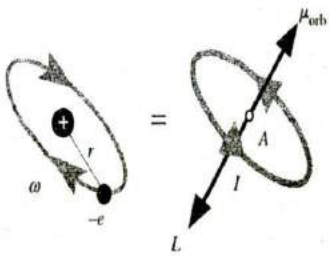


Figure 8.4 An orbiting electron is equivalent to a magnetic dipole moment μ_{orb}.

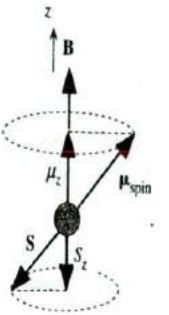


Figure 8.5 The spin magnetic moment precesses about an external magnetic field along z and has a value μ_z along z.

orbital motions and spins of *all* the electrons. Electrons in closed subshells, however, do not contribute to the overall magnetic moment because for every electron with a given **L** (or **S**), there is another one with an opposite **L** (or **S**). The reason is that the direction of **L** is space quantized by m_ℓ and all negative and positive values of m_ℓ are occupied in a closed shell. Similarly, there are as many electrons spinning up as there are spinning down, so there is no net electron spin in a closed shell and no net $\boldsymbol{\mu}_{spin}$. Thus, only **unfilled subshells** contribute to the overall magnetic moment of an atom.

Consider an atom that has closed inner shells and a single electron in an *s* orbital ($\ell = 0$). This means that the orbital magnetic moment is zero and the atom has a magnetic moment due to the spin of the electron alone, $\boldsymbol{\mu}_{atom} = \boldsymbol{\mu}_{spin}$. In the presence of an external magnetic field along the z direction, the magnetic moment cannot simply rotate and align with the field because quantum mechanics requires the spin angular momentum to be space quantized, that is, S_z (the component of **S** along z) must be $m_s\hbar$ where $m_s = \pm\frac{1}{2}$ is the spin magnetic quantum number. The torque experienced by the spinning electron causes the spin magnetic moment to precess about the external magnetic field, as shown in Figure 8.5. This precession is such that $S_z = -\frac{1}{2}\hbar$ and leads to an average magnetic moment μ_z along the field given by Equation 8.3 with S_z, that is,

Magnetic moment along the field

$$\mu_z = -\frac{e}{m_e}S_z = -\frac{e}{m_e}(m_s\hbar) = \frac{e\hbar}{2m_e} = \beta \qquad [8.4]$$

The quantity $\beta = e\hbar/2m_e$ is called the **Bohr magneton** and has the value 9.27×10^{-24} A m^2 or J T^{-1}.

Thus, the spin of a single electron has a magnetic moment of one Bohr magneton along the field.

8.1.3 MAGNETIZATION VECTOR M

Consider a tightly wound long solenoid, ideally infinitely long, with free space (or vacuum) as the medium inside the solenoid, as shown in Figure 8.6a. The magnetic field inside the solenoid is denoted by $\mathbf{B}_o$ to specifically identify this field as in free space.

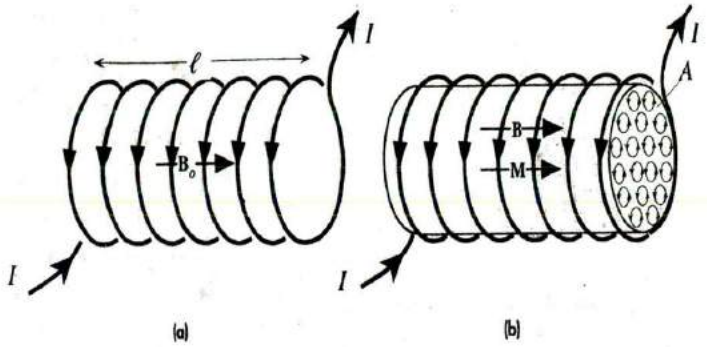


Figure 8.6
(a) Consider a long solenoid. With free space as the medium inside, the magnetic field is $\mathbf{B}_o$.
(b) A material medium inserted into the solenoid develops a magnetization **M**.

This field depends on the current I through the solenoid wire and the number of turns per unit length n and is given by[2]

$$B_o = \mu_o n I = \mu_o I' \qquad [8.5]$$

Free space field inside solenoid

where I' is the current per unit length of the solenoid, that is, $I' = nI$, and μ_o is the absolute permeability of free space in henries per meter, H m^{-1}.

If we now place a cylindrical material medium to fill the inside of this solenoid, as in Figure 8.6b, we find that the magnetic field has changed. The new magnetic field in the presence of a medium is denoted as **B**. We will take $\mathbf{B}_o$ to be the applied magnetic field into which the material medium is placed.

Each atom of the material responds to the applied field $\mathbf{B}_o$ and develops, or acquires, a net magnetic moment $\boldsymbol{\mu}_m$ along the applied field. We can view each magnetic moment $\boldsymbol{\mu}_m$ as the result of the precession of each atomic magnetic moment about $\mathbf{B}_o$. The medium therefore develops a net magnetic moment along the field and becomes **magnetized.** The magnetic vector **M** describes the extent of magnetization of the medium. **M** is defined as the **magnetic dipole moment per unit volume.** Suppose that there are N atoms in a small volume ΔV and each atom i has a magnetic moment $\boldsymbol{\mu}_{mi}$ (where $i = 1$ to N). Then **M** is defined by

$$\mathbf{M} = \frac{1}{\Delta V}\sum_{i=1}^{N} \boldsymbol{\mu}_{mi} = n_{at}\boldsymbol{\mu}_{av} \qquad [8.6]$$

Magnetization vector

where n_{at} is the number of atoms per unit volume and $\boldsymbol{\mu}_{av}$ is the average magnetic moment per atom. We can assume that each atom acquires a magnetic moment $\boldsymbol{\mu}_{av}$ along $\mathbf{B}_o$. Each of these magnetic moments along $\mathbf{B}_o$ can be viewed as an elementary current loop at the atomic scale, as schematically depicted in Figure 8.6b. These elementary

[2] The proof of this comes out from Ampere's law and can be found in any text on electromagnetism.

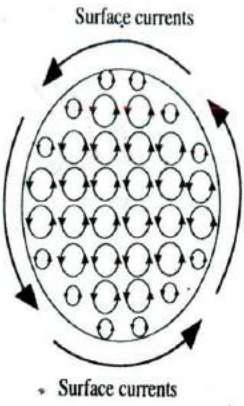


Figure 8.7 Elementary current loops result in surface currents.
There is no internal current, as adjacent currents on neighboring loops are in opposite directions.

current loops are due to electronic currents within the atom and arise from both orbital and spin motions of the electrons. Each current loop has its current plane normal to $\mathbf{B}_0$.

Consider a cross section of the magnetized medium, as in Figure 8.7. All the elementary current loops in this plane have the current circulation in the same direction inasmuch as each atom acquires the same magnetic moment $\boldsymbol{\mu}_{av}$. All neighboring loops in the bulk have adjacent currents in opposite directions that cancel each other, as apparent in Figure 8.7. Thus, there are no net bulk currents, or internal currents, within the bulk of the material. However, the currents at the surface in the surface loops cannot be canceled and this leads to a net **surface current,** as depicted in Figure 8.7. The surface currents are induced by the magnetization of the medium by the applied magnetic field and therefore depend on the magnetization M of the specimen.

From the definition of M, the total magnetic moment of the cylindrical specimen is

$$\text{Total magnetic moment} = M\,(\text{Volume}) = MA\ell$$

Suppose that the magnetization current on the surface per unit length of the specimen is I_m. Then the total circulating surface current is $I_m\ell$ and the total magnetic moment of the specimen, by definition, is

$$\text{Total magnetic moment} = (\text{Total current}) \times (\text{Cross-sectional area}) = I_m\ell A$$

Magnetization and surface currents

Equating the two total magnetic moments, we find

$$M = I_m \qquad [8.7]$$

We derived this for a particular sample geometry, a cylindrical specimen, in which M is along the axis of the cylindrical specimen and I_m flows in a plane perpendicular to M. The relationship, however, is more general, as derived in more advanced texts. It should be emphasized that the magnetization current I_m is not due to the flow of free charge carriers, as in a current-carrying copper wire, but due to localized electronic currents within the atoms of the solid at the surface. Equation 8.7 states that we can represent the magnetization of a medium by a surface current per unit length I_m that is equal to M.

8.1.4 MAGNETIZING FIELD OR MAGNETIC FIELD INTENSITY H

The magnetized specimen in Figure 8.6b placed inside the solenoid develops magnetization currents on the surface. It therefore behaves like a solenoid. We can now regard the solenoid with medium inside, as depicted in Figure 8.8. The magnetic field within the medium now arises from not only the conduction current per unit length I' in the solenoid wires but also from the magnetization current I_m on the surface. The magnetic field B inside the solenoid is now given by the usual solenoid expression but with a current that includes both I' and I_m, as shown in Figure 8.8:

$$B = \mu_o(I' + I_m) = B_o + \mu_o M$$

This relationship is generally valid and can be written in vector form as

Magnetic field in a magnetized medium

$$\mathbf{B} = \mathbf{B}_o + \mu_o \mathbf{M} \qquad [8.8]$$

The field at a point inside a magnetized material is the sum of the applied field $\mathbf{B}_o$ and a contribution from the magnetization $\mathbf{M}$ of the material. The magnetization arises from the application of $\mathbf{B}_o$ due to the current of free carriers in the solenoid wires, called the **conduction current,** which we can externally adjust. It becomes useful to introduce a vector field that represents the effect of the external or conduction current alone. In general, $\mathbf{B} - \mu_o\mathbf{M}$ at a point is the contribution of the external currents alone to the magnetic field at that point inside the material that we called $\mathbf{B}_o$. $\mathbf{B} - \mu_o\mathbf{M}$ represents a magnetizing field because it is the field of the external currents that magnetize the material. The **magnetizing field H** is defined as

Definition of the magnetizing field

$$\mathbf{H} = \frac{1}{\mu_o}\mathbf{B} - \mathbf{M} \qquad [8.9]$$

or

$$\mathbf{H} = \frac{1}{\mu_o}\mathbf{B}_o$$

Definition of the magnetizing field

The magnetizing field is also known as the **magnetic field intensity** and is measured in A m^{-1}. The reason for the division by μ_o is that the resulting vector field $\mathbf{H}$ becomes simply related to the external conduction currents (through Ampere's law). Since in the solenoid $\mathbf{B}_o$ is $\mu_o nI$, we see that the magnetizing field in a solenoid is

$$H = nI = \text{Total conduction current per unit length} \qquad [8.10]$$

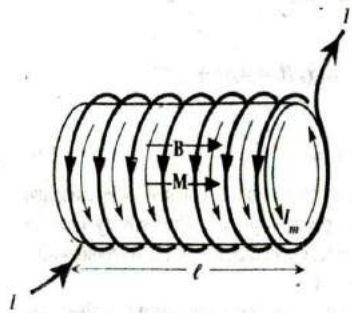


Figure 8.8 The field **B** in the material inside the solenoid is due to the conduction current I through the wires and the magnetization current I_m on the surface of the magnetized medium, or $\mathbf{B} = \mathbf{B}_o + \mu_o\mathbf{M}$.

It is generally helpful to imagine **H** as the *cause* and **B** as the *effect*. The cause **H** depends only on the external conduction currents, whereas the effect **B** depends on the magnetization **M** of matter.

8.1.5 MAGNETIC PERMEABILITY AND MAGNETIC SUSCEPTIBILITY

Suppose that at a point P in a material, the magnetic field is **B** and the magnetizing field is **H**. We let $\mathbf{B}_o$ be the magnetic field at P in the absence of any material (*i.e.*, in free space). The magnetic permeability of the medium at P is defined as the magnetic field per unit magnetizing field,

Definition of magnetic permeability

$$\mu = \frac{B}{H} \tag{8.11}$$

It relates the effect B to the cause H at the same point P inside a material. In simple qualitative terms, μ represents to what extent a medium is permeable by magnetic fields. Relative permeability μ_r of a medium is the fractional increase in the magnetic field with respect to the field in free space when a material medium is introduced. For example, suppose that the field in a solenoid with free space in it is B_o but with material inserted it is B. Then μ_r is defined by

Definition of relative permeability

$$\mu_r = \frac{B}{B_o} = \frac{B}{\mu_o H} \tag{8.12}$$

From Equations 8.11 and 8.12, clearly,

$$\mu = \mu_o \mu_r$$

The magnetization **M** produced in a material depends on the net magnetic field **B**. It would be natural to proceed as in dielectrics by relating **M** to **B** analogously to relating P (polarization) to $\mathcal{E}$ (electric field). However, for historic reasons, **M** is related to **H**, the magnetizing field. Suppose that the medium is isotropic (same properties in all directions), then magnetic susceptibility χ_m of the medium is defined simply by

Definition of magnetic susceptibility

$$\mathbf{M} = \chi_m \mathbf{H} \tag{8.13}$$

This relationship is not obeyed by all magnetic materials. For example, as we will see later, ferromagnetic materials do not obey Equation 8.12. Since the magnetic field

$$\mathbf{B} = \mu_o(\mathbf{H} + \mathbf{M})$$

we have

$$B = \mu_o H + \mu_o M = \mu_o H + \mu_o \chi_m H = \mu_o(1 + \chi_m)H$$

Relative permeability and susceptibility

and

$$\mu_r = 1 + \chi_m \tag{8.14}$$

The presence of a magnetizable material is conveniently accounted for by using the relative permeability μ_r, or $(1 + \chi_m)$, to simply multiply μ_o. Alternatively, one can simply replace μ_o with $\mu = \mu_o \mu_r$. For example, the inductance of the solenoid with a magnetic medium inside increases by a factor of μ_r.

Table 8.1 provides a summary of various important magnetic quantities, their definitions, and units.

Table 8.1 Magnetic quantities and their units

Magnetic Quantity	Symbol	Definition	Units	Comment
Magnetic field; magnetic induction	**B**	$\mathbf{F} = q\mathbf{v} \times \mathbf{B}$	T = tesla = webers m^{-2}	Produced by moving charges or currents, acts on moving charges or currents.
Magnetic flux	Φ	$\Delta\Phi = B_{\text{normal}}\,\Delta A$	Wb = weber	$\Delta\Phi$ is flux through ΔA and B_{normal} is normal to ΔA. Total flux through any closed surface is zero.
Magnetic dipole moment	$\boldsymbol{\mu}_m$	$\mu_m = IA$	$A\,m^2$	Experiences a torque in **B** and a net force in a nonuniform **B**.
Bohr magneton	β	$\beta = e\hbar/2m_e$	$A\,m^2$ or $J\,T^{-1}$	Magnetic moment due to the spin of the electron. $\beta = 9.27 \times 10^{-24}\,A\,m^2$
Magnetization vector	**M**	Magnetic moment per unit volume	$A\,m^{-1}$	Net magnetic moment in a material per unit volume.
Magnetizing field; magnetic field intensity	**H**	$\mathbf{H} = \mathbf{B}/\mu_o - \mathbf{M}$	$A\,m^{-1}$	**H** is due to external conduction currents only and is the cause of **B** in a material.
Magnetic susceptibility	χ_m	$\mathbf{M} = \chi_m \mathbf{H}$	None	Relates the magnetization of a material to the magnetizing field **H**.
Absolute permeability	μ_o	$c = [\varepsilon_o \mu_o]^{-1/2}$	$H\,m^{-1} = Wb\,m^{-1}\,A^{-1}$	A fundamental constant in magnetism. In free space $\mu_o = B/H$.
Relative permeability	μ_r	$\mu_r = B/\mu_o H$	None	
Magnetic permeability	μ	$\mu = \mu_o \mu_r$	$H\,m^{-1}$	Not to be confused with magnetic moment.
Inductance	L	$L = \Phi_{\text{total}}/I$	H (henries)	Total flux threaded per unit current.
Magnetostatic energy density	E_{vol}	$dE_{\text{vol}} = H\,dB$	$J\,m^{-3}$	dE_{vol} is the energy required per unit volume in changing B by dB.

EXAMPLE 8.1

AMPERE'S LAW AND THE INDUCTANCE OF A TOROIDAL COIL Ampere's law provides a relationship between the conduction current I and the magnetic field intensity H threading this current. The conduction current I is the current due to the flow of free charge carriers through a conductor and not due to the magnetization of any medium. Consider an arbitrary closed path C around a conductor carrying a current I, as shown in Figure 8.9. The tangential component of **H** to the curve C at point P is H_t. If dl is an infinitesimally small path length of C at P, as shown in Figure 8.9, then the summation of $H_t\,dl$ around the path C gives the conduction current enclosed within C. This is **Ampere's law,**

$$\oint_C H_t\,dl = I \qquad [8.15]$$ *Ampere's law*

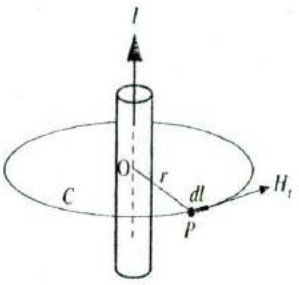


Figure 8.9 Ampere's circuital law.

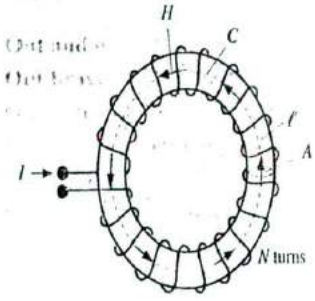


Figure 8.10 A toroidal coil with N turns.

Consider the toroidal coil with N turns shown in Figure 8.10. First assume that the toroid core is air ($\mu_r \approx 1$). Suppose that the current through the coils is I. By symmetry, the magnetic field intensity H inside the toroidal core is the same everywhere and is directed along the circumference. Suppose that ℓ is the length of the mean circumference C. The current is linked N times by the circumference C, so Equation 8.15 is

$$\oint_C H_l \, dl = H\ell = NI$$

or

$$H = \frac{NI}{\ell}$$

The magnetic field B_a with air as core material is then simply

$$B_a = \mu_a H = \frac{\mu_a NI}{\ell}$$

When the toroidal coil has a magnetic medium with a relative permeability μ_r, the magnetic field intensity is still H because the conduction current I has not changed. But the magnetic field B is now different than B_a and is given by

Magnetic field inside toroidal coil

$$B = \mu_a \mu_r H = \frac{\mu_a \mu_r NI}{\ell}$$

If A is the cross-sectional area of the toroid, then the total flux Φ through the core is BA or $\mu_a \mu_r NAI/\ell$. The current I in Figure 8.10 threads the flux N times. The inductance L of the toroidal coil, by definition, is then

Inductance of toroidal coil

$$L = \frac{\text{Total flux threaded}}{\text{Current}} = \frac{N\Phi}{I} = \frac{\mu_a \mu_r N^2 A}{\ell}$$

Having a magnetic material as the toroid core increases the inductance by a factor of μ_r in the same way a dielectric material increases the capacitance by a factor of ε_r.

EXAMPLE 8.6 **MAGNETOSTATIC ENERGY PER UNIT VOLUME** Consider a toroidal coil with N turns that is energized from a voltage supply through a rheostat, as shown in Figure 8.11. The core of the toroid may be any material. Suppose that by adjusting the rheostat we increase the current i

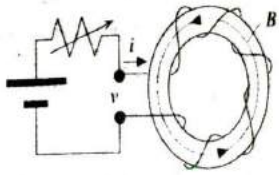


Figure 8.11 Energy required to magnetize a toroidal coil.

supplied to the coil. The current i produces magnetic flux Φ in the core, which is BA, where B is the magnetic field and A is the cross-sectional area. We can now use Ampere's law for H to relate the current i to H, as in Example 8.1. If ℓ is the mean circumference, then

$$H\ell = Ni \tag{8.16}$$

The changing current means that the flux is also changing (both increasing). We know from Faraday's law that a changing flux that threads a circuit generates a voltage v in that circuit given by the rate of change of total threaded flux, or $N\Phi$. Lenz's law makes the polarity of the induced voltage oppose the applied voltage. Suppose that in a time interval δt seconds, the magnetic field within the core changes by δB; then $\delta\Phi = A\delta B$ and

$$v = \frac{\delta(\text{Total flux threaded})}{\delta t} = \frac{N\delta\Phi}{\delta t} = NA\frac{\delta B}{\delta t} \tag{8.17}$$

The battery has to supply the current i against this induced voltage v, which means that it has to do electrical work iv every second. In other words, the battery has to do work $iv\,\delta t$ in a time interval δt to supply the necessary current to increase the magnetic field by δB. The electric energy δE that is input into the coil in time δt is then, using Equations 8.16 and 8.17,

$$\delta E = iv\,\delta t = \left(\frac{H\ell}{N}\right)\left(NA\frac{\delta B}{\delta t}\right)\delta t = (A\ell)H\,\delta B$$

This energy δE is the work done in increasing the field in the core by δB. The volume of the toroid is $A\ell$. Therefore, the total energy or work required per unit volume to increase the magnetic field from an initial value B_1 to a final value B_2 in the toroid is

Work done per unit volume during magnetization

$$E_{\text{vol}} = \int_{B_1}^{B_2} H\,dB \tag{8.18}$$

where the integration limits are determined by the initial and final magnetic field. This is the expression for calculating the **energy density** (energy per unit volume) required to change the field from B_1 to B_2. It should be emphasized that Equation 8.18 is valid for *any medium*. We conclude that an incremental energy density of $dE_{\text{vol}} = H\,dB$ is required to increase the magnetic field by dB at a point in any medium including free space.

We can now consider a core material that we can represent by a *constant* relative permeability μ_r. This means we can exclude those materials that do not have a linear relationship between B and H, such as ferromagnetic and ferrimagnetic materials, which we will discuss later. If the core is free space or air, then $\mu_r = 1$.

Suppose that we increase the current in Figure 8.11 from zero to some final value I so that the magnetic field changes from zero to some final value B. Since the medium has a constant relative permeability μ_r, we can write

$$B = \mu_r\mu_o H$$

and use this in Equation 8.18 to integrate and find the energy per unit volume needed to establish the field B or field intensity H

Energy density of a magnetic field

$$E_{vol} = \frac{1}{2}\mu_r\mu_o H^2 = \frac{B^2}{2\mu_r\mu_o} \qquad [8.19]$$

This is the energy absorbed from the battery per unit volume of core medium to establish the magnetic field. This energy is stored in the magnetic field and is called **magnetostatic energy density.** It is a form of magnetic potential energy. If we were to suddenly remove the battery and short those terminals, the current will continue to flow for a short while (determined by L/R) and do external work in heating the resistor. This external work comes from the stored energy in the magnetic field. If the medium is free space, or air, then the energy density is

Magnetostatic energy density in free space

$$E_{vol}(\text{air}) = \frac{1}{2}\mu_o H^2 = \frac{B^2}{2\mu_o}$$

A magnetic field of 2 T corresponds to a magnetostatic energy density of 1.6 MJ m^{-3} or 1.6 J cm^{-3}. The energy in a magnetic field of 2 T in a 1 cm^3 volume (size of a thimble) has the work ability (potential energy) to raise an average-sized apple by 5 feet. We should note that as long as the core material is linear, that is, μ_r is independent of the magnetic field itself, magnetostatic energy density can also be written as

Magnetostatic energy in a linear magnetic medium

$$E_{vol} = \frac{1}{2}HB \qquad [8.20]$$

8.2 MAGNETIC MATERIAL CLASSIFICATIONS

In general, magnetic materials are classified into five distinct groups: diamagnetic, paramagnetic, ferromagnetic, antiferromagnetic, and ferrimagnetic. Table 8.2 provides a summary of the magnetic properties of these classes of materials.

8.2.1 DIAMAGNETISM

Typical diamagnetic materials have a magnetic susceptibility that is negative and small. For example, the silicon crystal is diamagnetic with $\chi_m = -5.2 \times 10^{-6}$. The relative permeability of diamagnetic materials is slightly less than unity. When a diamagnetic substance such as a silicon crystal is placed in a magnetic field, the magnetization vector **M** in the material is in the *opposite* direction to the applied field $\mu_o\mathbf{H}$ and the resulting field **B** within the material is less than $\mu_o\mathbf{H}$. The negative susceptibility can be interpreted as the diamagnetic substance trying to expel the applied field from the material. When a diamagnetic specimen is placed in a nonuniform magnetic field, the magnetization **M** of the material is in the opposite direction to **B** and the specimen experiences a net force toward smaller fields, as depicted in Figure 8.12. A substance exhibits diamagnetism whenever the constituent atoms in the material have closed subshells and shells. This means that each constituent atom has no permanent magnetic moment in the absence of an applied field. Covalent

Table 8.2 Classification of magnetic materials

Type	χ_m (typical values)	χ_m versus T	Comments and Examples
Diamagnetic	Negative and small (-10^{-6})	T independent	Atoms of the material have closed shells. Organic materials, *e.g.*, many polymers; covalent solids, *e.g.*, Si, Ge, diamond; some ionic solids, *e.g.*, alkalihalides; some metals, *e.g.*, Cu, Ag, Au.
	Negative and large (-1)	Below a critical temperature	Superconductors
Paramagnetic	Positive and small (10^{-5}–10^{-4})	Independent of T	Due to the alignment of spins of conduction electrons. Alkali and transition metals.
	Positive and small (10^{-5})	Curie or Curie–Weiss law, $\chi_m = C/(T - T_C)$	Materials in which the constituent atoms have a permanent magnetic moment, *e.g.*, gaseous and liquid oxygen; ferromagnets (Fe), antiferromagnets (Cr), and ferrimagnets (Fe_3O_4) at high temperatures.
Ferromagnetic	Positive and very large	Ferromagnetic below and paramagnetic above the Curie temperature	May possess a large permanent magnetization even in the absence of an applied field. Some transition and rare earth metals, Fe, Co, Ni, Gd, Dy.
Antiferromagnetic	Positive and small	Antiferromagnetic below and paramagnetic above the Néel temperature	Mainly salts and oxides of transition metals, *e.g.*, MnO, NiO, MnF_2, and some transition metals, α–Cr, Mn.
Ferrimagnetic	Positive and very large	Ferrimagnetic below and paramagnetic above the Curie temperature	May possess a large permanent magnetization even in the absence of an applied field. Ferrites.

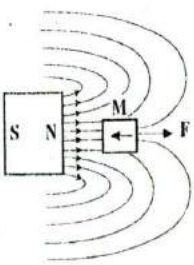


Figure 8.12 A diamagnetic material placed in a nonuniform magnetic field experiences a force toward smaller fields. This repels the diamagnetic material away from a permanent magnet.

crystals and many ionic crystals are typical diamagnetic materials because the constituent atoms have no unfilled subshells. Superconductors, as we will discuss later, are perfect diamagnets with $\chi_m = -1$ and totally expel the applied field from the material.

8.2.2 PARAMAGNETISM

Paramagnetic materials have a small positive magnetic susceptibility. For example, oxygen gas is paramagnetic with $\chi_m = 2.1 \times 10^{-6}$ at atmospheric pressure and room temperature. Each oxygen molecule has a net magnetic dipole moment $\boldsymbol{\mu}_{mol}$. In the absence of an applied field, these molecular moments are randomly oriented due to the random collisions of the molecules, as depicted in Figure 8.13a. The magnetization of the gas is zero. In the presence of an applied field, the molecular magnetic moments take various alignments with the field, as illustrated in Figure 8.13b. The degree of alignment of $\boldsymbol{\mu}_{mol}$ with the applied field and hence magnetization **M** increases with the strength of the applied field $\mu_0\mathbf{H}$. Magnetization M typically decreases with increasing temperature because at higher temperatures there are more molecular collisions, which destroy the alignments of molecular magnetic moments with the applied field. When a paramagnetic substance is placed in a nonuniform magnetic field, the induced magnetization **M** is along **B** and there is a net force toward greater fields. For example, when liquid oxygen is poured close to a strong magnet, as depicted in Figure 8.14, the liquid becomes attracted to the magnet.

Many metals are also paramagnetic, such as magnesium with $\chi_m = 1.2 \times 10^{-5}$. The origin of paramagnetism (called **Pauli spin paramagnetism**) in these metals is due to the alignment of the majority of spins of conduction electrons with the field.

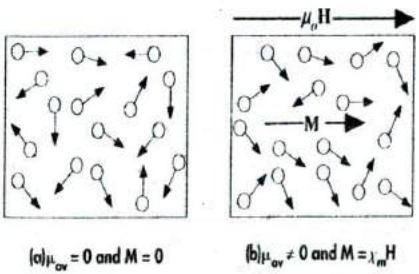


Figure 8.13
(a) In a paramagnetic material, each individual atom possesses a permanent magnetic moment, but due to thermal agitation there is no average moment per atom and **M** = 0.
(b) In the presence of an applied field, individual magnetic moments take alignments along the applied field and **M** is finite and along **B**.

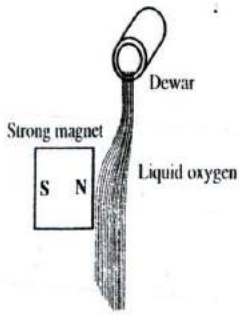


Figure 8.14 A paramagnetic material placed in a nonuniform magnetic field experiences a force toward greater fields.
This attracts the paramagnetic material (e.g., liquid oxygen) toward a permanent magnet.

8.2.3 FERROMAGNETISM

Ferromagnetic materials such as iron can possess large permanent magnetizations even in the absence of an applied magnetic field. The magnetic susceptibility χ_m is typically positive and very large (even infinite) and, further, depends on the applied field intensity. The relationship between the magnetization **M** and the applied magnetic field μ_o**H** is highly nonlinear. At sufficiently high fields, the magnetization **M** of the ferromagnet saturates. The origin of ferromagnetism is the quantum mechanical exchange interaction (discussed later) between the constituent atoms that results in regions of the material possessing permanent magnetization. Figure 8.15 depicts a region of the Fe crystal, called a **magnetic domain,** that has a net magnetization vector **M** due to the alignment of the magnetic moments of all Fe atoms in this region. This crystal domain has **magnetic ordering** as all the atomic magnetic moments have been aligned parallel to each other. Ferromagnetism occurs below a critical temperature called the Curie temperature T_C. At temperatures above T_C, ferromagnetism is lost and the material becomes paramagnetic.

8.2.4 ANTIFERROMAGNETISM

Antiferromagnetic materials such as chromium have a small but positive susceptibility. They cannot possess any magnetization in the absence of an applied field, in contrast to ferromagnets. Antiferromagnetic materials possess a magnetic ordering in which the magnetic moments of alternating atoms in the crystals align in opposite directions, as schematically depicted in Figure 8.16. The opposite alignments of atomic magnetic moments are due to quantum mechanical exchange forces (described later in Section 8.3). The net result is that in the absence of an applied field, there is no net magnetization. Antiferromagnetism occurs below a critical temperature called the **Néel temperature** T_N. Above T_N, antiferromagnetic material becomes paramagnetic.

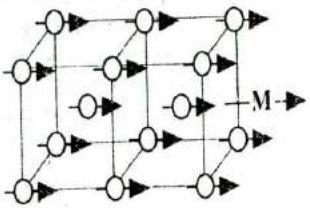


Figure 8.15 In a magnetized region of a ferromagnetic material such as iron, all the magnetic moments are spontaneously aligned in the same direction.
There is a strong magnetization vector **M** even in the absence of an applied field.

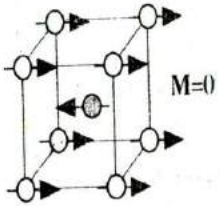


Figure 8.16 In this antiferromagnetic BCC crystal (Cr), the magnetic moment of the center atom is canceled by the magnetic moments of the corner atoms (one-eighth of the corner atom belongs to the unit cell).

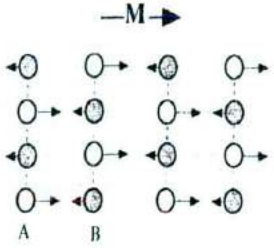


Figure 8.17 Illustration of magnetic ordering in the ferrimagnetic crystal.
All A atoms have their spins aligned in one direction and all B atoms have their spins aligned in the opposite direction. As the magnetic moment of an A atom is greater than that of a B atom, there is net magnetization **M** in the crystal.

8.2.5 FERRIMAGNETISM

Ferrimagnetic materials such as ferrites (*e.g.*, Fe_3O_4) exhibit magnetic behavior similar to ferromagnetism below a critical temperature called the Curie temperature T_C. Above T_C they become paramagnetic. The origin of ferrimagnetism is based on magnetic ordering, as schematically illustrated in Figure 8.17. All A atoms have their spins aligned in one direction and all B atoms have their spins aligned in the opposite direction. As the magnetic moment of an A atom is greater than that of a B atom, there is net magnetization **M** in the crystal. Unlike the antiferromagnetic case, the oppositely directed magnetic moments have different magnitudes and do not cancel. The net effect is that the crystal can possess magnetization even in the absence of an applied field. Since ferrimagnetic materials are typically nonconducting and therefore do not suffer from eddy current losses, they are widely used in high-frequency electronics applications.

All useful magnetic materials in electrical engineering are invariably ferromagnetic or ferrimagnetic.

8.3 FERROMAGNETISM ORIGIN AND THE EXCHANGE INTERACTION

The transition metals iron, cobalt, and nickel are all ferromagnetic at room temperature. The rare earth metals gadolinium and dysprosium are ferromagnetic below room temperature. Ferromagnetic materials can exhibit permanent magnetization even in the absence of an applied field; that is, they possess a susceptibility that is infinite.

In a magnetized iron crystal, all the atomic magnetic moments are aligned in the same direction, as illustrated in Figure 8.15, where the moments in this case have all been aligned along the [100] direction, which gives net magnetization along this direction. It may be thought that the reason for the alignment of the moments is the magnetic forces between the moments, just as bar magnets will tend to align head to tail in an SNSN . . . fashion. This is not, however, the cause, as the magnetic potential energy of interaction is small, indeed smaller than the thermal energy.

The iron atom has the electron structure $[Ar]3d^64s^2$. An isolated iron atom has only the $3d$ subshell with four of the five orbitals unfilled. By virtue of Hund's rule, the electrons try to align their spins so that the five $3d$ orbitals contain two paired electrons

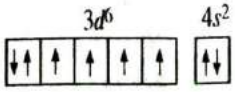


Figure 8.18 The isolated Fe atom has four unpaired spins and a spin magnetic moment of 4β.

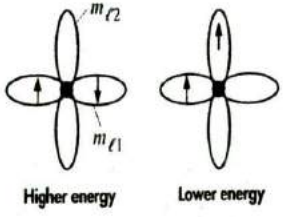


Figure 8.19 Hund's rule for an atom with many electrons is based on the exchange interaction.

and four unpaired electrons, as in Figure 8.18. The isolated atom has four parallel electron spins and hence a spin magnetic moment of 4β.

The origin of Hund's rule, visualized in Figure 8.19, lies in the fact that when the spins are parallel (same m_s), as a requirement of the Pauli exclusion principle, the electrons must occupy orbitals with different m_ℓ and hence possess different spatial distributions (recall that m_ℓ determines the orientation of an orbit). Different m_ℓ values result in a smaller Coulombic repulsion energy between the electrons compared with the case where the electrons have opposite spins (different m_s), where they would be in the same orbital (same m_ℓ), that is, in the same spatial region. It is apparent that even though the interaction energy between the electrons has nothing to do with magnetic forces, it does depend nonetheless on the orientations of their spins (m_s), or on their spin magnetic moments, and it is less when the spins are parallel. Two electrons parallel their spins not because of the direct magnetic interaction between the spin magnetic moments but because of the **Pauli exclusion principle** and the **electrostatic interaction energy.** Together they constitute what is known as an **exchange interaction,** which forces two electrons to take m_s and m_ℓ values that result in the minimum of electrostatic energy. In an atom, the exchange interaction therefore forces two electrons to take the same m_s but different m_ℓ if this can be done within the Pauli exclusion principle. This is the reason an isolated Fe atom has four unpaired spins in the $3d$ subshell.

In the crystal, of course, the outer electrons are no longer strictly confined to their parent Fe atoms, particularly the $4s$ electrons. The electrons now have wavefunctions that belong to the whole solid. Something like Hund's rule also operates at the crystal level for Fe, Co, and Ni. If two $3d$ electrons parallel their spins and occupy different wavefunctions (and hence different negative charge distributions), the resulting mutual Coulombic repulsion between them and also with all the other electrons and the attraction to the positive Fe ions result in an overall reduction of potential energy. This reduction in energy is again due to the exchange interaction and is a direct consequence of the Pauli exclusion principle and the Coulombic forces. Thus, the majority of $3d$ electrons spontaneously parallel their spins without the need for the application of an external magnetic field. The number of electrons that actually parallel their spins depends on the strength of the exchange interaction, and for the iron crystal this turns out to be about 2.2 electrons per atom. Since typically the wavefunctions

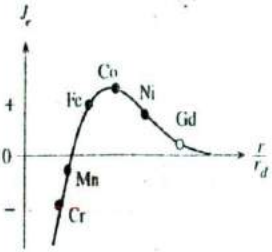


Figure 8.20 The exchange integral as a function of r/r_d, where r is the interatomic distance and r_d the radius of the d orbit (or the average d subshell radius). Cr to Ni are transition metals. For Gd, the x axis is r/r_f, where r_f is the radius of the f orbit.

of the 3*d* electrons in the whole iron crystal show localization around the iron ions, some people prefer to view the 3*d* electrons as spending the majority of their time around Fe atoms, which explains the reason for drawing the magnetized iron crystal as in Figure 8.15.

It may be thought that all solids should follow the example of Fe and become spontaneously ferromagnetic since paralleling spins would result in different spatial distributions of negative charge and probably a reduction in the electrostatic energy, but this is not generally the case at all. We know that, in the case of covalent bonding, the electrons have the lowest energy when the two electrons spin in opposite directions. In covalent bonding in molecules, the exchange interaction does not reduce the energy. Making the electron spins parallel leads to spatial negative charge distributions that result in a net mutual electrostatic repulsion between the positive nuclei.

In the simplest case, for two atoms only, the exchange energy depends on the interatomic separation between two interacting atoms and the relative spins of the two outer electrons (labeled as 1 and 2). From quantum mechanics, the exchange interaction can be represented in terms of an exchange energy E_{ex} as

$$E_{ex} = -2J_e \mathbf{S}_1 \cdot \mathbf{S}_2 \qquad [8.21]$$

where $\mathbf{S}_1$ and $\mathbf{S}_2$ are the spin angular momenta of the two electrons and J_e is a numerical quantity called the **exchange integral** that involves integrating the wavefunctions with the various potential energy interaction terms. It therefore depends on the electrostatic interactions and hence on the interatomic distance. For the majority of solids, J_e is negative, so the exchange energy is negative if $\mathbf{S}_1$ and $\mathbf{S}_2$ are in the opposite directions, that is, the spins are antiparallel (as we found in covalent bonding). This is the antiferromagnetic state. For Fe, Co, and Ni, however, J_e is positive. E_{ex} is then negative if $\mathbf{S}_1$ and $\mathbf{S}_2$ are parallel. Spins of the 3*d* electrons on the Fe atoms therefore spontaneously align in the same direction to reduce the exchange energy. This spontaneous magnetization is the phenomenon of ferromagnetism. Figure 8.20 illustrates how J_e changes with the ratio of interatomic separation to the radius of the 3*d* subshell (r/r_d). For the transition metals Fe, Co, and Ni, the r/r_d is such that J_e is positive.[3] In all other cases, it is negative and does not produce ferromagnetic behavior. It should be

[3] According to H. P. Myers, *Introductory Solid State Physics* 2nd ed., London: Taylor and Francis Ltd., 1997, p. 362, there have been no theoretical calculations of the exchange integral J_e for any real magnetic substance.

mentioned that Mn, which is not ferromagnetic, can be alloyed with other elements to increase r/r_d and hence endow ferromagnetism in the alloy.

EXAMPLE 8.3

SATURATION MAGNETIZATION IN IRON The maximum magnetization, called **saturation magnetization** M_{sat}, in iron is about 1.75×10^6 A m^{-1}. This corresponds to all possible net spins aligning parallel to each other. Calculate the effective number of Bohr magnetons per atom that would give M_{sat}, given that the density and relative atomic mass of iron are 7.86 g cm^{-3} and 55.85, respectively.

SOLUTION

The number of Fe atoms per unit volume is

$$n_{at} = \frac{\rho N_A}{M_{at}} = \frac{(7.86 \times 10^3 \text{ kg m}^{-3})(6.022 \times 10^{23} \text{ mol}^{-1})}{55.85 \times 10^{-3} \text{ kg mol}^{-1}}$$

$$= 8.48 \times 10^{28} \text{ atoms m}^{-3}$$

If each Fe atom contributes x number of net spins, then since each net spin has a magnetic moment of β, we have,

$$M_{sat} = n_{at}(x\beta)$$

so

$$x = \frac{M_{sat}}{n_{at}\beta} = \frac{1.75 \times 10^6}{(8.48 \times 10^{28})(9.27 \times 10^{-24})} \approx 2.2$$

In the solid, each Fe atom contributes only 2.2 Bohr magnetons to the magnetization even though the isolated Fe atom has 4 Bohr magnetons. There is no orbital contribution to the magnetic moment per atom in the solid because all the outer electrons, 3*d* and 4*s* electrons, can be viewed as belonging to the whole crystal, or being in an energy band, rather than orbiting individual atoms. A 3*d* electron is attracted by various Fe ions in the crystal and therefore does not experience a central force, in contrast to the 3*d* electron in the isolated Fe atom that orbits the nucleus. The orbital momentum in the crystal is said to be quenched.

We should note that when the magnetization is saturated, all atomic magnetic moments are aligned. The resulting magnetic field within the iron specimen in the absence of an applied magnetizing field ($H = 0$) is

$$B_{sat} = \mu_o M_{sat} = 2.2 \text{ T}$$

8.4 SATURATION MAGNETIZATION AND CURIE TEMPERATURE

The maximum magnetization in a ferromagnet when all the atomic magnetic moments have been aligned as much as possible is called the saturation magnetization M_{sat}. In the iron crystal, for example, this corresponds to each Fe atom with an effective spin magnetic moment of 2.2 Bohr magnetons aligning in the same direction to give a magnetic field $\mu_o M_{sat}$ or 2.2 T. As we increase the temperature, lattice vibrations become more energetic, which leads to a frequent disruption of the alignments of the spins. The spins cannot align perfectly with each other as the temperature increases due to lattice vibrations

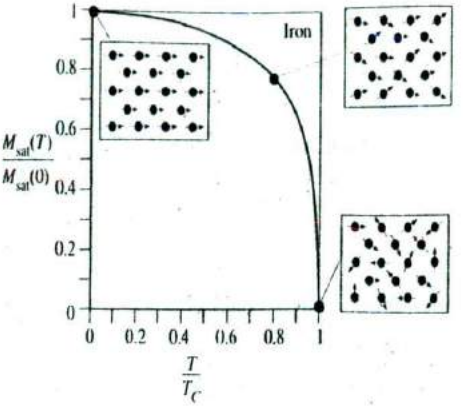


Figure 8.21 Normalized saturated magnetization versus reduced temperature T/T_C where T_C is the Curie temperature (1043 K).

randomly agitating the individual spins. When an energetic lattice vibration passes through a spin site, the energy in the vibration may be sufficient to disorientate the spin of the atom. The ferromagnetic behavior disappears at a critical temperature called the **Curie temperature,** denoted by T_C, when the thermal energy of lattice vibrations in the crystal can overcome the potential energy of the exchange interaction and hence destroy the spin alignments. Above the Curie temperature, the crystal behaves as if it were paramagnetic. The saturation magnetization M_{sat}, therefore, decreases from its maximum value $M_{sat}(0)$ at absolute zero of temperature to zero at the Curie temperature. Figure 8.21 shows the dependence of M_{sat} on the temperature when M_{sat} has been normalized to $M_{sat}(0)$ and the temperature is the reduced temperature, that is, T/T_C. At $T/T_C = 1$, $M_{sat} = 0$. When plotted in this way, the ferromagnets cobalt and nickel follow closely the observed behavior for iron. We should note that since for iron $T_C = 1043$ K, at room temperature, $T/T_C = 0.29$ and M_{sat} is very close to its value at $M_{sat}(0)$.

Since at the Curie temperature, the thermal energy, of the order of kT_C, is sufficient to overcome the energy of the exchange interaction E_{ex} that aligns the spins, we can take kT_C as an order of magnitude estimate of E_{ex}. For iron, E_{ex} is ~0.09 eV and for cobalt this is ~0.1 eV.

Table 8.3 summarizes some of the important properties of the ferromagnets Fe, Co, Ni, and Gd (rare earth metal).

Table 8.3 Properties of the ferromagnets Fe, Co, Ni, and Gd

	Fe	Co	Ni	Gd
Crystal structure	BCC	HCP	FCC	HCP
Bohr magnetons per atom	2.22	1.72	0.60	7.1
$M_{sat}(0)$ (MA m^{-1})	1.75	1.45	0.50	2.0
$B_{sat} = \mu_0 M_{sat}(T)$	2.2	1.82	0.64	2.5
T_C	770 °C	1127 °C	358 °C	16 °C
	1043 K	1400 K	631 K	289 K

8.5 MAGNETIC DOMAINS: FERROMAGNETIC MATERIALS

8.5.1 MAGNETIC DOMAINS

A single crystal of iron does not necessarily possess a net permanent magnetization in the absence of an applied field. If a magnetized piece of iron is heated to a temperature above its Curie temperature and then allowed to cool in the absence of a magnetic field, it will possess no net magnetization. The reason for the absence of net magnetization is due to the formation of magnetic domains that effectively cancel each other, as discussed below. A **magnetic domain** is a region of the crystal in which all the spin magnetic moments are aligned to produce a magnetic moment in one direction only.

Figure 8.22a shows a single crystal of iron that has a permanent magnetization as a result of ferromagnetism (aligning of all atomic spins). The crystal is like a bar magnet with magnetic field lines around it. As we know, there is potential energy (*PE*), called **magnetostatic energy,** stored in a magnetic field, and we can reduce this energy in the external field by dividing the crystal into two domains where the magnetizations are in the opposite directions, as shown in Figure 8.22b. The external magnetic field lines are reduced and there is now less potential energy stored in the magnetic field. There are only field lines at the ends. This arrangement is energetically favorable because the magnetostatic energy has been reduced by decreasing the external field

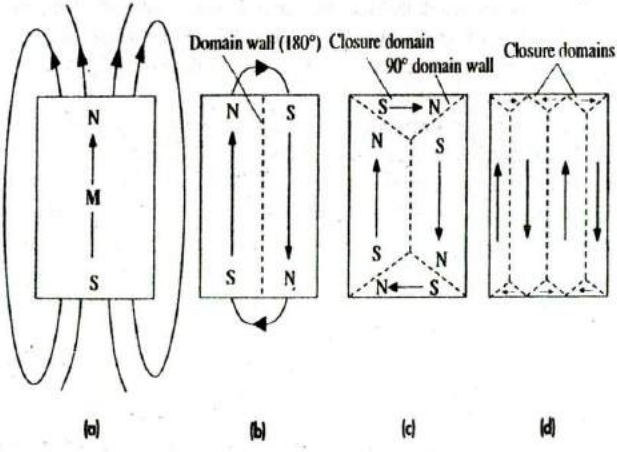


Figure 8.22

(a) Magnetized bar of ferromagnet in which there is only one domain and hence an external magnetic field.

(b) Formation of two domains with opposite magnetizations reduces the external field. There are, however, field lines at the ends.

(c) Domains of closure fitting at the ends eliminate the external fields at the ends.

(d) A specimen with several domains and closure domains. There is no external magnetic field and the specimen appears unmagnetized.

lines. However, there is now a boundary, called a **domain wall** (or **Bloch wall**), between the two domains where the magnetization changes from one direction to the opposite direction and hence the atomic spins do, also. It requires energy to rotate the atomic spin through 180° with respect to its neighbor because the exchange energy favors aligning neighboring atomic spins (0°). The wall in Figure 8.22b is a 180° wall inasmuch as the magnetization through the wall is rotated by 180°. It is apparent that the wall region where the neighboring atomic spins change their relative direction (or orientation) from one domain to the neighboring one has higher *PE* than the bulk of the domain, where all the atomic spins are aligned. As we will show below, the domain wall is not simply one atomic spacing but has a finite thickness, which for iron is typically of the order of 0.1 μm, or several hundred atomic spacings. The excess energy in the wall increases with the area of the wall.

The magnetostatic energy associated with the field lines at the ends in Figure 8.22b can be further reduced by eliminating these external field lines by closing the ends with sideway domains with magnetizations at 90°, as shown in Figure 8.22c. These end domains are **closure domains** and have walls that are 90° walls. The magnetization is rotated through 90° through the wall. Although we have reduced the magnetostatic energy, we have increased the potential energy in the walls by adding additional walls. The creation of magnetic domains continues (spontaneously) until the potential energy reduction in creating an additional domain is the same as the increase in creating an additional wall. The specimen then possesses minimum potential energy and is in equilibrium with no net magnetization. Figure 8.22d shows a specimen with several domains and no net magnetization. The sizes, shapes, and distributions of domains depend on a number of factors, including the size and shape of the whole specimen. For iron particles of dimensions less than of the order of 0.01 μm, the increase in the potential energy in creating a domain wall is too costly and these particles are single domains and hence always magnetized.

The magnetization of each domain is normally along one of the preferred directions in which the atomic spin alignments are easiest (the exchange interaction is the strongest). For iron, the magnetization is easiest along any one of six ⟨100⟩ directions (along cube edges), which are called **easy directions.** The domains have magnetizations along these easy directions. The magnetization of the crystal along an applied field occurs, in principle, by the growth of domains with magnetizations (or components of **M**) along the applied field (**H**), as illustrated in Figure 8.23a and b. For simplicity, the magnetizing field is taken along an easy direction. The Bloch wall between the domains A and B migrates toward the right, which enlarges the domain A and shrinks domain B, with the net result that the crystal has an effective magnetization **M** along **H**. The migration of the Bloch wall is caused by the spins in the wall, and also spins in section B adjacent to the wall, being gradually rotated by the applied field (they experience a torque). The magnetization process therefore involves the motions of Bloch walls in the crystal.

8.5.2 MAGNETOCRYSTALLINE ANISOTROPY

Ferromagnetic crystals characteristically exhibit magnetic anisotropy, which means that the magnetic properties are different along different crystal directions. In the case of iron (BCC), the spins in a domain are most easily aligned in any of the six [100] type

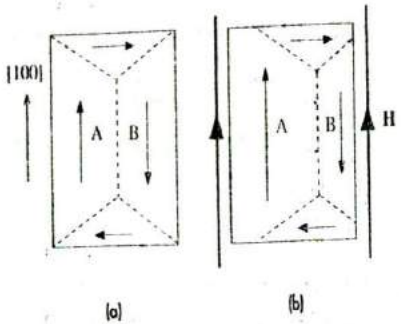


Figure 8.23
(a) An unmagnetized crystal of iron in the absence of an applied magnetic field. Domains A and B are the same size and have opposite magnetizations.
(b) When an external magnetic field is applied, the domain wall migrates into domain B, which enlarges A and shrinks B. The result is that the specimen now acquires net magnetization.

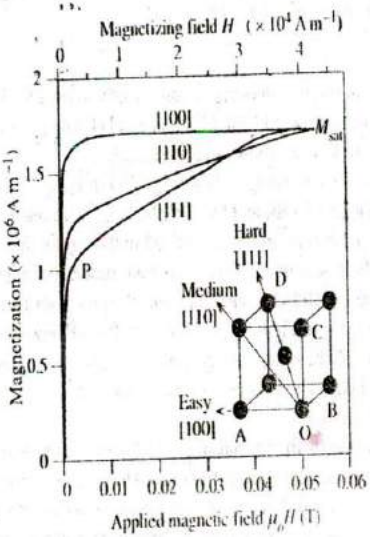


Figure 8.24 Magnetocrystalline anisotropy in a single iron crystal.
M versus *H* depends on the crystal direction and is easiest along [100] and hardest along [111].

directions, collectively labeled as ⟨100⟩, and correspond to the six edges of the cubic unit cell. The exchange interactions are such that spin magnetic moments are most easily aligned with each other if they all point in one of the six ⟨100⟩ directions. Thus ⟨100⟩ directions in the iron crystal constitute the easy directions for magnetization. When a magnetizing field **H** along a [100] direction is applied, as illustrated in Figure 8.23a and b, domain walls migrate to allow those domains (*e.g.*, A) with magnetizations along **H** to grow at the expense of those domains (*e.g.*, B) with magnetizations opposing **H**. The observed *M* versus *H* behavior is shown in Figure 8.24. Magnetization rapidly increases and saturates with an applied field of less than 0.01 T.

On the other hand, if we want to magnetize the crystal along the [111] direction by applying a field along this direction, then we have to apply a stronger field than that along [100]. This is clearly shown in Figure 8.24, where the resulting magnetization along [111] is smaller than that along [100] for the same magnitude of applied field. Indeed, saturation is reached at an applied field that is about a factor of 4 greater than

Table 8.4 Exchange interaction, magnetocrystalline anisotropy energy K, and saturation magnetostriction coefficient λ_{sat}

Material	Crystal	$E_{ex} \approx kT_C$ (meV)	Easy	Hard	K (mJ cm^{-3})	λ_{sat} ($\times 10^{-6}$)
Fe	BCC	90	<100>; cube edge	<111>; cube diagonal	48	20 [100] −20 [111]
Co	HCP	120	// to c axis	⊥ to c axis	450	
Ni	FCC	50	<111>; cube diagonal	<100>; cube edge	5	−46 [100] −24 [111]

NOTE: K is the magnitude of what is called the first anisotropy constant (K_1) and is approximately the magnitude of the anisotropy energy E_a; is an estimate from kT_C, where T_C is the Curie temperature. All approximate values are from various sources. (Further data can be found in D. Jiles, *Introduction to Magnetism and Magnetic Materials*, London, England: Chapman and Hall, 1991.)

that along [100]. The [111] direction in the iron crystal is consequently known as the **hard direction.** The M versus H behavior along [100], [110], and [111] directions in an iron crystal and the associated anisotropy are shown in Figure 8.24.

When an external field is applied along the diagonal direction OD in Figure 8.24, initially all those domains with **M** along OA, OB, and OC, that is, those with magnetization components along OD, grow by consuming those with **M** in the wrong direction and eventually take over the whole specimen. This is an easy process (similar to the process along [100]) and requires small fields and represents the processes from 0 to P on the magnetization curve for [111] in Figure 8.24. However, from P onwards, the magnetizations in the domains have to be rotated away from their easy directions, that is, from OA, OB, and OC toward OD. This process consumes substantial energy and hence needs much stronger applied fields.

It is apparent that the magnetization of the crystal along [100] needs the least energy, whereas that along [111] consumes the greatest energy. The excess energy required to magnetize a unit volume of a crystal in a particular direction with respect to that in the easy direction is called the **magnetocrystalline anisotropy energy** and is denoted by K. For iron, the anisotropy energy is zero for [100] and largest for the [111] direction, about 48 kJ m^{-3} or 3.5×10^{-4} eV per atom. For cobalt, which has the HCP crystal structure, the anisotropy energy is at least an order of magnitude greater. Table 8.4 summarizes the easy and hard directions, and the anisotropy energy K for the hard direction.

8.5.3 DOMAIN WALLS

We recall that the spin magnetic moments rotate across a domain wall. We mentioned that the wall is not simply one atomic spacing wide, as this would mean two neighboring spins being at 180° to each other and hence possessing excessive exchange interaction. A schematic illustration of the structure of a typical 180° Bloch wall, between two domains A and B, is depicted in Figure 8.25. It can be seen that the neighboring spin magnetic moments are rotated gradually, and over several hundred atomic spacings the magnetic moment reaches a rotation of 180°. Exchange

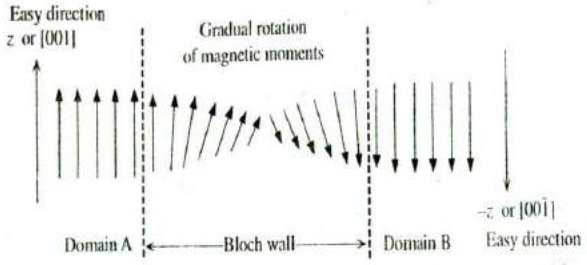


Figure 8.25 In a Bloch wall, the neighboring spin magnetic moments rotate gradually, and it takes several hundred atomic spacings to rotate the magnetic moment by 180°.

forces between neighboring atomic spins favor very little relative rotation. Had it been left to exchange forces alone, relative rotation of neighboring spins would be so minute that the wall would have to be very thick (infinitely thick) to achieve a 180° rotation.

However, magnetic moments that are oriented away from the easy direction possess excess energy, called the anisotropy energy (K). If the wall is thick, then it will contain many magnetic moments rotated away from the easy direction and there would be a substantial anisotropy energy in the wall. Minimum anisotropy energy in the wall is obtained when the magnetic moment changes direction by 180° from the easy direction along $+z$ to that along $-z$ in Figure 8.25 without any intermediate rotations away from z. This requires a wall of one atomic spacing. In reality, the wall thickness is a compromise between the exchange energy, demanding a thick wall, and anisotropy energy, demanding a thin wall. The equilibrium wall thickness is that which minimizes the total potential energy, which is the sum of the exchange energy *and* the anisotropy energy within the wall. This thickness turns out to be ~0.1 μm for iron and less for cobalt, in which the anisotropy energy is greater.

EXAMPLE 8.4

MAGNETIC DOMAIN WALL ENERGY AND THICKNESS The Bloch wall energy and thickness depend on two main factors: the exchange energy E_{ex} (J atom^{-1}) and magnetocrystalline energy K (J m^{-3}). Suppose that we consider a Bloch wall of unit area, and thickness δ, and calculate the potential energy U_{wall} in this wall due to the exchange energy and the magnetocrystalline anisotropy energy. The spins change by 180° across the thickness δ of the Bloch wall as in Figure 8.25. The contribution $U_{exchange}$ from the exchange energy arises because it takes energy to rotate one spin with respect to another. If the thickness δ is large, then the angular change from one spin to the next will be small, and the exchange energy contribution $U_{exchange}$ will also be small. Thus, $U_{exchange}$ is inversely proportional to δ. $U_{exchange}$ is also directly proportional to E_{ex} which gauges the magnitude of this exchange energy; it costs E_{ex} to rotate the two spins 180° to each other. Thus, $U_{exchange} \propto E_{ex}/\delta$.

The anisotropy energy contribution $U_{anisotropy}$ arises from having spins point away from the easy direction. If the thickness δ is large, there are more and more spin moments that are aligned away from the easy direction, and the anisotropy energy contribution $U_{anisotropy}$ is also large. Thus, $U_{anisotropy}$ is proportional to δ, and also, obviously, to the anisotropy energy K that gauges the magnitude of this energy. Thus, $U_{anisotropy} \propto K\delta$.

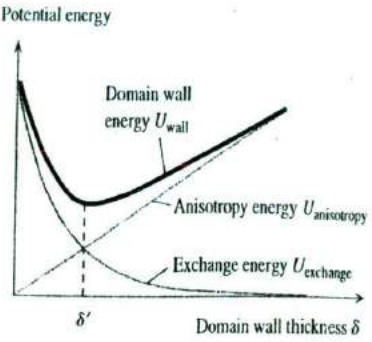


Figure 8.26 The potential energy of a domain wall depends on the exchange and anisotropy energies.

Figure 8.26 shows the contributions of the exchange and anisotropy energies, $U_{exchange}$ and $U_{anisotropy}$, to the total Bloch wall energy as a function of wall thickness δ. It is clear that exchange and anisotropy energies have opposite (or conflicting) requirements on the wall thickness. There is, however, an optimum thickness δ' that *minimizes* the Bloch wall energy, that is, a thickness that balances the requirements of exchange and anisotropy forces.

If the interatomic spacing is a, then there would be $N = \delta/a$ atomic layers in the wall. Since the spin moment angle changes by 180° across δ, we can calculate the relative spin orientations $(180°/N)$ of adjacent atomic layers, and hence we can find the exact contributions of exchange and anisotropy energies. We do not need the exact mathematics, but the final result is that the potential energy U_{wall} per unit area of the wall is approximately

Potential energy of a Bloch wall

$$U_{wall} \approx \frac{\pi^2 E_{ex}}{2a\delta} + K\delta$$

The first term on the right is the exchange energy contribution (proportional to E_{ex}/δ), and the second is the anisotropy energy contribution (proportional to $K\delta$); both have the features we discussed.

Show that the minimum energy occurs when the wall has the thickness

Bloch wall thickness

$$\delta' = \left(\frac{\pi^2 E_{ex}}{2aK}\right)^{1/2}$$

Taking $E_{ex} \approx kT_C$, where T_C is the Curie temperature, and for iron, $K \approx 50$ kJ m^{-3}, and $a \approx 0.3$ nm, estimate the thickness of a Bloch wall and its energy per unit area.

SOLUTION

We can differentiate U_{wall} with respect to δ,

$$\frac{dU_{wall}}{d\delta} = -\frac{\pi^2 E_{ex}}{2a\delta^2} + K$$

and then set it to zero for $\delta = \delta'$ to find,

$$\delta' = \left(\frac{\pi^2 E_{ex}}{2aK}\right)^{1/2}$$

Since $T_C = 1043$ K, $E_{ex} = kT_C = (1.38 \times 10^{-23}\ \text{J K}^{-1})(1043\ \text{K}) = 1.4 \times 10^{-20}$ J, so that

$$\delta' = \left(\frac{\pi^2 E_{ex}}{2aK}\right)^{1/2} = \left[\frac{\pi^2(1.4 \times 10^{-20})}{2(0.3 \times 10^{-9})(50,000)}\right]^{1/2} = 6.8 \times 10^{-8}\text{m} \quad \text{or} \quad 68\ \text{nm}$$

$$\text{and} \quad U_{\text{wall}} = \frac{\pi^2 E_{\text{ex}}}{2a\delta'} + K\delta' = \frac{\pi^2(1.4 \times 10^{-20})}{2(0.3 \times 10^{-9})(6.8 \times 10^{-8})} + (50 \times 10^3)(6.8 \times 10^{-8})$$

$$= 0.007 \text{ J m}^{-2} \quad \text{or} \quad 7 \text{ mJ m}^{-2}$$

A better calculation gives δ' and U_{wall} as 40 nm and 3 mJ m^{-2}, respectively, about the same order of magnitude.[4] The Bloch wall thickness is roughly 70 nm or $\delta/a = 230$ atomic layers. It is left as an exercise to show that when $\delta = \delta'$, the exchange and anisotropy energy contributions are equal.

8.5.4 MAGNETOSTRICTION

If we were to strain a ferromagnetic crystal (by applying a suitable stress) along a certain direction, we would change the interatomic spacing not only along this direction but also in other directions and hence change the exchange interactions between the atomic spins. This would lead to a change in the magnetization properties of the crystal. In the converse effect, the magnetization of the crystal generates strains or changes in the physical dimensions of the crystal. For example, in very qualitative terms, when an iron crystal is magnetized along the [111] direction by a strong field, the atomic spins within domains are rotated from their easy directions toward the hard [111] direction. These electron spin rotations involve changes in the electron charge distributions around the atoms and therefore affect the interatomic bonding and hence the interatomic spacing. When an iron crystal is placed in a magnetic field along an easy direction [100], it gets longer along this direction but contracts in the transverse directions [010] and [001], as depicted in Figure 8.27. The reverse is true for nickel. The longitudinal strain $\Delta\ell/\ell$ along the direction of magnetization is called the **magnetostrictive constant,** denoted by λ. The magnetostrictive constant depends on the crystal direction and may be positive (extension) or negative (contraction). Further, λ depends on the applied field and can even change sign as the field is increased; for example, λ for iron along the [110] direction is initially positive and then, at higher fields, becomes negative. When the crystal reaches saturation magnetization, λ also reaches saturation, called **saturation magnetostriction strain** λ_{sat}, which is typically 10^{-6}–10^{-5}. Table 8.4 summarizes the λ_{sat} values for Fe and Ni along the easy and hard directions. The crystal lattice strain energy associated with magnetostriction is called the **magnetostrictive energy,** which is typically less than the anisotropy energy.

Magnetostriction is responsible for the transformer hum noise one hears near power transformers. As the core of a transformer is magnetized one way and then in the opposite direction under an alternating voltage, the alternating changes in the longitudinal strain vibrate the surrounding environment, air, oil, and so forth, and generate an acoustic noise at twice the main frequency, or 120 Hz, and its harmonics. (Why?)

The magnetostrictive constant can be controlled by alloying metals. For example, λ_{sat} along the easy direction for nickel is negative and for iron it is positive, but for the alloy 85% Ni–15% Fe, it is zero. In certain magnetic materials, λ can be quite large,

[4] See, for example, D. Jiles, *Introduction to Magnetism and Magnetic Materials*, London, England: Chapman and Hall, 1991.

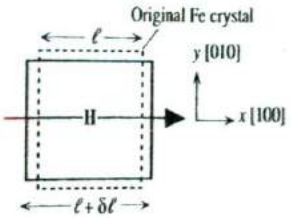


Figure 8.27 Magnetostriction means that the iron crystal in a magnetic field along x, an easy direction, elongates along x but contracts in the transverse directions (in low fields).

greater than 10^{-4}, which has opened up new areas of sensor applications based on the magnetostriction effect. For example, it may be possible to develop torque sensors for automotive steering applications by using Co-ferrite type magnetic materials[5] (*e.g.*, $CoO–Fe_2O_3$ or similar compounds) that have λ_{sat} of the order of 10^{-4}.

8.5.5 DOMAIN WALL MOTION

The magnetization of a single ferromagnetic crystal involves the motions of domain boundaries to allow the favorably oriented domains to grow at the expense of domains with magnetizations directed away from the field (Figure 8.23). The motion of a domain wall in a crystal is affected by crystal imperfections and impurities and is not smooth. For example, in a 90° Bloch wall, the magnetization changes direction by 90° across the boundary. Due to magnetostriction (Figure 8.27), there is a change in the distortion of the lattice across the 90° boundary, which leads to a complicated strain and hence stress distribution around this boundary. We also know that crystal imperfections such as dislocations and point defects also have strain and stress distributions around them. Domain walls and crystal imperfections therefore interact with each other. Dislocations are line defects that have a substantial volume of strained lattice around them. Figure 8.28 visualizes a dislocation with tensile and compressive strains around it and a domain wall that has a tensile strain on the side of the dislocation. If the wall gets close to the dislocation, the tensile and compressive strains cancel, which results in an unstrained lattice and hence a lower strain energy. This energetically favorable arrangement keeps the domain boundary close to the dislocation. It now takes greater magnetic field to snap away the boundary from the dislocation. Domain walls also interact with nonmagnetic impurities and inclusions. For example, an inclusion that finds itself in a domain becomes magnetized and develops south and north poles, as shown in Figure 8.29a. If the domain wall were to intersect the inclusion and if there were to be two neighboring domains around the inclusion, as in Figure 8.29b, then the magnetostatic energy would be lowered—energetically a favorable event. This reduction in magnetostatic potential energy means that it now takes greater force to move the domain wall past the impurity, as if the wall were "pinned" by the impurity.

The motion of a domain wall in a crystal is therefore not smooth but rather jerky. The wall becomes pinned somewhere by a defect or an impurity and then needs a

[5] See, for example, D. Jiles and C. C. H. Lo, *Sensors and Actuators*, **A106**, 3, 2003.

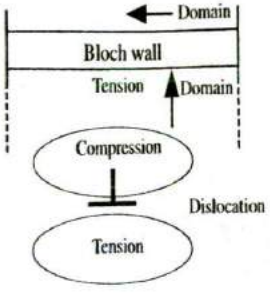


Figure 8.28 Stress and strain distributions around a dislocation and near a domain wall.

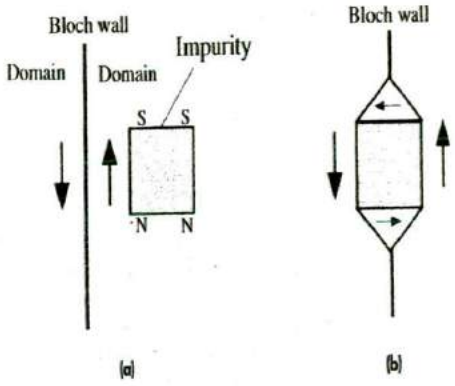


Figure 8.29 Interaction of a Bloch wall with a nonmagnetic (no permanent magnetization) inclusion.
(a) The inclusion becomes magnetized and there is magnetostatic energy.
(b) This arrangement has lower potential energy and is thus favorable.

greater applied field to break free. Once it snaps off, the domain wall is moved until it is attracted by another type of imperfection, where it is held until the field increases further to snap it away again. Each time the domain wall is snapped loose, lattice vibrations are generated, which means loss of energy as heat. The whole domain wall motion is nonreversible and involves energy losses as heat to the crystal.

8.5.6 POLYCRYSTALLINE MATERIALS AND THE M VERSUS H BEHAVIOR

The majority of the magnetic materials used in engineering are polycrystalline and therefore have a microstructure that consists of many grains of various sizes and orientations depending on the preparation and thermal history of the component. In an unmagnetized polycrystalline sample, each crystal grain will possess domains, as depicted in Figure 8.30. The domain structure in each grain will depend on the size and shape of the grain and, to some extent, on the magnetizations in neighboring grains. Although very small grains, perhaps smaller than 0.1 μm, may be single domains, in most cases the majority of the grains will have many domains. Overall, the structure will possess no net magnetization, provided that it was not previously subjected to an applied magnetic field. We can assume that the component was heated to a temperature above the Curie point and then allowed to cool to room temperature without an applied field.

Suppose that we start applying a very small external magnetic field ($\mu_o H$) along some direction, which we can arbitrarily label as $+x$. The domain walls within various grains begin to move small distances, and favorably oriented domains (those with a component of M along $+x$) grow a little larger at the expense of those pointing away from the field, as indicated by point a in Figure 8.31. The domain walls that are pinned by imperfections tend to bow out. There is a very small but net magnetization

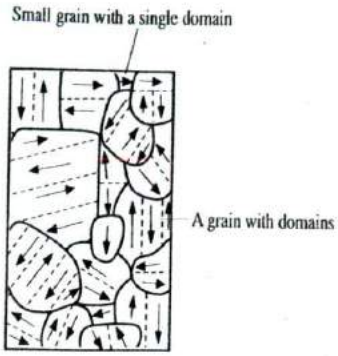


Figure 8.30 Schematic illustration of magnetic domains in the grains of an unmagnetized polycrystalline iron sample.
Very small grains have single domains.

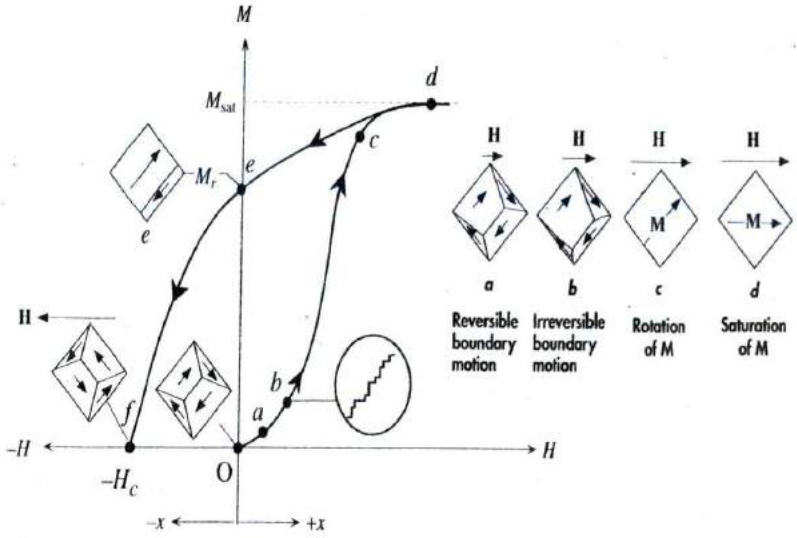


Figure 8.31 *M* versus *H* behavior of a previously unmagnetized polycrystalline iron specimen.
An example grain in the unmagnetized specimen is that at O.
(a) Under very small fields, the domain boundary motion is reversible.
(b) The boundary motions are irreversible and occur in sudden jerks.
(c) Nearly all the grains are single domains with saturation magnetizations in the easy directions.
(d) Magnetizations in individual grains have to be rotated to align with the field *H*.
(e) When the field is removed, the specimen returns along *d* to *e*.
(f) To demagnetize the specimen, we have to apply a magnetizing field of H_c in the reverse direction.

along the field, as indicated by the *Oa* region in the magnetization versus magnetizing field (*M* versus *H*) behavior in Figure 8.31. As we increase the magnetizing field, the domain motions extend larger distances, as shown for point *b* in Figure 8.31, and walls encounter various obstacles such as crystal imperfections, impurities, second phases, and so on, which tend to attract the walls and thereby hinder their motions. A

domain wall that is stuck (or pinned) at an imperfection at a given field cannot move until the field increases sufficiently to provide the necessary force to snap the wall free, which then suddenly surges forward to the next obstacle. As a wall suddenly snaps free and shoots forward to the next obstacle, essentially two causes lead to heat generation. Sudden changes in the lattice distortion, due to magnetostriction, create lattice waves that carry off some of the energy. Sudden changes in the magnetization induce eddy currents that dissipate energy via Joule heating (domains have a finite electrical resistance). These processes involve energy conversion to heat and are irreversible. Sudden jerks in the wall motions lead to small jumps in the magnetization of the specimen as the magnetizing field is increased; the phenomenon is known as the **Barkhausen effect.** If we could examine the magnetization precisely with a highly sensitive instrument, we would see jumps in the M versus H behavior, as shown in the inset in Figure 8.31.

As we increase the field, magnetization continues to increase by jerky domain wall motions that enlarge domains with favorably oriented magnetizations and shrink away those with magnetizations pointing away from the applied field. Eventually domain wall motions leave each crystal grain with a single domain and magnetization in one of the easy directions, as indicated by point c in Figure 8.31. Although some grains would be oriented to have the easy direction and hence M along the applied field, the magnetization in many grains will be pointing at some angle to H as shown for point c in Figure 8.31. From then until point d, the increase in the applied field forces the magnetization in a grain, such as that at point c to rotate toward the direction of H. Eventually the applied field is sufficiently strong to align M along H, and the specimen reaches saturation magnetization M_{sat}, directed along H or $+x$, as at point d in Figure 8.31.

If we were to decrease and remove the magnetizing field, the magnetization in each grain would rotate to align parallel with the nearest easy direction in that grain. Further, in some grains, additional small domains may develop that reduce the magnetization within that grain, as indicated at point e in Figure 8.31. This process, from point d to point e, leaves the specimen with a permanent magnetization, called the **remanent** or **residual magnetization** and denoted by M_r.

If we were now to apply a magnetizing field in the reverse direction $-x$, the magnetization of the specimen, still along $+x$, would decrease and eventually, at a sufficiently large applied field M would be zero and the sample would have been totally demagnetized. This is shown as point f in Figure 8.31. The magnetizing field H_c required to totally demagnetize the sample is called the **coercivity** or the **coercive field.** It represents the resistance of the sample to demagnetization. We should note that at point f in Figure 8.31, the sample again has grains with many domains, which means that during the demagnetization process, from point e to point f, new domains had to be generated. The demagnetization process invariably involves the nucleation of various domains at various crystal imperfections to cancel the overall magnetization. The treatment of the nucleation of domains is beyond the scope of this book; we will nonetheless, accept it as required process for the demagnetization of the crystal grains.

If we continue to increase the applied magnetic field along $-x$, as illustrated in Figure 8.32a, the process from point f onward becomes similar to that described for magnetization from point a to point d in Figure 8.31 along $+x$ except that it is now directed along $-x$. At point g, the sample reaches saturation magnetization along the $-x$ direction. The full M versus H behavior as the magnetizing field is cycled between

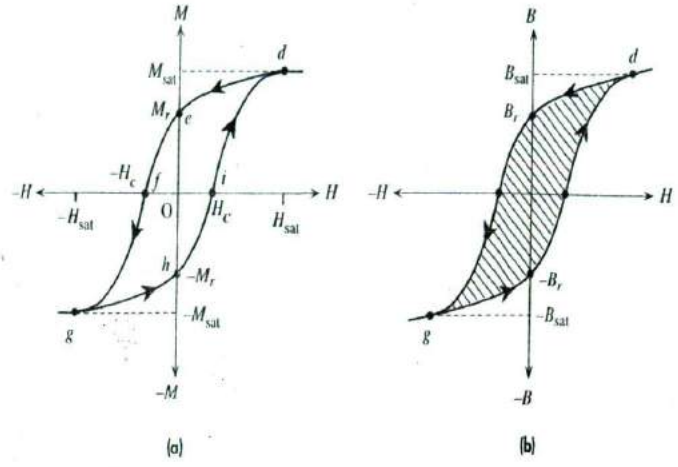


Figure 8.32

(a) A typical M versus H hysteresis curve.

(b) The corresponding B versus H hysteresis curve. The shaded area inside the hysteresis loop is the energy loss per unit volume per cycle.

$+x$ to $-x$ has a closed loop shape, shown in Figure 8.32a, called the **hysteresis loop.** We observe that in both $+x$ and $-x$ directions, the magnetization reaches saturation M_{sat} when H reaches H_{sat}, and on removing the applied field, the specimen retains an amount of permanent magnetization, represented by points e and h and denoted by M_r. The necessary applied field of magnitude H_c that is needed to demagnetize the specimen is the coercivity (or coercive field), which is represented by points f and i. The initial magnetization curve, $Oabcd$ in Figure 8.31, which starts from an unmagnetized state, is called the **initial magnetization curve.**

We can, of course, monitor the magnetic field B instead of M, as in Figure 8.32b, where

$$B = \mu_o M + \mu_o H$$

which leads to a hysteresis loop in the B versus H behavior. The very slight increase in B with H when M is in saturation is due to the permeability of free space ($\mu_o H$). The area enclosed within the B versus H hysteresis loop, shown as the hatched region in Figure 8.32b, represents the energy dissipated per unit volume per cycle of applied field variation.

Suppose we do not take a magnetic material to saturation but still subject the specimen to a cyclic applied field alternating between the $+x$ and $-x$ directions. Then the hysteresis loop would be different than that when the sample is taken all the way to saturation, as shown in Figure 8.33. The magnetic field in the material does not reach B_{sat} (corresponding to M_{sat}) but instead reaches some maximum value B_m when the magnetizing field is H_m. There is still a hysteresis effect because the magnetization and demagnetization processes are nonreversible and do not retrace each other. The shape of the hysteresis

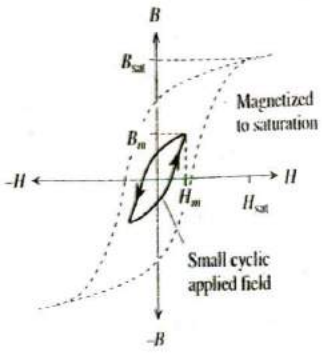


Figure 8.33 The B versus H hysteresis loop depends on the magnitude of the applied field in addition to the material and sample shape and size.

loop depends on the magnitude of the applied field in addition to the material and sample shape and size. The area enclosed within the loop is still the energy dissipated per unit volume per cycle of applied field oscillation. The hysteresis loop taken to saturation, as in Figure 8.32a and b, is called the **saturation (major) hysteresis loop.** It is apparent from Figure 8.33 that the remanence and coercivity exhibited by the sample depend on the B–H loop. The quoted values normally correspond to the saturation hysteresis loop.

Ferrimagnetic materials exhibit properties that closely resemble those of ferromagnetic materials. One can again identify distinct magnetic domains and domain wall motions during magnetization and demagnetization that also lead to B–H hysteresis curves with the same characteristic parameters, namely, saturation magnetization (M_{sat} and B_{sat} at H_{sat}), remanence (M_r and B_r), coercivity (H_c), hysteresis loss, and so on.

8.5.7 DEMAGNETIZATION

The B–H hysteresis curves, as in Figure 8.32b, that are commonly given for magnetic materials represent B versus H behavior observed under repeated cycling. The applied field intensity H is cycled back and forward between the $-x$ and $+x$ directions. If we were to try and demagnetize a specimen with a remanent magnetization at point e in Figure 8.34 by applying a reverse field intensity, then the magnetization would move along from point e to point f. If at point f we were to suddenly switch off the applied field, we would find that B does not actually remain zero but recovers along f-e' to point e' and attains some value B_r'. The main reason is that small domain wall motions are reversible and as soon as the field is removed, there is some reversible domain wall motion "bouncing back" the magnetization along f-e'. We can anticipate this recovery and remove the field intensity at some point f' so that the sample recovers along f'O and the magnetization ends up being zero. However, to remove the field intensity at point f', we need to know not only the exact B–H behavior but also the exact location of point f' (or the recovery behavior). The simplest method to demagnetize the sample is first to cycle H with ample magnitude to reach saturation and then to continue cycling H but with a gradually decreasing magnitude, as depicted in Figure 8.35. As H is cycled with a decreasing magnitude, the sample traces out smaller and smaller B–H loops until the B–H loops are so small that they end up at the origin when H reaches

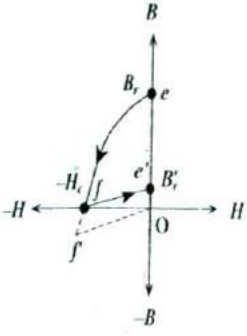


Figure 8.34 Removal of the demagnetizing field at point f does not necessarily result in zero magnetization as the sample recovers along f–e'.

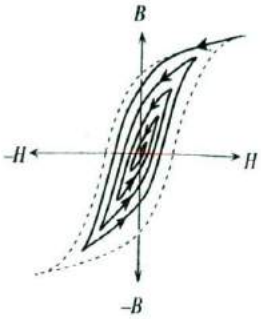


Figure 8.35 A magnetized specimen can be demagnetized by cycling the field intensity with a decreasing magnitude, that is, tracing out smaller and smaller B–H loops until the origin is reached, $H = 0$.

zero. The demagnetization process in Figure 8.35 is commonly known as **deperming.** Undesirable magnetization of various magnetic devices such as recording heads is typically removed by this deperming process (for example, a demagnetizing gun brought close to a magnetized recording head implements deperming by applying a cycled H with decreasing magnitude).

EXAMPLE 8.5 **ENERGY DISSIPATED PER UNIT VOLUME AND THE HYSTERESIS LOOP** Consider a toroidal coil with an iron core that is energized from a voltage supply through a rheostat, as shown in Figure 8.11. Suppose that by adjusting the rheostat we can adjust the current i supplied to the coil and hence the magnetizing field H in the core material. H and i are simply related by Ampere's law. However, the magnetic field B in the core is determined by the B–H characteristics of the core material. From electromagnetism (see Example 8.2), we know that the battery has to do work dE_{vol} per unit volume of core material to increase the magnetic field by dB, where

$$dE_{vol} = H\,dB$$

so that the total energy or work involved per unit volume in changing the magnetic field from an initial value B_1 to a final value B_2 in the core is

Work done per unit volume during magnetization

$$E_{vol} = \int_{B_1}^{B_2} H\,dB \qquad [8.22]$$

where the integration limits are determined by the initial and final magnetic fields.

Equation 8.22 corresponds to the area between the B–H curve and the B axis between B_1 and B_2. Suppose that we take the iron core in the toroid from point P on the hysteresis curve to Q, as shown in Figure 8.36. This is a magnetization process for which energy is put into the sample. The work done per unit volume from P to Q is the area $PQRS$, shown as hatched. On returning the sample to the same initial magnetization (same magnetic field B as we had at P), taking it from Q to S, energy is returned from the core into the electric circuit. This energy per unit volume is the area QRS, shown as gray, and is less than $PQRS$ during magnetization. The difference is the energy dissipated in the sample as heat (moving domain walls and so on) and

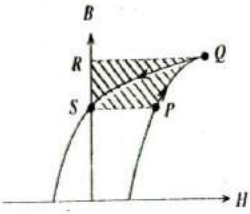


Figure 8.36 The area between the *B–H* curve and the *B* axis is the energy absorbed per unit volume in magnetization or released during demagnetization.

corresponds to the hysteresis loop area *PQS*. Over one full cycle, the energy dissipated per unit volume is the total hysteresis loop area.

The hysteresis loop and hence the energy dissipated per unit volume per cycle depend not only on the core material but also on the magnitude of the magnetic field (B_m), as apparent in Figure 8.33. For example, for magnetic steels used in transformer cores, the hysteresis **power loss** P_h per unit volume of core is empirically expressed in terms of the maximum magnetic field B_m and the ac frequency f as[6]

Hysteresis power loss per m³

$$P_h = K f B_m^n \qquad [8.23]$$

where K is a constant that depends on the core material (typically, $K = 150.7$), f is the ac frequency (Hz), B_m is the maximum magnetic field (T) in the core (assumed to be in the range 0.1–1.5 T), and $n = 1.6$. According to Equation 8.23, the hysteresis loss can be decreased by operating the transformer with a reduced magnetic field.

8.6 SOFT AND HARD MAGNETIC MATERIALS

8.6.1 DEFINITIONS

Based on their *B–H* behavior, engineering materials are typically classified into soft and hard magnetic materials. Their typical *B–H* hysteresis curves are shown in Figure 8.37. Soft magnetic materials are easy to magnetize and demagnetize and hence require relatively low magnetic field intensities. Put differently, their *B–H* loops are narrow, as shown in Figure 8.37. The hysteresis loop has a small area, so the hysteresis power loss per cycle is small. Soft magnetic materials are typically suitable for applications where repeated cycles of magnetization and demagnetization are involved, as in electric motors, transformers, and inductors, where the magnetic field varies cyclically. These applications also require low hysteresis losses, or small hysteresis loop area. Electromagnetic relays that have to be turned on and off require the relay iron to be magnetized and demagnetized and therefore need soft magnetic materials.

Hard magnetic materials, on the other hand, are difficult to magnetize and demagnetize and hence require relatively large magnetic field intensities, as apparent in Figure 8.37. Their *B–H* curves are broad and almost rectangular. They possess relatively large coercivities, which means that they need large applied fields to be demagnetized. The coercive field for hard materials can be millions of times greater than those for soft

[6] This is the power engineers Steinmetz equation for commercial magnetic steels. It has been applied not only to silicon irons (Fe + few percent Si) but also to a wide range of magnetic materials.

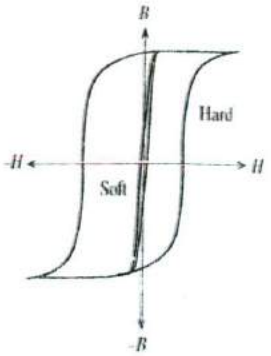


Figure 8.37 Soft and hard magnetic materials.

magnetic materials. Their characteristics make hard magnetic materials useful as permanent magnets in a variety of applications. It is also clear that the magnetization can be switched from one very persistent direction to another very persistent direction, from $+B_r$ to $-B_r$, by a suitably large magnetizing field intensity. As the coercivity is strong, both the states $+B_r$ and $-B_r$ persist until a suitable (large) magnetic field intensity switches the field from one direction to the other. It is apparent that hard magnetic materials can also be used in magnetic storage of digital data, where the states $+B_r$ and $-B_r$ can be made to represent 1 and 0 (or vice versa).

8.6.2 INITIAL AND MAXIMUM PERMEABILITY

It is useful to characterize the magnetization of a material by a relative permeability μ_r, since this simplifies magnetic calculations. For example, inductance calculations become straightforward if one could represent the magnetic material by μ_r alone. But it is clear from Figure 8.38a that

$$\mu_r = \frac{B}{\mu_0 H}$$

is not even approximately constant because it depends on the applied field and the magnetic history of the sample. Nonetheless, we still find it useful to specify a relative permeability to compare various materials and even use it in various calculations. The definition $\mu_r = B/(\mu_0 H)$ represents the slope of the straight line from the origin O to the point P, as shown in Figure 8.38a. This is a maximum when the line becomes a tangent to the B–H curve at P, as in the figure. Any other line from O to the B–H curve that is not a tangent does not yield a maximum relative permeability (the mathematical proof is left to the reader, though the argument is intuitively acceptable from the figure). The **maximum relative permeability**, as defined in Figure 8.38a, is denoted by $\mu_{r,\max}$ and serves as a useful magnetic parameter. The point P in Figure 8.38a that defines the maximum permeability corresponds to what is called the "knee" of the B–H curve. Many transformers are designed to operate with the maximum magnetic field in the core reaching this knee point. For pure iron, $\mu_{r,\max}$ is less than 10^4, but for certain

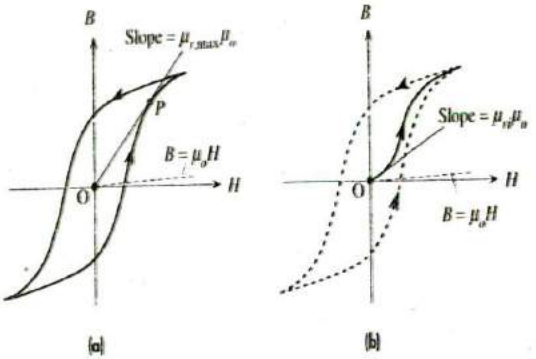


Figure 8.38 Definitions of (a) maximum permeability and (b) initial permeability.

soft magnetic materials such as supermalloys (a nickel–iron alloy), the values of $\mu_{r,max}$ can be as high as 10^6.

Initial relative permeability, denoted as μ_{ri}, represents the initial slope of the initial B versus H curve as the material is first magnetized from an unmagnetized state, as illustrated in Figure 8.38b. This definition is useful for soft magnetic materials that are employed at very low magnetic fields (*e.g.*, small signals in electronics and communications engineering). In practice, weak applied magnetic fields where μ_{ri} is useful are typically less than 10^{-4} T. In contrast, $\mu_{r,max}$ is useful when the magnetic field in the material is not far removed from saturation. Initial relative permeability of a magnetically soft material can vary by orders of magnitude. For example, μ_{ri} for iron is 150, whereas for supernumetal-200, a commercial alloy of nickel and iron, it is about 2×10^5.

8.7 SOFT MAGNETIC MATERIALS: EXAMPLES AND USES

Table 8.5 identifies what properties are desirable in soft magnetic materials and also lists some typical examples with various applications. An *ideal* soft magnetic material would have zero coercivity (H_c), a very large saturation magnetization (B_{sat}), zero remanent magnetization (B_r), zero hysteresis loss, and very large $\mu_{r,max}$ and μ_{ri}. A number of example materials, from pure iron to ferrites, which are ferrimagnetic, are listed in Table 8.5. Pure iron, although soft, is normally not used in electric machines (except in a few specific relay-type applications) because its good conductivity allows large eddy currents to be induced under varying fields. Induced eddy currents in the iron lead to Joule losses (RI^2), which are undesirable. The addition of a few percentages of silicon to iron (silicon–iron), known typically as silicon–steels, increases the resistivity and hence reduces the eddy current losses. Silicon–iron is widely used in power transformers and electric machinery.

The nickle–iron alloys with compositions around 77% Ni–23% Fe constitute an important class of soft magnetic materials with low coercivity, low hysteresis losses, and high permeabilities (μ_{ri} and $\mu_{r,max}$). High μ_{ri} makes these alloys particularly useful in low magnetic field applications that are typically found in high-frequency work in

Table 8.5 Selected soft magnetic materials and some typical values and applications

Magnetic Material	$\mu_o H_c$ (T)	B_{sat} (T)	B_r (T)	μ_{ri}	$\mu_{r,max}$	W_h	Typical Applications
Ideal soft	0	Large	0	Large	Large	0	Transformer cores, inductors, electric machines, electromagnet cores, relays, magnetic recording heads.
Iron (commercial) grade, 0.2% impurities)	$<10^{-4}$	2.2	<0.1	150	10^4	250	Large eddy current losses. Generally not preferred in electric machinery except in some specific applications (*e.g.*, some electromagnets and relays).
Silicon iron (Fe: 2–4% Si)	$<10^{-4}$	2.0	0.5–1	10^3	10^4–4×10^5	30–100	Higher resistivity and hence lower eddy current losses. Wide range of electric machinery (*e.g.*, transformers).
Supermalloy (79% Ni–15.5% Fe–5% Mo–0.5% Mn)	2×10^{-7}	0.7–0.8	<0.1	10^5	10^6	<0.5	High permeability, low-loss electric devices, *e.g.*, specialty transformers, magnetic amplifiers.
78 Permalloy (78.5% Ni–21.5% Fe)	5×10^{-6}	0.86	<0.1	8×10^3	10^5	<0.1	Low-loss electric devices, audio transformers, HF transformers, recording heads, filters.
Glassy metals, Fe–Si–B	2×10^{-6}	1.6	$<10^{-6}$	—	10^5	20	Low-loss transformer cores.
Ferrites, Mn–Zn ferrite	10^{-5}	0.4	<0.01	2×10^3	5×10^3	<0.01	HF low-loss applications. Low conductivity ensures negligible eddy current losses. HF transformers, inductors (*e.g.*, pot cores, E and U cores), recording heads.

NOTE: W_h is the hysteresis loss, energy dissipated per unit volume per cycle in hysteresis losses, J m^{-3} cycle^{-1}, typically at $B_m = 1$ T.

electronics (*e.g.*, audio and wide-band transformers). They have found many engineering uses in sensitive relays, pulse and wide-band transformers, current transformers, magnetic recording heads, magnetic shielding, and so forth. Alloying iron with nickel increases the resistivity and hence reduces eddy current losses. The magnetocrystalline anisotropy energy is least at these nickel compositions, which leads to easier domain wall motions and hence smaller hysteresis losses. There are a number of commercial nickel–iron alloys whose application depends on the exact composition (which may also have a few percentages of Mo, Cu, or Cr) and the method of preparation (*e.g.*, mechanical rolling). For example, supermalloy (79% Ni–16% Fe–5% Co) has $\mu_{ri} \approx 10^5$, compared with commercial grade iron, which has μ_{ri} less than 10^3.

Amorphous magnetic metals, as the name implies, have no crystal structure (they only have short-range order) and consequently possess no crystalline imperfections such as grain boundaries and dislocations. They are prepared by rapid solidification of the melt by using special techniques such as melt spinning (as described in Chapter 1). Typically they are thin ribbons by virtue of their preparation method. Since they have no crystal structure, they also have no magnetocrystalline anisotropy energy, which means that all

directions are easy. The absence of magnetocrystalline anisotropy and usual crystalline defects which normally impede domain wall motions, leads to low coercivities and hence to soft magnetic properties. The coercivity, however, is not zero inasmuch as there is still some magnetic anisotropy due to the directional nature of the strains frozen in the metal during rapid solidification. By virtue of their disordered structure, these metallic glasses also have higher resistivities and hence they have smaller eddy current losses. Although they are ideally suited for various transformer and electric machinery applications, their limited size and shape, at present, prevent their use in power applications.

Ferrites are ferrimagnetic materials that are typically oxides of mixed transition metals, one of which is iron. For example, Mn ferrite is $MnFe_2O_4$ and MgZn ferrite is $Mn_{1-x}Zn_xFe_2O_4$. They are normally insulators and therefore do not suffer from eddy current losses. They are ideal as magnetic materials for high-frequency work where eddy current losses would prevent the use of any material with a reasonable conductivity. Although they can have high initial permeabilities and low losses, they do not possess as large saturation magnetizations as ferromagnets, and further, their useful temperature range (determined by the Curie temperature) is lower. There are many types of commercial ferrites available depending on the application, tolerable losses, and the required upper frequency of operation. MnZn ferrites, for example, have high initial permeabilities (*e.g.*, 2×10^3) but are only useful up to about 1 MHz, whereas NiZn ferrites have lower initial permeability (*e.g.*, 10^2) but can be used up to 200 MHz. Generally, the initial permeability in the high-frequency region decreases with frequency.

Garnets are ferrimagnetic materials that are typically used at the highest frequencies that cover the microwave range (1–300 GHz). The yttrium iron garnet, YIG, which is $Y_3Fe_5O_{12}$, is one of the simplest garnets with a very low hysteresis loss at microwave frequencies. Garnets have excellent dielectric properties with high resistivities and hence low losses. The main disadvantages are the low saturation magnetization, which is 0.18 T for YIG, and low Curie temperature, 280 °C for YIG. The compositions of garnets depend on the properties required for the particular microwave application. For example, $Y_{2.1}Gd_{0.98}Fe_5O_{12}$ is a garnet that is used in X-band (8–12 GHz) three-port circulators handling high microwave powers (*e.g.*, peak power 200 kW and average power 200 W).

EXAMPLE 8.6

AN INDUCTOR WITH A FERRITE CORE Consider a toroidal coil with a ferrite core. Suppose that the coil has 200 turns and is used in HF work with small signals. The mean diameter of the toroid is 2.5 cm and the core diameter is 0.5 cm. If the core is a MnZn ferrite, what is the approximate inductance of the coil?

SOLUTION

The inductance L of a toroidal coil is given by

$$L = \frac{\mu_{ri}\mu_o N^2 A}{\ell}$$

so

$$L = \frac{(2 \times 10^3)(4\pi \times 10^{-7}\,\mathrm{H\,m^{-1}})(200)^2 \pi \left(\frac{0.005}{2}\,\mathrm{m}\right)^2}{(\pi\, 0.025\,\mathrm{m})} = 0.025\ \mathrm{H} \quad \text{or} \quad 25\ \mathrm{mH}$$

Had the core been air, the inductance would have been 1.26×10^{-5} H or 12.6 μH. The main assumption is that B is uniform in the core, and this will be only so if the diameter of the toroid (2.5 cm) is much greater than the core diameter (0.5 cm). Here this ratio is 5 and the calculation is only approximate.

8.8 HARD MAGNETIC MATERIALS: EXAMPLES AND USES

An ideal hard magnetic material, as summarized in Table 8.6, has very large coercivity and remanent magnetic field. Further, since they are used as permanent magnets, the energy stored per unit volume in the external magnetic field should be as large as possible since this is the energy available to do work. This energy density (J m^{-3}) in the external field depends on the maximum value of the product BH in the second quadrant of the B–H characteristics and is denoted as $(BH)_{max}$. It corresponds to the largest rectangular area that fits the B–H curve in the second quadrant, as shown in Figure 8.39.

When the size of a ferromagnetic sample falls below a certain critical dimension, of the order of 0.1 μm for cobalt, the whole sample becomes a single domain, as depicted in Figure 8.40, because the cost of energy in generating a domain wall is too high compared with the reduction in external magnetostatic energy. These small particle-like pieces of magnets are called **single domain fine particles.** Their magnetic

Table 8.6 Hard magnetic materials and typical values

Magnetic Material	$\mu_0 H_c$ (T)	B_r (T)	$(BH)_{max}$ (kJ m^{-3})	Examples and Uses
Ideal hard	Large	Large	Large	Permanent magnets in various applications.
Alnico (Fe–Al–Ni–Co–Cu)	0.19	0.9	50	Wide range of permanent magnet applications.
Alnico (Columnar)	0.075	1.35	60	
Strontium ferrite (anisotropic)	0.3–0.4	0.36–0.43	24–34	Starter motors, dc motors, loudspeakers, telephone receivers, various toys.
Rare earth cobalt, *e.g.*, Sm_2Co_{17} (sintered)	0.62–1.1	1.1	150–240	Servo motors, stepper motors, couplings, clutches, quality audio headphones.
NdFeB magnets	0.9–1.0	1.0–1.2	200–275	Wide range of applications, small motors (*e.g.*, in hand tools), walkman equipment, CD motors, MRI body scanners, computer applications.
Hard particles, γ-Fe_2O_3	0.03	0.2		Audio and video tapes, floppy disks.

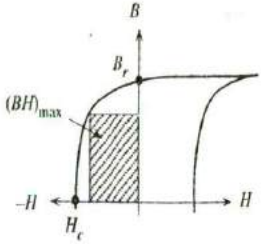


Figure 8.39 Hard magnetic materials and $(BH)_{max}$.

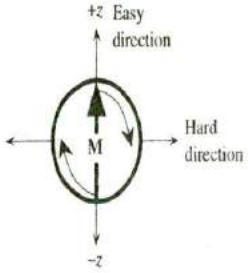


Figure 8.40 A single domain fine particle.

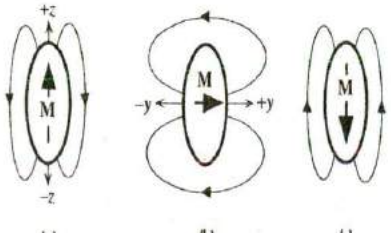


Figure 8.41 A single domain elongated particle. Due to shape anisotropy, magnetization prefers to be along the long axis as in (a). Work has to be done to change **M** from (a) to (b) to (c).

properties depend not only on the crystal structure of the particle but also on the shape of the particle because different shapes give rise to different external magnetic fields. For a spherical iron particle, the magnetization **M** will be in an easy direction, for example, along [100] taken along $+z$. To reverse the magnetization from $+z$ to $-z$ by an applied field, we have to rotate the spins around past the hard direction, as shown in Figure 8.40, since we cannot generate reverse domains (or move domain walls). The rotation of magnetization involves substantial work due to the magnetocrystalline anisotropic energy, and the result is high coercivity. The higher the magnetocrystalline anisotropy energy, the greater the coercivity. The energy involved in creating a domain wall increases with the magnetocrystalline anisotropy energy. The critical size below which a particle becomes a single domain therefore increases with the crystalline anisotropy. Barium ferrite crystals have the hexagonal structure and hence have a high degree of magnetocrystalline anisotropy. Critical size for single domain barium ferrite particles is about 1–1.5 μm, and the coercivity $\mu_o H_c$ of small particles can be as high as 0.3 T, compared with values 0.02–0.1 T in multidomain barium ferrite pieces.

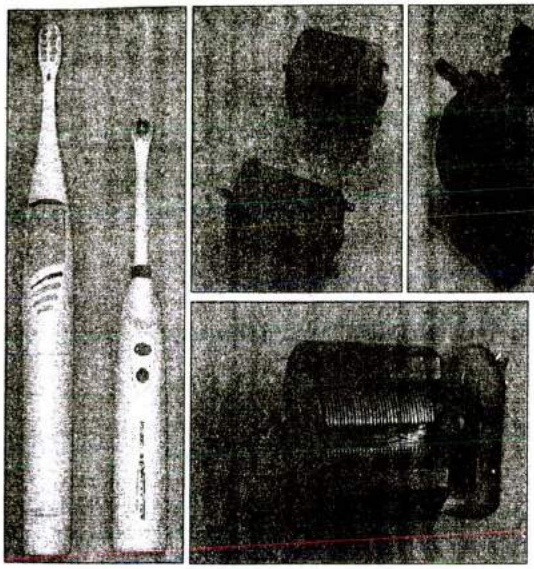
Particles that are not spherical may even have higher coercivity as a result of shape anisotropy. Consider an ellipsoid (elongated) fine particle, shown in Figure 8.41a. If the magnetization **M** is along the long axis (along z), then the potential energy in the external magnetic field is less than if **M** were along the minor axis (along y), as compared in Figure 8.41a and b. Thus, we have to do work to rotate **M** from the long to the short axis, or from Figure 8.41a to b. An elongated fine particle therefore has its magnetization along its length, and the effect is called **shape anisotropy.** If we have to

reverse the magnetization from $+z$ to $-z$ by applying a reverse field, then we can only do so by rotating the magnetization, as shown in Figure 8.41a to c. **M** has to be rotated around through the minor axis, and this involves substantial work. Thus the coercivity is high. In general, the greater the elongation of the particle with respect to its width, the higher the coercivity. Small spherical Fe–Cr–Co particles have a coercivity $\mu_o H_c$ at most 0.02 T, but elongated and aligned particles can have a coercivity as high as 0.075 T due to shape anisotropy.

High coercivity magnets can be fabricated by having elongated fine particles dispersed by precipitation in a structure. Fine particles will be single domains. Alnico is a popular permanent magnet material that is an alloy of the metals Al, Ni, Co, and Fe (hence the name). Its microstructure consists of fine elongated Fe–Co rich particles, called the α'-phase, dispersed in a matrix that is Ni–Al rich and called the α-phase. The structure is obtained by an appropriate heat treatment that allows fine α' particles to precipitate out from a solid solution of the alloy. The α' particles are strongly magnetic, whereas the α-phase matrix is weakly magnetic. When the heat treatment is carried out in the presence of a strong applied magnetic field, the α' particles that are formed have their elongations (or lengths) and hence their magnetizations along the applied field. The demagnetization process requires the rotations of the magnetizations in single domain elongated α' particles, which is a difficult process (shape anisotropy), and hence the coercivity is high. The main drawback of the Alnico magnet is that the alloy is mechanically hard and brittle and cannot be shaped except by casting or sintering before heat treatment. There are, however, other alloy permanent magnets that can be machined.

A variety of permanent magnets are made by compacting high-coercivity particles by using powder metallurgy (*e.g.*, powder pressing or sintering). The particles are magnetically hard because they are sufficiently small for each to be of single domain or they possess substantial shape anisotropy (elongated particles may be ferromagnetic alloys, *e.g.*, Fe–Co, or various hard ferrites). These are generically called powdered solid permanent magnets. An important class is the **ceramic magnets** that are made by compacting barium ferrite, $BaFe_{12}O_9$, or strontium ferrite, $SrFe_{12}O_9$, particles. The barium ferrite has the hexagonal crystal structure with a large magnetocrystalline anisotropy, which means that barium ferrite particles have high coercivity. The ceramic magnet is typically formed by wet pressing ferrite powder in the presence of a magnetizing field, which allows the easy directions of the particles to be aligned, and then drying and carefully sintering the ceramic. They are used in many low-cost applications.

Rare earth cobalt permanent magnets based on samarium–cobalt (Sm–Co) alloys have very high $(BH)_{max}$ values and are widely used in many applications such as dc motors, stepper and servo motors, traveling wave tubes, klystrons, and gyroscopes. The intermetallic compound $SmCo_5$ has a hexagonal crystal structure with high magnetocrystalline anisotropy and hence high coercivity. The $SmCo_5$ powder is pressed in the presence of an applied magnetic field to align the magnetizations of the particles. This is followed by careful sintering to produce a solid powder magnet. The Sm_2Co_{15} magnets are more recent and have particularly high values of $(BH)_{max}$ up to about 240 kJ m^{-3}. Sm_2Co_{15} is actually a generic name and the alloy may contain other transition metals substituting for some of the Co atoms.



One important application of permanent magnets is in small dc motors. Toothbrushes that operate from batteries use dc motors with strong permanent magnets to get the required torque to drive the brushes.

The more recent **neodymium–iron–boron,** NdFeB, powdered solid magnets can have very large $(HB)_{max}$ values up to about 275 kJ m^{-3}. The tetragonal crystal structure has the easy direction along the long axis and possesses high magnetocrystalline anisotropic energy. This means that we need a substantial amount of work to rotate the magnetization around through the hard direction, and hence the coercivity is also high. The main drawback is the lower Curie temperature, typically around 300 °C, whereas for Alnico and rare earth cobalt magnets, the Curie temperatures are above 700 °C. Another method of preparing NdFeB magnets is by the recrystallization of amorphous NdFeB at an elevated temperature in an applied field. The grains in the recrystallized structure are sufficiently small to be single domain grains and therefore possess high coercivity.

EXAMPLE 8.7

$(BH)_{max}$ FOR A PERMANENT MAGNET Consider the permanent magnet in Figure 8.42. There is a small air gap of length ℓ_g where there is an external magnetic field that is available to do work. For example, if we were to insert an appropriate coil in the gap and pass a current through the coil, it would rotate as in a moving coil panel meter. Show that the magnetic energy per unit volume stored in the gap is proportional to the maximum value of BH. How does $(BH)_{max}$ vary with the magnetizing field?

SOLUTION

Let ℓ_m be the mean length of the magnet from one end to the other, as shown in Figure 8.42. We assume that the cross-sectional area A is constant throughout. There are no windings around the magnet and no current, $I = 0$. Ampere's law for H involves integrating H along a closed path or around the mean path length $\ell_m + \ell_g$. Suppose that H_m and H_g are the magnetic

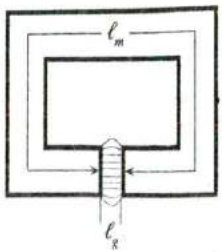


Figure 8.42 A permanent magnet with a small air gap.

field intensities in the permanent magnet and in the gap, respectively. Then $H\,d\ell$ integrated around $\ell_m + \ell_g$ is

$$\oint H\,d\ell = H_m\ell_m + H_g\ell_g = 0$$

so that

$$H_g = -H_m\frac{\ell_m}{\ell_g}$$

and hence

B–H for air gap

$$B_g = -\mu_o\frac{\ell_m}{\ell_g}H_m \qquad [8.24]$$

Equation 8.24 is a relationship between B_g in the gap and H_m in the magnet. In addition, we have the B–H relationship for the magnetic material itself between the magnetic field B_m and intensity H_m in the magnet, that is,

B–H for magnet material

$$B_m = f(H_m) \qquad [8.25]$$

The magnetic flux in the magnet and in the air gap must be continuous. Since we assumed a uniform cross-sectional area, the continuity of flux across the air gap implies that $B_m = B_g$. Thus we need to equate Equation 8.24 to Equation 8.25. Equation 8.24 is a straight line with a negative slope in a B_g versus H_m plot, as shown in Figure 8.43a. Equation 8.25 is, of course, the B–H characteristics of the material. The two intersect at point P, as shown in Figure 8.43a, where $B_g = B_m = B'_m$ and $H_m = H'_m$.

We know that there is magnetic energy in the air gap given by

$$\begin{aligned} E_{mag} &= (\text{Gap volume})(\text{Magnetic energy density in the gap}) \\ &= (A\ell_g)\left(\frac{1}{2}B_gH_g\right) = \frac{1}{2}(A\ell_g)B'_mH'_m\left(\frac{\ell_m}{\ell_g}\right) \\ &= \frac{1}{2}(A\ell_m)B'_mH'_m \\ &= \frac{1}{2}(\text{Magnet volume})B'_mH'_m \end{aligned}$$

Energy in air gap of a magnet

$$[8.26]$$

Thus, the external magnetic energy depends on the magnet volume and the *product* of B'_m and H'_m of the magnet characteristics at the operating point P. For a given magnet size, the magnetic energy in the gap is proportional to the rectangular area $B'_mH'_m$, $OB'_mPH'_m$ in Figure 8.43a,

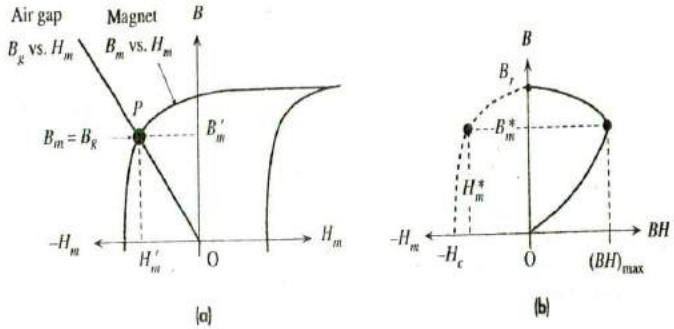


Figure 8.43

(a) Point P represents the operating point of the magnet and determines the magnetic field inside and outside the magnet.

(b) Energy density in the gap is proportional to BH, and for a given geometry and size of gap, this is a maximum at a particular magnetic field B^*_m or B^*_g.

and we have to maximize this area for the best energy extraction. Figure 8.43b shows how the product BH varies with B in a typical magnetic material. BH is maximum at $(BH)_{max}$, when the magnetic field is B^*_m and the field intensity is H^*_m. We can appropriately choose the air-gap size to operate at these values, in which case we will be only limited by the $(BH)_{max}$ available for that magnetic material. It is clear that $(BH)_{max}$ is a good figure of merit for comparing hard magnetic materials. According to Table 8.6, we can extract four to five times more work from a rare earth cobalt magnet than from an Alnico magnet of the same size if we were not limited by economics and weight. It should be mentioned that Equation 8.26 is only approximate as it neglects all fringe fields.

8.9 SUPERCONDUCTIVITY

8.9.1 ZERO RESISTANCE AND THE MEISSNER EFFECT

In 1911 Kamerlingh Onnes at the University of Leiden in Holland observed that when a sample of mercury is cooled to below 4.2 K, its resistivity totally vanishes and the material behaves as a **superconductor,** exhibiting no resistance to current flow. Other experiments since then have shown that there are many such substances, not simply metals, that exhibit superconductivity when cooled below a **critical temperature** T_c that depends on the material. On the other hand, there are also many conductors, including some with the highest conductivities such as silver, gold, and copper, that do not exhibit superconductivity. The resistivity of these **normal conductors** at low temperatures is limited by scattering from impurities and crystal defects and saturates at a finite value determined by the residual resistivity. The two distinctly different types of behavior are depicted in Figure 8.44. Between 1911 and 1986, many different metals and metal alloys had been studied, and the highest

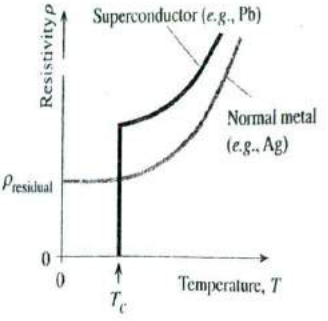
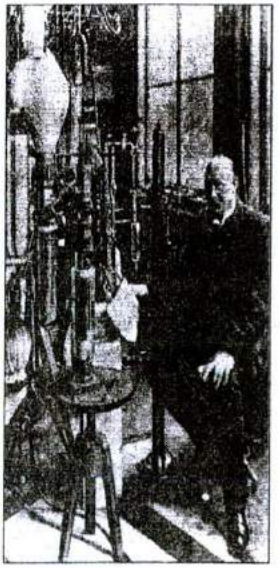


Figure 8.44 A superconductor such as lead evinces a transition to zero resistivity at a critical temperature T_c (7.2 K for Pb).
A normal conductor such as silver exhibits residual resistivity down to lowest temperatures.

recorded critical temperature was about 23 K in a niobium–germanium compound (Nb_3Ge) whose superconductivity was discovered in the early 1970s. In 1986 Bednorz and Müller, at IBM Research Laboratories in Zürich, discovered that a copper oxide–based ceramic-type compound La–Ba–Cu–O, which normally has high resistivity, becomes superconducting when cooled below 35 K. Following this Nobel prize–winning discovery, a variety of copper oxide–based compounds (called cuprate ceramics) have been synthesized and studied. In 1987 it was found that yttrium barium copper oxide (Y–Ba–Cu–O) becomes superconducting at a critical temperature of 95 K, which is above the boiling point of nitrogen (77 K). This discovery was particularly significant because liquid nitrogen is an inexpensive cryogent that is readily liquified and easy to use compared with cryogent liquids that had to be used in the

Superconductivity, zero resistance below a certain critical temperature, was discovered by a Dutch physicist, Heike Kamerlingh Onnes, in 1911. Kamerlingh Onnes and one of his graduate students found that the resistance of frozen mercury simply vanished at 4.15 K; Kamerlingh Onnes won the Nobel prize in 1913.

SOURCE: © Rijksmuseum voor de Geschiedenis der Natuurwetenschappen, courtesy AIP Emilio Segrè Visual Archives.





John Bardeen, Leon N. Cooper, and John Robert Schrieffer, in Nobel prize ceremony (1972). They received the Nobel prize for the explanation of superconductivity in terms of Cooper pairs.

| SOURCE: AIP Emilio Segrè Visual Archives.

"My belief is that the pairing condensation is what Mother Nature had in mind when she created these fascinating high-T_c systems." Robert Schrieffer (1991)

past (liquid helium). At present the highest critical temperature for a superconductor is around 130 K (−143 °C) for Hg–Ba–Ca–Cu–O. These superconductors with T_c above ~30 K are now typically referred as **high-T_c superconductors.** The quest for a near-room-temperature superconductor goes on, with many scientists around the world trying different materials, or synthesizing them, to raise T_c even higher. There are already commercial devices utilizing high-T_c superconductors, for example, thin-film SQUIDs[7] that can accurately measure very small magnetic fluxes, high-Q filters, and resonant cavities in microwave communications.

The vanishing of resistivity is not the only characteristic of a superconductor. A superconductor cannot be viewed simply as a substance that has infinite conductivity below its critical temperature. A superconductor below its critical temperature expels all the magnetic field from the bulk of the sample as if it were a perfectly diamagnetic substance. This phenomenon is known as the **Meissner effect.** Suppose that we place a superconducting material in a magnetic field above T_c. The magnetic field lines will penetrate the sample, as we expect for any low μ_r medium. However, when the superconductor is cooled below T_c, it rejects all the magnetic flux in the sample, as depicted in Figure 8.45. The superconductor develops a magnetization M by developing surface currents, such that M and the applied field cancel everywhere inside

[7] SQUID is a superconducting quantum interference device that can detect very small magnetic fluxes.

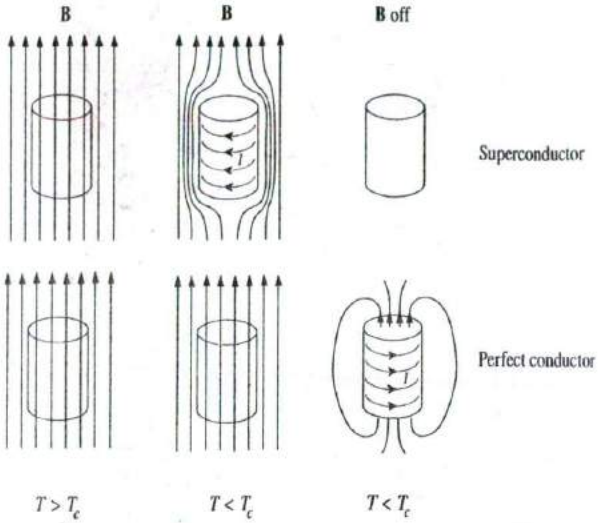
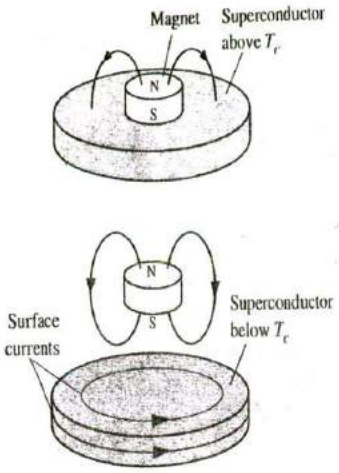


Figure 8.45 The Meissner effect.
A superconductor cooled below its critical temperature expels all magnetic field lines from the bulk by setting up a surface current. A perfect conductor ($\sigma = \infty$) shows no Meissner effect.

the sample. Put differently $\mu_o M$ is in the *opposite* direction to the applied field and equal to it in magnitude. Thus, below T_c a superconductor is a perfectly diamagnetic substance ($\chi_m = -1$). This should be contrasted with the behavior of a perfect conductor, which only exhibits infinite conductivity, or $\rho = 0$, below T_c. If we place a perfect conductor in a magnetic field and then cool it below T_c, the magnetic field is not rejected. These two types of behavior are identified in Figure 8.45. If we switch off the field, the field around the superconductor simply disappears. But switching off the field means there is a decreasing applied field. This change in the field induces currents in the perfect conductor by virtue of Faraday's law of induction. These currents generate a magnetic field that opposes the change (Lenz's law); in other words, they generate a field along the same direction as the applied field to reenforce the decreasing field. As the current can be sustained ($\rho = 0$) without Joule dissipation, it keeps on flowing and maintaining the magnetic field. The two final situations are shown in Figure 8.45 and distinguish the Meissner effect, a distinct characteristic of a superconductor, from the behavior of a perfect conductor ($\rho = 0$ only). The photograph showing the levitation of a magnet above the surface of a superconductor (Figure 8.46) is the direct result of the Meissner effect: the exclusion of the magnet's magnetic fields from the interior of the superconductor.

The transition from the normal state to the superconducting state as the temperature falls below the critical temperature has similarities with phase transitions such as solid to liquid or liquid to vapor changes. At the critical temperature, there is a sharp change in the heat capacity as one would observe for any phase change. In the superconducting



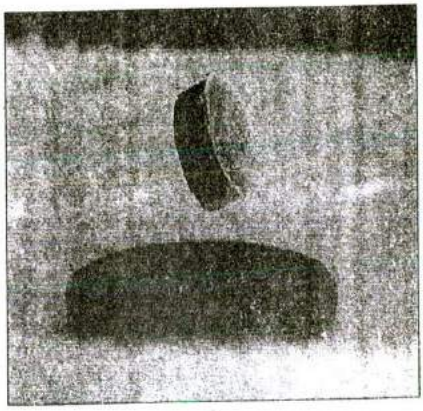


Figure 8.46

Left: A magnet over a superconductor becomes levitated. The superconductor is a perfect diamagnet which means that there can be no magnetic field inside the superconductor.

Right: Photograph of a magnet levitating above a superconductor immersed in liquid nitrogen (77 K). This is the Meissner effect.

| SOURCE: Photo courtesy of Professor Paul C. W. Chu.

state, we cannot treat a conduction electron in isolation. The electrons behave collectively and thereby impart the superconducting characteristics to the substance, as discussed later.

8.9.2 TYPE I AND TYPE II SUPERCONDUCTORS

The superconductivity below the critical temperature has been observed to disappear in the presence of an applied magnetic field exceeding a critical value denoted by B_c. This critical field depends on the temperature and is a characteristic of the material. Figure 8.47 shows the dependence of the critical field on the temperature. The critical field is maximum, $B_c(0)$, when $T = 0$ K (obtained by extrapolation[8]). As long as the applied field is below B_c at that temperature, the material is in the superconducting state, but when the field exceeds B_c, the material reverts to the normal state. We know that in the superconducting state, the applied magnetic field lines are expelled from the sample and the phenomenon is called the Meissner effect. The external field, in fact, does penetrate the sample from the surface into the bulk, but the magnitude of this penetrating field decreases exponentially from the surface. If the field at the surface of the sample is B_0, then at a distance x from the surface, the field is

[8] There is a third law to thermodynamics that is not as emphasized as the first two laws, which dominate all branches of engineering. That is, one can never reach the absolute zero of temperature.

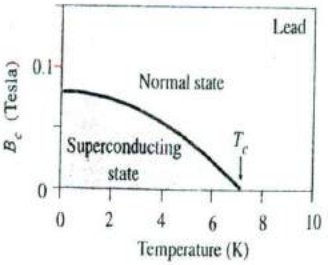


Figure 8.47 The critical field versus temperature in Type I superconductors.

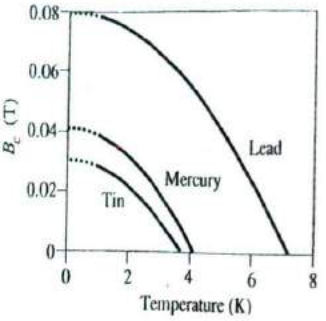


Figure 8.48 The critical field versus temperature in three examples of Type I superconductors.

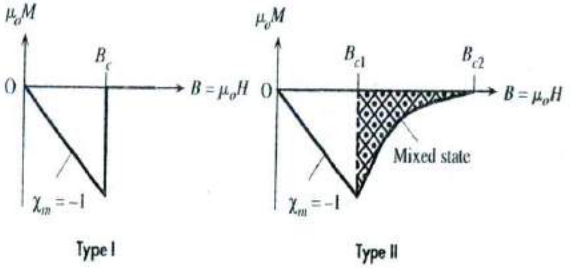


Figure 8.49 Characteristics of Type I and Type II superconductors. $B = \mu_o H$ is the applied field and M is the overall magnetization of the sample. Field inside the sample, $B_{inside} = \mu_o H + \mu_o M$, which is zero only for $B < B_c$ (Type I) and $B < B_{c1}$ (Type II).

given by an exponential decay,

$$B(x) = B_o \exp\left(-\frac{x}{\lambda}\right)$$

where λ is a "characteristic length" of penetration, called the **penetration depth,** and depends on the temperature and T_c (or the material). At the critical temperature, the penetration length is infinite and any magnetic field can penetrate the sample and destroy the superconducting state. Near absolute zero of temperature, however, typical penetration depths are 10–100 nm. Figure 8.48 shows the B_c versus T behavior for three example superconductors, tin, mercury, and lead.

Superconductors are classified into two types, called Type I and Type II, based on their diamagnetic properties. In Type I superconductors, as the applied magnetic field B increases, so does the opposing magnetization M until the field reaches the critical field B_c, whereupon the superconductivity disappears. At that point, the perfect diamagnetic behavior, the Meissner effect, is lost, as illustrated in Figure 8.49. A Type I superconductor below B_c is in the **Meissner state,** where it excludes all the magnetic

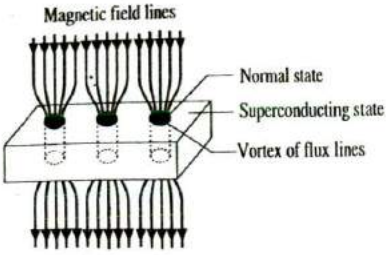


Figure 8.50 The mixed or vortex state in a Type II superconductor.

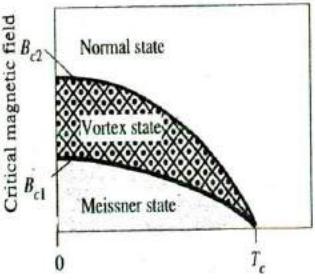


Figure 8.51 Temperature dependence of B_{c1} and B_{c2}.

flux from the interior of the sample. Above B_c it is in the normal state, where the magnetic flux penetrates the sample as it would normally and the conductivity is finite.

In the case of Type II superconductors, the transition does not occur sharply from the Meissner state to the normal state but goes through an intermediate phase in which the applied field is able to pierce through certain local regions of the sample. As the magnetic field increases, initially the sample behaves as a perfect diamagnet exhibiting the Meissner effect and rejecting all the magnetic flux. When the applied field increases beyond a critical field denoted as B_{c1}, the **lower critical field,** the magnetic flux lines are no longer totally expelled from the sample. The overall magnetization M in the sample opposes the field, but its magnitude does not cancel the field everywhere. As the field increases, M gets smaller and more flux lines pierce through the sample until at B_{c2}, the **upper critical field,** all field lines penetrate the sample and superconductivity disappears. This behavior is shown in Figure 8.49. Type II superconductors therefore have two critical fields B_{c1} and B_{c2}.

When the applied field is between B_{c1} and B_{c2}, the field lines pierce through the sample through tubular local regions, as pictured in Figure 8.50. The sample develops local small cylindrical (filamentary) regions of normal state in a matrix of superconducting state and the magnetic flux lines go though these filaments of local normal state, as shown in Figure 8.50. The state between B_{c1} and B_{c2} is called the **mixed state** (or **vortex state**) because there are two states—normal and superconducting—mixed in the same sample. The filaments of normal state have finite conductivity and a quantized amount of flux through them. Each filament is a **vortex** of flux lines (hence the name *vortex state*). It should be apparent that there should be currents circulating around the walls of vortices. These circulating currents ensure that the magnetic flux through the superconducting matrix is zero. The sample overall has infinite conductivity due to the superconducting regions. Figure 8.51 shows the dependence of B_{c1} and B_{c2} on the temperature and identifies the regions of Meissner, mixed, and normal states. All engineering applications of superconductors invariably use Type II materials because B_{c2} is typically much greater than B_c found in Type I materials and, furthermore, the critical temperatures of Type II materials are higher than those of Type I. Many superconductors, including the recent high-T_c superconductors, are of Type II. Table 8.7 summarizes the characteristics of selected Type I and Type II superconductors.

Table 8.7 Examples of Type I and Type II superconductors

Type I	Sn	Hg	Ta	V	Pb	Nb
T_c (K)	3.72	4.15	4.47	5.40	7.19	9.2
B_c (T)	0.030	0.041	0.083	0.14	0.08	0.198

Type II	Nb_3Sn	Nb_3Ge	$Ba_{2-x}Br_xCuO_4$	Y–Ba–Cu–O ($YBa_2Cu_3O_7$)	Bi–Sr–Ca–Cu–O ($Bi_2Sr_2Ca_2Cu_3O_{10}$)	Hg–Ba–Ca–Cu–O
T_c (K)	18.05	23.2	30–35	93–95	122	130–135
B_{c2} (Tesla) at 0 K	24.5	38	~150	~300		
J_c (A cm^{-2}) at 0 K	~10^7			10^4–10^7		

NOTE: Critical fields are close to absolute zero, obtained by extrapolation. Type I for pure, clean elements.

8.3 CRITICAL CURRENT DENSITY

Another important characteristic feature of the superconducting state is that when the current density through the sample exceeds a critical value J_c, it is found that superconductivity disappears. This is not surprising since the current through the superconductor will itself generate a magnetic field and at sufficiently high current densities, the magnetic field at the surface of the sample will exceed the critical field and extinguish superconductivity. This plausible direct relation between B_c and J_c is only true for Type I superconductors, whereas in Type II superconductors, J_c depends in a complicated way on the interaction between the current and the flux vortices. New high-T_c superconductors have exceedingly high critical fields, as apparent in Table 8.7, that do not seem to necessarily translate to high critical current densities. The critical current density in Type II superconductors depends not only on the temperature and the applied magnetic field but also on the preparation and hence the microstructure (*e.g.*, polycrystallinity) of the superconductor material. Critical current densities in new high-T_c superconductors vary widely with preparation conditions. For example, in Y–Ba–Cu–O, J_c may be greater than 10^7 A cm^{-2} in some carefully prepared thin films and single crystals but around 10^3–10^6 A cm^{-2} in some of the polycrystalline bulk material (*e.g.*, sintered bulk samples). In Nb_3Sn, used in superconducting solenoid magnets, on the other hand, J_c is close to 10^7 A cm^{-2} at near 0 K.

The critical current density is important in engineering because it limits the total current that can be passed through a superconducting wire or a device. The limits of superconductivity are therefore defined by the critical temperature T_c, critical magnetic field B_c (or B_{c2}), and critical current density J_c. These constitute a surface in a three-dimensional plot, as shown in Figure 8.52, which separates the superconducting state from the normal state. Any operating point (T_1, B_1, J_1) inside this surface is in the superconducting state. When the cuprate ceramic superconductors were first discovered, their J_c values were too low to allow immediate significant applications in engineering.

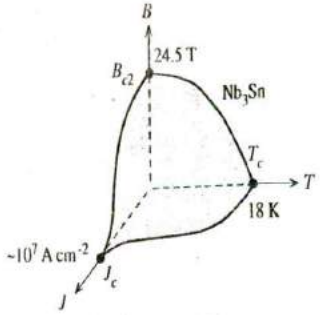


Figure 8.52 The critical surface for a niobium–tin alloy, which is a Type II superconductor.

Their synthesis over the last 10 years has advanced to a level that we can now benefit from large critical currents and fields. Over the same temperature range, ceramic cuprate superconductors now easily outperform the traditional superconductors. There are already a number of applications of these high-T_c superconductors in the commercial market.

EXAMPLE 8.8

SUPERCONDUCTING SOLENOIDS[9] Superconducting solenoid magnets can produce very large magnetic fields up to ~15 T or so, whereas the magnetic fields available from a ferromagnetic core solenoid is limited to ~2 T. High field magnets used in magnetic resonance imaging are based on superconducting solenoids wound using a superconducting wire. They are operated around 4 K with expensive liquid helium as the cryogen. These superconducting wires are typically Nb_3Sn or NbTi alloy filaments embedded in a copper matrix. A very large current, several hundred amperes, is passed through the solenoid winding to obtain the necessary high magnetic fields. There is, of course, no Joule heating once the current is flowing in the superconducting state. The main problem is the large forces and hence stresses in the coil due to large currents. Two wires carrying currents in the opposite direction repel each other, and the force is proportional to I^2. Thus the magnetic forces between the wires of the coil give rise to outward radial forces trying to "blow open" the solenoid, as depicted in Figure 8.53. The forces between neighboring wires are attractive and hence give rise to compressional forces squeezing the solenoid axially. The solenoid has to have a proper mechanical support structure around it to prevent mechanical fracture and failure due to large forces between the windings. The copper matrix serves as mechanical support to cushion against the stresses as well as a good thermal conductor in the event that superconductivity is inadvertently lost during operation.

Suppose that we have a superconducting solenoid that is 10 cm in diameter and 1 m in length and has 500 turns of Nb_3Sn wire, whose critical field B_c at 4.2 K (liquid He temperature) is about 20 T and critical current density J_c is 3×10^6 A cm^{-2}. What is the current necessary to set up a field of 5 T at the center of a solenoid? What is the approximate energy stored in the

[9] Designing a superconducting solenoid is by no means trivial, and the enthusiastic student is referred to a very readable description given by James D. Doss, *Engineer's Guide to High Temperature Superconductivity*, New York: John Wiley & Sons, 1989, ch. 4. Photographs and descriptions of catastrophic failure in high field solenoids can be found in an article by G. Broebinger, A. Passner, and J. Bevk, "Building World-Record Magnets" in *Scientific American*, June 1995, pp. 59–66.

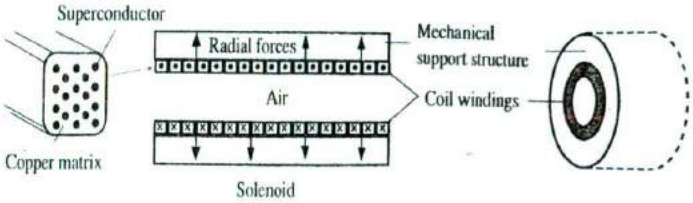
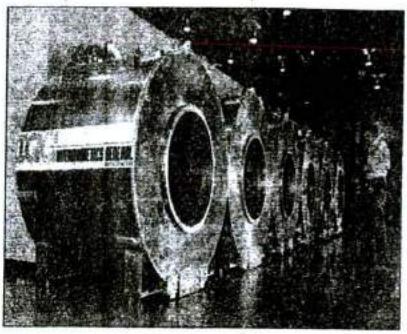


Figure 8.53 A solenoid carrying a current experiences radial forces pushing the coil apart and axial forces compressing the coil.



Superconducting electromagnets used on MRI. Operates with liquid He, providing a magnetic field 0.5–1.5 T.

SOURCE: Courtesy of IGC Magnet Business group.

solenoid? Assume that the critical current density decreases linearly with the applied field. Further, assume also that the field across the diameter of the solenoid is approximately uniform (field at the windings is the same as that at the center).

SOLUTION

We can assume that we have a long solenoid, that is, length (100 cm) ≫ diameter (10 cm). The field at the center of a long solenoid is given by

$$B = \frac{\mu_o N I}{\ell}$$

so the current necessary for $B = 5$ T is

$$I = \frac{B\ell}{\mu_o N} = \frac{(5)(1)}{(4\pi \times 10^{-7})(500)} = 7958 \text{ A} \quad \text{or} \quad 7.96 \text{ kA}$$

As the coil is 1 m and there are 500 turns, the coil wire radius must be 1 mm. If all the cross section of the wire were of superconducting medium, then the corresponding current density would be

$$J_{\text{wire}} = \frac{I}{\pi r^2} = \frac{7958}{\pi (0.001)^2} = 2.5 \times 10^9 \text{ A m}^{-2} \quad \text{or} \quad 2.5 \times 10^5 \text{ A cm}^{-2}$$

The actual current density through the superconductors will be greater than this as the wires are embedded in a metal matrix. Suppose that 20 percent by cross-sectional area (and hence as volume percentage) is the superconductor; then the actual current density through the

superconductor is

$$J_{super} = \frac{J_{wire}}{0.2} = 1.25 \times 10^6 \text{ A cm}^{-2}$$

We now need the critical current density J_c' at a field of 5 T. Assuming J_c decreases linearly with the applied field and vanishes when $B = B_c$, we can find J_c', from linear interpolation

$$J_c' = J_c \frac{B_c - B}{B_c} = (3 \times 10^6 \text{ A cm}^{-2}) \frac{20 \text{ T} - 5 \text{ T}}{20 \text{ T}} = 2.25 \times 10^6 \text{ A cm}^{-2}$$

The actual current density J_{super} through the superconductors is less than this critical value J_c'. We can assume that the superconducting solenoid will operate "safely" (with all other designs correctly implemented). It should be emphasized that accurate and reliable calculations will involve the actual J_c–B_c–T_c surface, as in Figure 8.52 for the given material.

Since the field in the solenoid is $B = 5$ T, assuming that this is uniform along the axis and the core is air, the energy density or energy per unit volume is

$$E_{vol} = \frac{B^2}{2\mu_o} = \frac{5^2}{2(4\pi \times 10^{-7})} = 9.95 \times 10^6 \text{ J m}^{-3}$$

so the total energy

$$E = E_{vol}[\text{volume}] = (9.95 \times 10^6 \text{ J m}^{-3})[(1 \text{ m})(\pi 0.05^2 \text{ m}^2)]$$
$$= 7.81 \times 10^4 \text{ J} \quad \text{or} \quad 78.1 \text{ kJ}$$

If all this energy can be converted to electrical work, it would light a 100 W lamp for 13 min (and if converted to mechanical work, it could lift an 8 ton truck by 1 m).

8.10 SUPERCONDUCTIVITY ORIGIN

Although superconductivity was discovered in 1911, the understanding of its origin did not emerge until 1957 when Bardeen, Cooper, and Schrieffer formulated the theory (called the **BCS theory**) in terms of quantum mechanics. The quantum mechanical treatment is certainly beyond the scope of this book, but one can nonetheless grasp an intuitive understanding, as follows. The cardinal idea is that, at sufficiently low temperatures, two oppositely spinning and oppositely traveling electrons can attract each other indirectly through the deformation of the crystal lattice of positive metal ions. The idea is illustrated pictorially in Figure 8.54. The electron 1 distorts the lattice

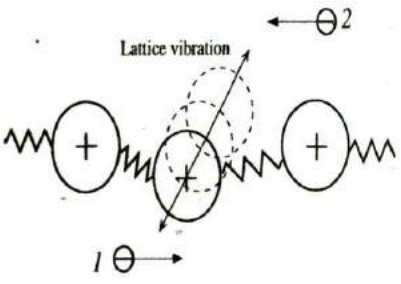


Figure 8.54 A pictorial and intuitive view of an indirect attraction between two oppositely traveling electrons via lattice distortion and vibration.

around it and changes its vibrations as it passes through this region. Random thermal vibrations of the lattice at low temperatures are not strong enough to randomize this induced lattice distortion and vibration. The vibrations of this distorted region now look differently to another electron, 2, passing by. This second electron feels a "net" attractive force due to the slight displacements of positive metal ions from their equilibrium positions. The two electrons interact indirectly through the deformations and vibrations of the lattice of positive ions. This indirect interaction at sufficiently low temperatures is able to overcome the mutual Coulombic repulsion between the electrons and hence bind the two electrons to each other. The two electrons are called a **Cooper pair.** The intuitive diagram in Figure 8.54, of course, does not even convey the intuition why the spins of the electrons should be opposite. The requirement of opposite spins comes from the formal quantum mechanical theory. The net spin of the Cooper pair is zero and their net linear momentum is also zero. There is a further significance to the pairing of electron spins in the Cooper pair. As a quasi-particle, or an entity, the Cooper pair has no net spin and hence the Cooper pairs do not obey the Fermi–Dirac statistics.[10] They can therefore all "condense" to the *lowest energy* state and possess one single wavefunction that can describe the whole collection of Cooper pairs. All the paired electrons are described collectively by a single coherent wavefunction Ψ, which extends over the whole sample. A crystal imperfection cannot simply scatter a single Cooper pair because all the pairs behave as a single entity—like a "huge molecule." Scattering one pair involves scattering all, which is simply not possible. An analogy may help. One can scatter an individual football player running on his own. But if all the team members got together and moved forward arm in arm as a rigid line, then the scattering of any one now is impossible, as the rest will hold him in the line and continue to move forward (don't forget, it's only an analogy!). Superconductivity is said to be a macroscopic manifestation of quantum mechanics. The BCS theory has had good success with traditional superconductors, but there seems to be some doubt about its applicability to the new high-T_c superconductors. There are a number of high-T_c superconductivity theories at present, and the interested student can easily find additional reading on the subject.

ADDITIONAL TOPICS

8.11 ENERGY BAND DIAGRAMS AND MAGNETISM

8.11.1 PAULI SPIN PARAMAGNETISM

Consider a paramagnetic metal such as sodium. The paramagnetism arises from the alignment of the spins of conduction electrons with the applied magnetic field. A conduction electron in a metal has an extended wave function and does not orbit any particular metal ion. The conduction electron's magnetic moment arises from the electron spin alone, and $\boldsymbol{\mu}_{spin}$ is in the opposite direction to the spin; $\boldsymbol{\mu}_{spin}$ can be either up

[10] In fact, the Cooper pair without a net spin behaves as if is were a **boson** particle.

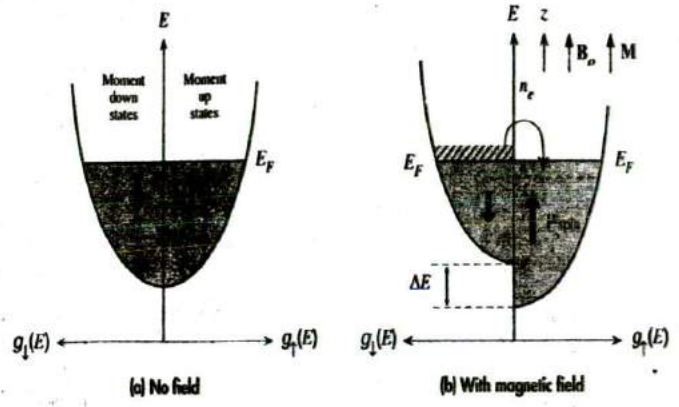


Figure 8.55 Pauli spin paramagnetism in metals due to conduction electrons.

($m_s = -\frac{1}{2}$) or down ($m_s = +\frac{1}{2}$). In the absence of a magnetic field, the energies of magnetic moment up and down states (or wavefunctions) are the same and there are as many electrons with magnetic moment up as there are with magnetic moment down. Figure 8.55a shows the density of states (number of states per unit energy per unit volume) for states with magnetic moment up (↑), denoted as $g_\uparrow(E)$, and for states with magnetic moment down (↓), denoted as $g_\downarrow(E)$. Both states have the same energy and both are equally occupied. All energy levels up to the Fermi energy E_F are occupied as shown in Figure 8.55a. Effectively we are viewing the energy band of the metal as two subbands corresponding to magnetic moment up and down bands. The bands overlap in the absence of a field and are indistinguishable.

Consider what happens in the presence of an applied field B_a along the z direction. If a conduction electron's magnetic moment μ_z is *along* the field (aligned with the field), then it has a lower potential energy. Thus, those electron wavefunctions with a magnetic moment up have lower energy, whereas those wavefunctions with a magnetic moment down have higher energy. In the presence of a field B_a, therefore, all states with magnetic moment up, and hence $g_\uparrow(E)$, are lowered in energy by βB_a where β is the Bohr magneton. All states with magnetic moment down, and hence $g_\downarrow(E)$, are raised by βB_a. Both shifts are shown in Figure 8.55b. Those electrons with magnetic moment down near E_F in the $g_\downarrow(E)$ band can now find lower energy states in the $g_\uparrow(E)$ band and hence flip their spins and transfer to the $g_\uparrow(E)$ band. There are now more electrons in states with magnetic moment up in the $g_\uparrow(E)$ band than in the $g_\downarrow(E)$ band. When averaged over all conduction electrons there is now a net magnetic moment per conduction electron along the z direction or the applied field.

To find the net magnetic moment per conduction electron we have to find how many electrons transfer from the $g_\downarrow(E)$ band to the $g_\uparrow(E)$ band. The energy separation ΔE between the magnetic moment down and up states is $2\beta B_a$. All electrons, n_e per unit volume, in the $g_\downarrow(E)$ band around E_F within an energy range $\frac{1}{2}\Delta E$ transfer to the $g_\uparrow(E)$ band. ΔE is small, so n_e is approximately $g_\downarrow(E_F)(\frac{1}{2}\Delta E)$ or $\frac{1}{2}\, g(E_F)(\frac{1}{2}\Delta E)$ because $g(E_F)$ includes states with spin up and down, that is, $\frac{1}{2}\, g(E_F) = g_\downarrow(E_F)$. The magnetic

moment down band decreases by n_ϵ and the magnetic moment up band increases by n_ϵ and the net magnetic moment per unit volume is

$$M \approx 2n_\epsilon\mu_\epsilon = 2\left[\tfrac{1}{2}g(E_F)\left(\tfrac{1}{2}\Delta E\right)\right]\beta$$
$$= 2\left[\tfrac{1}{2}g(E_F)\left(\tfrac{1}{2}2\beta B_a\right)\right]\beta = \beta^2 g(E_F)B_a$$

Pauli spin paramagnetism

Using $B_a = \mu_a H$ and the definition $\chi_m = M/H$, the paramagnetic susceptibility is

$$\chi_{para} \approx \mu_a\beta^2 g(E_F)$$

We see that the density of states at the Fermi level determines the susceptibility.

PAULI SPIN PARAMAGNETISM OF SODIUM The Fermi energy of sodium, E_F, is 3.15 eV. Using the density of states $g(E)$ expression for the free conduction electrons in a metal, evaluate the paramagnetic susceptibility of sodium and compare with the experimental value of 9.1×10^{-6}.

SOLUTION

The density of states $g(E)$ in the free electron model is

$$g(E) = (8\pi 2^{1/2})\left(\frac{m_e}{h^2}\right)^{3/2} E^{1/2}$$

We have to evaluate $g(E)$ at the Fermi energy $E = E_F = 3.15$ eV,

$$g(E_F) = (8\pi 2^{1/2})\left(\frac{9.1 \times 10^{-31}}{(6.626 \times 10^{-34})^2}\right)^{3/2}(3.15 \times 1.6 \times 10^{-19})^{1/2} = 7.54 \times 10^{46}\ \text{J}^{-1}\,\text{m}^{-3}$$

Paramagnetic susceptibility is

$$\chi_{para} = \mu_a\beta^2 g(E_F) = (4\pi \times 10^{-7})(9.27 \times 10^{-24})^2(7.54 \times 10^{46}) = 8.16 \times 10^{-6}$$

We need to subtract the diamagnetic from the calculated paramagnetic susceptibility to obtain the net susceptibility, which would decrease the calculated value slightly. Nonetheless, given the approximate nature of the theory, the calculated value is not far out from the measured value.

8.11.2 ENERGY BAND MODEL OF FERROMAGNETISM

The energy band model of paramagnetism can be extended to explain ferromagnetism. Once we start using the energy band model, we are essentially assigning all the valence (outer shell) electrons of the atoms to a collective sharing among *all* the atoms; they no longer belong to their individual parents. These valence electrons now belong to the whole crystal. (The model is also known as the *itinerant electron model*.)

Recall that in a ferromagnetic crystal there is an internal magnetization, even in the absence of an applied field, due to a net number of unpaired spins; that is, overall, the crystal has more electrons with spins up than with spins down. The reason is the exchange energy, which causes the spin magnetic moments of two electrons to line up parallel to each other so that their energy is lowered in much the same way as Hund's

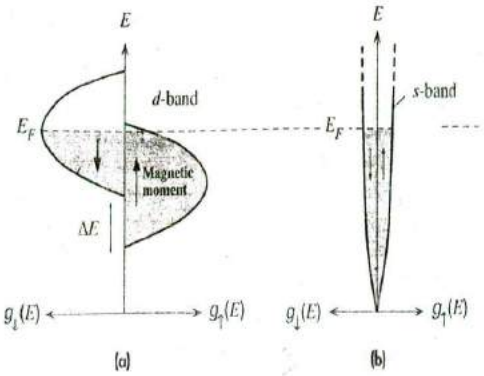


Figure 8.56 Energy band model of ferromagnetism.
(a) The split *d*-band.
(b) The *s*-band is not affected. The arrows in the bands are spin magnetic moments.

rule works within an atom. In magnetic metals such as Fe, Ni, and Co, there are two bands of interest, the *s*-band and the *d*-band. The two bands overlap but the *s*-band is much wider. We can represent the density of states for magnetic moment up and magnetic moment down states separately. Consider the *d*-band. The density of states $g_\uparrow(E)$ for magnetic moment up states is lowered by ΔE with respect to the density of states $g_\downarrow(E)$ for magnetic moment down states due to the exchange energy as shown in Figure 8.56a. The energy lowering ΔE for the *s*-band can be neglected as in Figure 8.56b. All the states up to the Fermi energy are occupied. For Fe, the *d*-band magnetic moment up states are filled almost to the top of the band (this band is 96 percent full), and magnetic moment down states are filled roughly halfway. Thus, there are many more electrons with moments up than moments down; put differently there are many electrons that have aligned their spins. The spin magnetic moment alignment of electrons is exactly what is needed to generate a net magnetization. (In some books, the spin magnetic moment down band is sketched lower than the spin magnetic moment up band in contrast to Figure 8.56a. Both sketches are correct since both would also result in a net number of electrons having their spins in parallel, and hence a net magnetization within the crystal. Another way to look at it is to realize that there are two bands: one band for the "majority of spins," and another band for the "minority spins.")

The *s*-band is filled up to E_F, and there are almost equal numbers of electrons with up and down moments in this band. The ferromagnetic effect arises from the behavior of electrons mainly in the *d*-band. Electrical conduction, on the other hand, is determined by electrons in the *s*-band. The reason is that the *s*-band is very wide compared with the *d*-band, and the electron effective mass in the *s*-band is very small. Thus, electrons have a much higher mobility in the *s*-band than in the *d*-band. When an *s*-electron is scattered (by phonons, impurities, defects, etc.) into the *d*-band, it does not make any significant contribution to conduction because the drift mobility is very small in this band. The spin of the electron cannot be flipped easily in a scattering process. An *s*-electron with its moment down can be easily scattered into the empty states in the corresponding moment-down *d*-band (there are many empty states at E_F), but the moment-up electron has no states in the moment-up

d-band into which it can be scattered. Conduction occurs by moment-up electrons; these are the *favored* electrons for conduction.

The band model is particularly useful in explaining the noninteger number of Bohr magnetons that give rise to the ferromagnetism. The isolated Fe atom has six $3d$ and two $4s$ electrons or 8 valence electrons. These electrons in the crystal become shared by all the atoms. If N is the number of atoms per unit volume, then one unit volume of crystal has $8N$ valence electrons. $8N$ electrons enter the s and d bands, filling states starting from the lowest energy.[11] The exact distribution of electrons depends on how many states are available at each energy as electrons fill the bands. We simply summarize the results of the filling process that is shown in Figure 8.56 for Fe:

$0.3N$ electrons in the moment-up s-band (N states available)

$0.3N$ electrons in the moment-down s-band (N states available)

$4.8N$ electrons in the moment-up d-band ($5N$ states available)

$2.6N$ electrons in the moment-down d-band ($5N$ states available)

To find how many electrons have parallel spin magnetic moments, we simply sum the above, which is $2.2N$ moment-up electrons per unit volume or $2.2N$ Bohr magnetons per unit volume, or 2.2 Bohr magnetons per atom. The saturation magnetization M_{sat} is then $(2.2N)\beta$ or 2.2 T. There is therefore a natural explanation for a noninteger number of spins per atom in the band model of ferromagnetism.

8.12 ANISOTROPIC AND GIANT MAGNETORESISTANCE

In general, **magnetoresistance** refers to the change in the resistance of a material (any material) when it is placed in a magnetic field. When a nonmagnetic metal such as copper is placed in a magnetic field, the change in its resistivity, and hence the sample resistance, is so small that it has no real practical use. When a magnetic metal, such as iron, is placed in a magnetic field, the change in the resistivity depends on the direction of the current flow with respect to the magnetic field. The resistivity $\rho_{//}$ for current flow parallel to the magnetic field decreases, and the resistivity $\rho_{\perp}$, perpendicular to the field, increases by roughly the same amount. The change in the resistivity due to the applied magnetic field is *anisotropic* (depends on the direction) and is called **anisotropic magnetoresistance (AMR).** The change in resistivity is limited to a few percent, but, nonetheless, is still useful. The physical origin of this phenomenon is based on the applied field being able to tilt the orbital angular momenta of the $3d$ electrons as shown in Figure 8.57a. The field rotates the $3d$ orbitals, which changes the scattering of the conduction electrons according to their direction of travel; hence $\rho_{//}$ and $\rho_{\perp}$ are different, as shown in Figure 8.57b.

[11] $8N$ is used to emphasize that all these valence electrons belong to the crystal, i.e., $8N \approx 7 \times 10^{24}$ cm^{-3}.

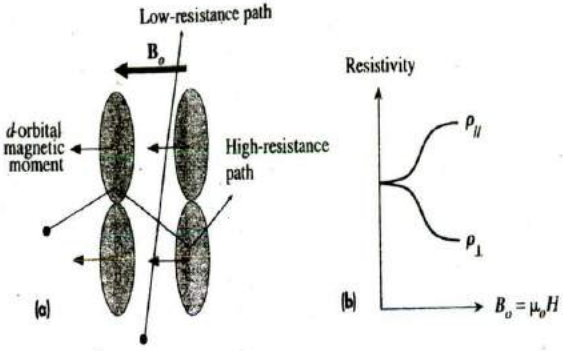


Figure 8.57

(a) The origin of anisotropic magnetoresistance (AMR). The electrons traveling along the field experience more scattering than those traveling perpendicular to the field.

(b) Resistivity depends on the current flow direction with respect to the applied magnetic field.

On the other hand, a very large magnetoresistance, called **giant magnetoresistance (GMR),** has been observed in certain special multilayer structures, which exhibit substantial changes in the resistance (*e.g.*, more than 10 percent) when a magnetic field is applied.[12] Even though GMR is a relatively new discovery (1988), it is already widely used in the read heads of hard disk drives. There are also various magnetic field sensors based on the GMR.

The special multilayer structure in its simplest form has two **ferromagnetic layers** (such as Fe or Co or their alloys, etc.) separated by a nonmagnetic transition metal layer (such as Cu), called the **spacer,** as shown in Figure 8.58a. The magnetic layers are thin (less than 10 nm), and the nonmagnetic layer is even thinner. The magnetizations of the two ferromagnetic layers are not random; they depend on the thickness of the spacer because the two layers are "coupled" indirectly through this thin spacer.[13] In the absence of an external field, two magnetic layers are coupled in such a way that their magnetizations are *antiparallel* or in opposite directions; this arrangement is also called an *antiferromagnetically* coupled configuration. We will use the notation FNA to represent the antiparallel configuration, where N stands for the nonmagnetic metal.

We can apply an external magnetic field to one of the layers and rotate its magnetization so that the two magnetizations are now in parallel as in Figure 8.58c. This parallel configuration is frequently called *ferromagnetically* coupled layers and is denoted as FNF. The two structures have a *giant* difference in their resistances, hence the term giant magnetoresistance. The resistance of the antiparallel FNA in Figure 8.58b structure is much higher than that of the parallel structure FNF in Figure 8.58c.

---

[12] GMR was discovered in the late 1980s by Peter Grünberg (Julich, Germany), and Albert Fert (University of Paris-Sud) and their coworkers. Magnetoresistance itself, however, has been well known, and dates back to Lord Kelvin's experiments in 1857.

[13] The physics of the coupling process between the two magnetic layers is an indirect exchange interaction, the details of which are not needed to understand the basics of the GMR phenomenon.

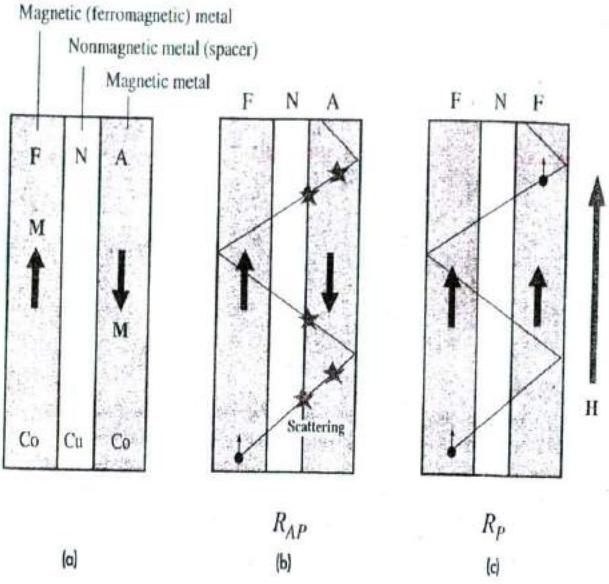


Figure 8.58 A highly simplified view of the principle of the giant magnetoresistance effect.
(a) The basic trilayer structure.
(b) Antiparallel magnetic layers with high resistance R_{AP}.
(c) An external field aligns layers; parallel alignment has a lower resistance R_P.

The current flow through this multilayer structure (whether along or perpendicular to the layers) will involve electrons crossing from one layer to another, passing through the interfaces. Recall that it is the electrons around the Fermi energy that are involved in the conduction and that their mean speed is orders of magnitude larger than the drift velocity. The electron trajectories are therefore not parallel to the current flow (and should not be confused with current flow lines).

Consider the antiparallel FNA structure. The magnetic moment up electron in the first magnetic layer is the *favored* conduction electron; that is, it suffers very little scattering. However, when this moment-up electron arrives at the A layer in which the magnetization is reversed, it finds itself with the wrong spin or wrong moment. It is now an *unfavored* electron and is subject to scattering. Thus, the moment-up electron suffers scattering not only in the bulk of A but, more significantly, as it crosses the N-layer into the A-layer, that is, at the interface as in Figure 8.58b. The antiparallel FNA structure therefore has a high resistance, denoted as R_{AP}. In contrast, when the magnetizations are parallel, the moment-up electron is the favored electron in both the layers and experiences very little scattering. The resistance R_P of this parallel (FNF) structure is smaller than R_{AP} ($R_P < R_{AP}$). The difference in the resistances R_P and R_{AP} in this simple trilayer is roughly 10 percent or less. But, in multilayered structures, which have a series of alternating magnetic and nonmagnetic layers (*e.g.*, 50 or more magnetic and nonmagnetic alternating layers as in FNANFANFA . . .), the change in the

Table 8.8 GMR effect in trilayers and multilayers

Sample	Structure and layer thicknesses	$\Delta R/R_P$ (%)	Temperature (K)
CoFe/CAgCu/CoFe	Trilayer	4-7	300
NiFe/Cu/Co	Trilayer, 10/2.5/2.2 nm (spin valve)	4.6	300
$Co_{90}Fe_{10}/Cu/Co_{90}Fe_{10}$	Trilayer, 4/2.5/0.8 nm (spin valve)	7	300
$[Co/Cu]_{100}$	100 layers of Co/Cu, 1 nm / 1 nm	80	300
$[Co/Co]_{60}$	60 layers Co/Cu, 0.8 nm / 0.83 nm	115	4.2

SOURCE: Data from P. Grünberg, *Sensors and Actuators*, **A91**, 153, 2001.

resistance can be impressively large, exceeding 100 percent at low temperature and 60–80 percent at room temperature.

The GMR effect is often measured by quoting the change in the resistance with respect to R_P,

Giant magnetoresistance effect

$$\left(\frac{\Delta R}{R_P}\right)_{GMR} = \frac{R_{AP} - R_P}{R_P}$$

Further, the magnetoresistance effect can be measured either by passing a current that flows in the plane of the layers or perpendicular to the plane. Most experiments use the first one, in what is known as **current in plane (CIP)** measurements; but the biggest change, however, is observed for currents perpendicular to the plane of the layers. Table 8.8 summarizes typically reported $\Delta R/R_P$ values for the GMR effect in simple trilayers and multilayers.

The structures with antiparallel and parallel magnetic alignments are obviously two extreme cases. If the angle between the magnetization vectors $\mathbf{M}_1$ and $\mathbf{M}_2$ of the two magnetic layers is θ, then the resistance of the structure depends on θ, with the minimum for $\theta = 0$ (FNF) and the maximum for $\theta = 180°$ (FNA) as shown in Figure 8.59. The fractional change in the resistance depends on θ as

GMR and relative magnetizations of magnetic layers

$$\frac{\Delta R}{R_P} = \left(\frac{\Delta R}{R_P}\right)_{max} \frac{1 - \cos\theta}{2}$$

As expected, the change is maximum when $\theta = 180°$.

One of the best applications of GMR is in a **spin valve,** in which the current flow is controlled by an external applied magnetic field. Stated differently, the resistance of the valve is controlled by an applied field. Figure 8.60a shows one possible simple spin valve structure. The magnetization of the Co magnetic layer is fixed, that is, *pinned,* by having this layer next to an antiferromagnetic layer, called the *pinning layer.* The exchange interaction between the ferromagnetic Co layer and the antiferromagnetic CoMn layer effectively pins the direction of the Co layer; it takes an enormous field to change the magnetization of the Co layer. A Cu spacer layer separates the Co and the next magnetic FeNi layer. The FeNi layer is called the *free* layer because its magnetization can be changed by an external magnetic field. Normally, in the absence of a

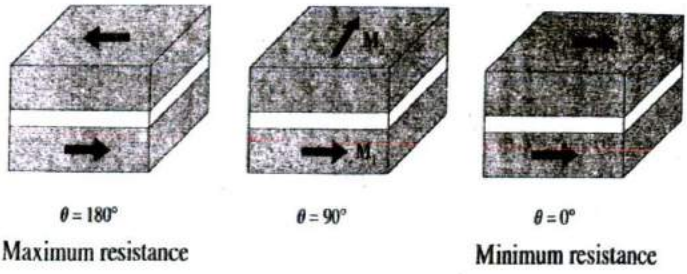


Figure 8.59 Resistance of the multilayer structure depends on the relative orientations of magnetization in the two magnetic layers.

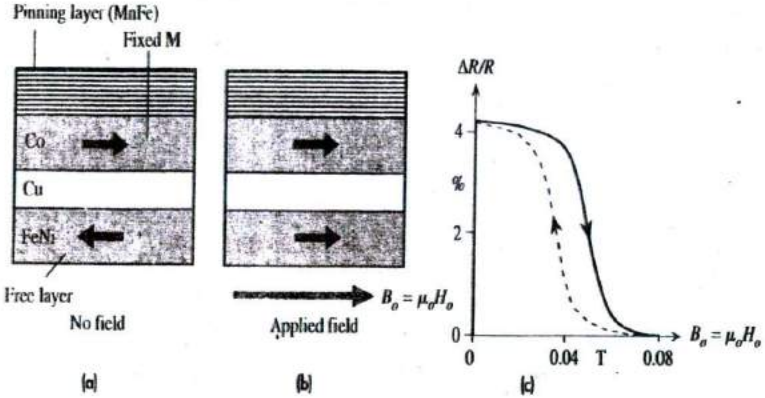


Figure 8.60 Principle of the spin valve.
(a) No applied field.
(b) Applied field has fully oriented the free-layer magnetization.
(c) Resistance change versus applied magnetic field (schematic) for a FeNi/Cu/FeNi spin valve.

field, the magnetization of the FeNi layer is antiparallel to the Co layer, and the structure has a high resistance R_{AP}. An applied external field $B_a = \mu_o H$ can rotate the FeNi layer's magnetization and can easily align FeNi's magnetization fully in parallel with that of Co so that the resistance becomes minimum, *i.e.*, R_P as in Figure 8.60b. It is clear that the external field can be used to control the flow of current through this structure. (The name spin valve reflects the fact that the valve operation relies on the spin of the electrons.) The free layer should be relatively soft to be able to respond to the applied field, whereas the pinned layer should have sufficient coercivity not to lose its magnetization. Figure 8.60c shows a typical magnetoresistance versus applied field characteristics for one particular type of spin valve. The spin valve exhibits hysteresis; that is, the signal ΔR versus H depends on the direction of magnetization as shown in the figure.

8.13 MAGNETIC RECORDING MATERIALS

General Principles of Magnetic Recording Outside electric machinery (mainly rotating machines and transformers), magnetic materials are most widely used in magnetic recording media to store information in either analog or digital form. The deep disappointment of accidently losing valuable stored information on the hard drive of one's computer is well known to most computer users. Magnetic materials in magnetic recording fall into two categories: those used in magnetic heads to record (write), play (read), and erase information, and those used in magnetic media in which the information is stored either permanently or until the next write requirement. The magnetic storage media can be flexible, as in audio and video cassettes and floppy disks, or it can be rigid, as in the hard disk of a computer hard drive. Even though magnetic recording appears in seemingly diverse applications (*e.g.*, audio tape recorders vis-à-vis computer hard drives), the basic principles are nonetheless quite similar.

As a very simple example, we will consider magnetic recording of a signal on an audio tape, as shown schematically in Figure 8.61. The tape is simply a polymer backing tape that has a thin coating of magnetic material on it, as described later. The information is converted into a current signal $i(t)$ that modulates the current around a toroid-type electromagnet with a very small air gap (around 1 μm). This gapped core electromagnet is the **inductive recording head.** The current modulates the magnetic field intensity in the core of the head and hence the field in the gap and around it. The recording of information is achieved by the **fringing magnetic field** around the gap region magnetizing the audio type passing under the head at a constant speed (the tape is usually in contact with the head). As the fringing field changes according to the current signal, so does the magnetization of the audio tape. This means that the electrical signal is stored as a spatial magnetic pattern in the tape. The fringing fields of the recording head modulate the magnetization in the tape in the direction of motion, put differently, along the length of the tape. This type of magnetic information storage is called **longitudinal recording.**

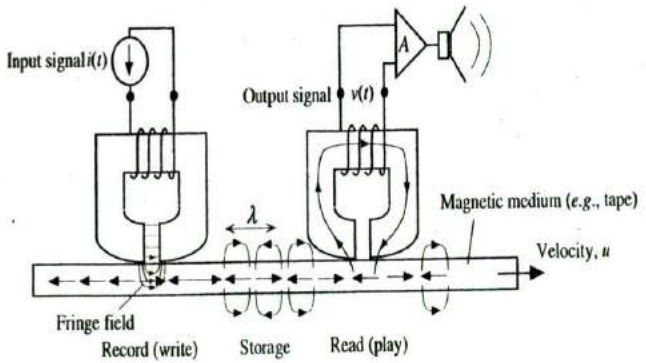


Figure 8.61 The principle of longitudinal magnetic recording on a flexible medium, for example, magnetic tape in an audio cassette.

The audio tape moves forward and passes under a second head, called the play (or read) head, that converts the spatial variations in the magnetization in the tape into a voltage signal that is amplified and appropriately conditioned for playback, as depicted in Figure 8.61. Of course, the same recording head can also serve as the play head, as is customarily the practice in various general audio recording equipment. The reading process is based on Faraday's law of induction. As the magnetized region in the tape passes under the play head, a portion of the magnetic field from this tape region penetrates into the core and flows around the whole core and hence links the coil. We should recall that magnetic fields prefer to flow in high permeability regions to which they are strongly attracted. The field thus loops around through the core of the head. It does so because the magnetic permeability of the core is very high. As the tape moves past the play head, the field linking the coil changes as different magnetized regions in the tape pass through. The changes in the magnetic flux linking the coil generate a voltage $v(t)$ that is proportional to the strength of the field and hence the magnetization in the tape under the head; the speed of the tape remains the same. Thus the spatial magnetic pattern (information) in the tape is converted into an output voltage signal as the tape is run through under the play head at a constant rate. It should be apparent that the spatial magnetic pattern in the tape is proportional to the current signal $i(t)$, whereas the output signal at the play head is the induced voltage $v(t)$.

Suppose that the input signal has a frequency f, or period $1/f$, and the speed of the tape is u. Then the magnetic pattern repeats at every $1/f$ seconds. During this time the tape advances by a distance $\Delta x = u/f$. This Δx represents a spatial wavelength λ that characterizes the repetition of the spatial magnetic pattern that represents the information. The smaller the λ, the greater the f and hence the greater the information that can be stored. Typical video tapes have λ in the submicron range (*e.g.*, 0.75 μm) to be able to store the high density of information in a video signal into a spatial magnetic pattern. The actual recording process in a video cassette recorder is more complicated and involves moving the heads helically across the film, which increases the relative tape speed and hence the induced voltage.

The recording of digital information is straightforward because the information in the form of ones and zeros involves only changes, or no changes, in the direction of magnetization along the tape. In the recording of analog signals, the audio signal is combined with an ac bias signal. However, the analog signal can also be stored as a digital signal by converting it, by an appropriate encoding procedure, to a digital signal.

Hard Disk Storage The basic principle of magnetic recording used in hard disk drives of computers is somewhat similar to the basic schematic illustration for recording on a tape in Figure 8.61, but with a few notable differences that allow high magnetic data storage capacity and a compact size. The basic principle of the magnetic hard disk drive storage is shown in Figure 8.62. The information storage medium is a thin film of magnetic material (described later) coated, for example, by sputtering, on a disk substrate, which rotates inside the hard drive. The information is recorded as magnetization patterns on this thin-film magnetic medium by an inductive write head, similar in principle to the recording head in Figure 8.61. Both the write and the read heads are in a single compact assembly that moves radially across the rotating disk to

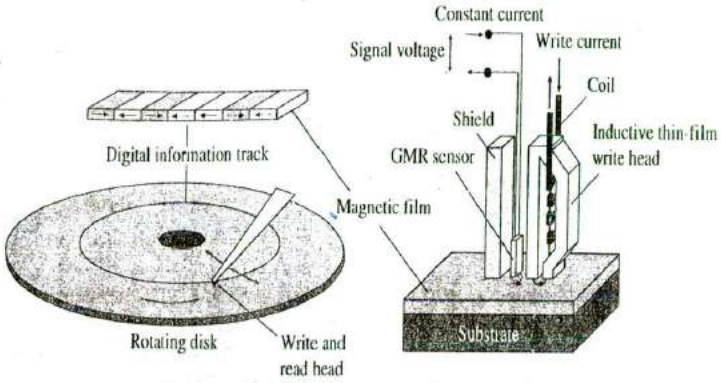
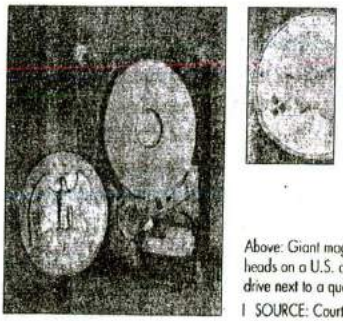


Figure 8.62 The principle of the hard disk drive magnetic recording. The write inductive head and the GMR read sensor have been integrated into a single tiny read/write head.



Above: Giant magnetoresistance (GMR) hard disk heads on a U.S. quarter. Left: A small hard disk drive next to a quarter coin—a microdrive.
SOURCE: Courtesy of IBM.

write or read the information into tracks, called **magnetic bit tracks,** on the magnetic medium. The total area storage density depends on the information density in the track and the track density on the disk. The read head is not an inductive head (as in Figure 8.61) but a tiny giant **magnetoresistance (GMR) sensor** whose resistance depends on an external magnetic field, as explained in Section 8.12. In this case, the field that influences the GMR sensor comes from that of the magnetized region of the disk that is under the GMR sensor. The principle of the GMR is shown in Figure 8.60. The GMR sensor is a multilayered thin-film device whose resistance changes by roughly 10 percent or so in response to an applied field. This change in the resistance generates the read signal. Normally a constant current is passed through the GMR sensor, and the read signal is the voltage variation across the sensor; this voltage is due to the resistance variation induced by the field from the magnetization pattern under the sensor.

There are two important reasons for using a GMR sensor instead of a conventional inductive read head. First is that the GMR sensor is so much smaller than the inductive head that it can probe a much smaller region of the magnetic medium; we can therefore squeeze more information into a given area on the magnetic storage medium. A

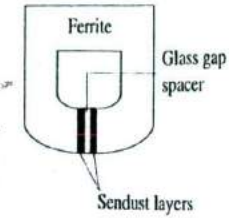


Figure 8.63 A simplified schematic illustration of a MIG (metal-in-gap) head.
The ferrite core has the poles coated with a ferromagnetic soft metal to enhance the head performance.

typical GMR sensor has a width that is something like 50 nm (~1000 times thinner than the human hair). Second, for the same size, GMR is much more sensitive than the inductive head. Thus, all hard drive read heads are tiny GMR sensors as indicated in Figure 8.62. The inductive write head is normally a **thin-film head,** which has a very small width. Consequently, the information can be written into a very small area on the magnetic storage medium. Usually the thin-film write head and the GMR sensor are integrated into the same structure for convenient write and read operations. The aforementioned basic principles still govern the operation of current magnetic hard drive storage devices.[14]

Recording Head Materials The material for the recording head must be magnetically soft so that its magnetization easily follows the input signal (current i or magnetic field intensity H). At the same time, it must provide a strong fringing magnetic field at the gap to magnetize the audio tape, that is, overcome the coercivity of the tape. This requires high saturation magnetization. Thus, the recording head needs small coercivity and large saturation magnetization, which requires soft magnetic materials with as large relative permeabilities as possible.

Typical materials that are used in recording heads are permalloys (Ni–Fe alloys), Sendust (Fe–Al–Si alloy), some sintered soft ferrites (*e.g.*, MnZn and NiZn ferrites), and, more recently, various magnetic amorphous metals such as CoZrNb alloys. Typically, metal-based heads (from permalloy, Sendust, or related materials) are made of laminated metal sheets (with thin insulation between them) to suppress eddy current losses at high frequencies. For high-frequency recording, generally ferrite heads are preferred since ferrites are insulators and suffer no eddy current losses. Ferrites however have low saturation magnetizations and require magnetic storage media of low coercivity. The main problem in ferrite recording heads is that the corners of the poles at the air gap become saturated first. Once saturated, the field around the gap is not proportional to the input current signal, and this degrades the quality of recording. This is overcome by coating the pole faces with a high magnetization metal alloy such as Sendust, or, more recently, a magnetic amorphous metal (*e.g.*, CoZrNb), as depicted in Figure 8.63. Since the magnetic metal alloy is only at the tips of the head, the eddy current losses are still small. This type of head where the poles of the ferrite core have a metal coating is called a metal-in-gap (MIG) head and is widely used in various

[14] One highly recommended book on magnetic recording is R. L. Comstock, *Introduction to Magnetism and Magnetic Recording*, New York: Wiley, 1999. See also R. L. Comstock, "Modern Magnetic Materials in Data Storage," *J. Mater. Sci: Mater. Electron.* **12**, 509, 2002.

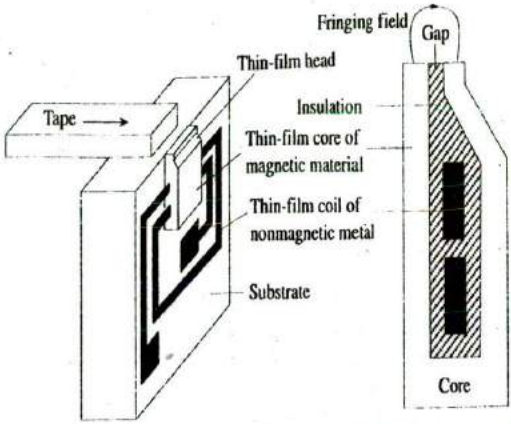


Figure 8.64 A highly simplified schematic illustration of the principle of a thin-film head.

recording applications. The gap distance itself also influences the extent of the fringing field around it and hence the field penetrating into the magnetic tape. The smaller the gap, the greater the fringing. The necessary fringing fields for proper recording on a tape require gap sizes around 1 μm or less.

More recently, recording head devices have been fabricated using thin films of various ferromagnetic metals or ferrite alloys that have sufficiently small eddy current losses to be useable at high frequencies. A highly simplified illustration of the principle of a thin-film head is shown in Figure 8.64. The head is manufactured by using typical thin-film deposition techniques, such as sputtering of the metal film in a vacuum chamber, photolithography, or some other method. The magnetic core is in the form of a thin film whose thickness is a few microns and whose width is about the same as the tape. The gap at the end of the core has the same width as the core, but its spacing is very small (*e.g.*, 0.25 μm) and generates the necessary fringing field. A spiral-type coil made by depositing a nonmagnetic metal thin film threads the core. The magnetic core is like a U-shaped core that is threaded by the metal strips of the coil. If the core is a metallic material, the coil metal is appropriately insulated from it by thin films of insulation.

Magnetic Storage Media Materials The properties of magnetic storage media such as magnetic tapes, floppy disks, and hard disks used in various magnetic recording applications (audio, video, digital, etc.) must be such that they are able to retain the spatial magnetization pattern written on them after they have passed the recording head. This requires high remanent magnetization M_r. High remanent magnetization is also important in the reading process because the magnetic flux that induces voltages in the read head depends on this remanent magnetization, given a particular speed of motion under the read head. Thus the read operation requires media with high M_r.

Further, it should be difficult to undesirably erase the magnetic information on the tape by demagnetizing it under stray fields, and this requires high coercivity H_c. A

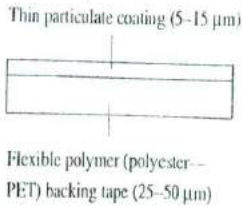


Figure 8.65 A magnetic tape is typically a magnetic coating on a flexible polymer (*e.g.*, PET) sheet in the form of a tape.

strong magnet passed over a floppy disk can destroy the information stored in it. The coercivity therefore determines the stability of the recording. The coercivity cannot be too high, however, as this would prevent the writing operation, that is, magnetization, under the recording head. One therefore has to find a compromise that allows the information to be written and at the same time retained without ease of demagnetization.

These two requirements, high M_r and medium-to-high H_c, lead to a choice of medium to hard magnetic materials as magnetic storage media. Typical flexible storage media (*e.g.*, audio or video tapes) use particulate coatings on flexible polymeric sheets or tapes, as pictured schematically in Figure 8.65. Elongated particles of various magnetic materials are magnetically hard due to a combination of two factors. First, these particles tend to be single domains and are hard due to the magnetocrystalline anisotropy energy. Second, they are also elongated, have a greater length to width ratio (aspect ratio), which means they are also hard due to shape anisotropy; they prefer to be magnetized along the length.

Typical particulate matter used in coatings are γ-Fe_2O_3, Co-modified γ-Fe_2O_3 or Co(γ-Fe_2O_3), CrO_2, and metallic particles (Fe), as summarized in Table 8.9. The overall magnetic properties of the particulate coating depend not only on the properties of the individual particles (which are hard) but also on the concentration of particles as well as their distribution in the coating. For example, as the packing density of particles increases, the saturation magnetization M_{sat} (total magnetic moment per unit volume) also increases, which is desirable, but the coercivity worsens. The concentrations of particles in the coating are typically between 5×10^{14} cm^{-3} (*e.g.*, floppy disk) and

Table 8.9 Selected examples of flexible magnetic storage media based on coatings of particulate matter: typical values

Particulate Matter	Typical Application	$\mu_0 M_r$ (T)	$\mu_0 H_c$ (T)	Comments
γ-Fe_2O_3	Audio tape (Type I)	0.16	0.036	Widely used particles.
γ-Fe_2O_3	Floppy disk	0.07	0.03	
Co(γ-Fe_2O_3)	Video tape	0.13	0.07	Cobalt-impregnated γ-Fe_2O_3 particles.
CrO_2	Audio tape (Type II)	0.16	0.05	More expensive than γ-Fe_2O_3.
CrO_2	Video tape	0.14	0.06	
Fe	Audio tape (Type IV)	0.30	0.11	High coercivity and magnetization. To avoid corrosion, the particles have to be treated (expensive).

5×10^{15} cm^{-3} (*e.g.*, video tape), which are sufficient to provide the necessary remanent field and maintain adequate coercivity.

The brown gamma iron oxide, γ-Fe_2O_3, is a metastable form of iron oxide that is ferrimagnetic and is prepared synthetically. Cobalt-treated γ-Fe_2O_3 particles have a small percentage of Co impregnated into the surface of the particles, which improves the magnetic hardness. Cobalt-impregnated γ-Fe_2O_3 particles are used in various video tapes. All these particles in Table 8.9 are needle shaped (elongated rod-like shapes) with length-to-diameter ratios greater than 5, which makes them substantially hard as a result of shape anisotropy. The needle-like particles are typically 0.3–0.6 μm in length and 0.05–0.1 μm in diameter. The particles are initially mixed into a lacquer-like resin binder that is then coated onto a thin polyester backing tape. When the resin coating solidifies, it forms a magnetic coating stuck on the backing tape. Typically between 20–40 percent of this magnetic coating is actually due to the magnetic particles.

Another form of magnetic storage medium is in the form of magnetic thin films deposited onto various hard substrates or even on a flexible plastic tape as in some video tapes. The hard disk in the hard drive of a computer, for example, is typically an aluminum disk that has a thin magnetic film (*e.g.*, CoPtCr) coated onto it. The deposition of the magnetic thin film may involve vacuum deposition techniques (*e.g.*, electron beam evaporation or sputtering) or electroplating. Typical film thicknesses are less than 50 nm. The advantage of using a thin-film coating is that they are solid films of a magnetic material, that is, almost 100 percent dense, whereas in a particulate medium, the packing density of magnetic particles is 20–40 percent. Consequently, thin magnetic films have higher saturation and remanent magnetizations, which enable a smaller area of the thin film to be used for storing the same information as that in a flexible medium. Thus there is an increase in the stored information density—a distinct advantage. Table 8.10 lists the characteristics of a few selected thin magnetic films used as magnetic storage media. Most thin films are alloys of Co because Co has a high degree of magnetocrystalline anisotropy and hence good coercivity H_c. Alloying Co with Cr provides good corrosion resistance and increases H_c. Alloying with Pt or Ta also increases H_c. The desired film properties can usually be obtained by alloying Co with other elements and optimizing the deposition conditions; this is an ongoing research area. The current commercial interest is to increase the storage density even

Table 8.10 Selected examples of thin films in magnetic storage media: typical values

Thin Film	Typical Deposition	$\mu_0 M_s$ (T)	$\mu_0 H_c$ (T)	Component and Typical or Potential Application
Co and rare earth	Sputtering in vacuum	0.7–0.8	0.65–0.07	Longitudinal magnetic recording media
Co(γ-Fe_2O_3)	Sputtering in vacuum	0.3	0.07–0.08	Longitudinal magnetic recording media
CoNiP	Electroplating	1	0.1	Longitudinal magnetic recording media, hard disks
CoCr alloys	Sputtering in vacuum	0.3–0.7	0.05–0.3	Longitudinal and perpendicular magnetic recording media, hard disks
CoPtCrB	Sputtering in vacuum	0.3–0.5	0.25–0.6	Longitudinal and perpendicular magnetic recording media, hard disks

further by using perpendicular magnetic recording in contrast to longitudinal recording. In perpendicular recording, the local magnetizations in the thin film are perpendicular to the surface of the film.

The magnetic coating on some video tapes may be in the form of a thin film deposited by vacuum evaporation of the magnetic material using an electron beam to heat it. Some recent video tapes have CoNi thin-film coatings that are evaporated by an electron beam onto a polyester (PET) tape.

8.14 JOSEPHSON EFFECT

The Josephson junction is a junction between two superconductors that are separated by a thin insulator (a few nanometers thick) as depicted in Figure 8.66. If the insulating barrier is sufficiently thin, then there is a probability that the Cooper pairs can tunnel across the junction. The wavefunction Ψ of the Cooper pair, however, changes phase by θ when it tunnels through the junction, not unexpected as the pair goes through a potential barrier. The maximum superconducting current I_c that can flow through this weak link depends on not only the thickness and area (size) of the insulator but also on the superconductor materials and the temperature. The current I, or the *supercurrent*, through the junction due to Cooper pair tunneling is determined by the phase angle θ,

Josephson junction supercurrent

$$I = I_c \sin\theta \qquad [8.27]$$

where I_c is the maximum current or the critical current. If the current through the junction is controlled by an external circuit, then the tunneling Cooper pairs on either side of the junction (in the superconductors) adjust their respective phases to maintain the phase change to satisfy Equation 8.27. If we plot the I–V characteristics of this junction as in Figure 8.67, we would find that for $I < I_c$, the behavior follows the vertical OC line with no voltage across the junction.

If the current through the junction exceeds I_c, then the Cooper pairs cannot tunnel through the insulator because Equation 8.27 cannot be satisfied. There is still a current through the junction, but it is due to the tunneling of normal, that is, single electrons as represented by the curve $OABD$ in Figure 8.67. Thus, the current switches from point C to point B and then follows the normal tunneling curve B to D. At point B, a *voltage*

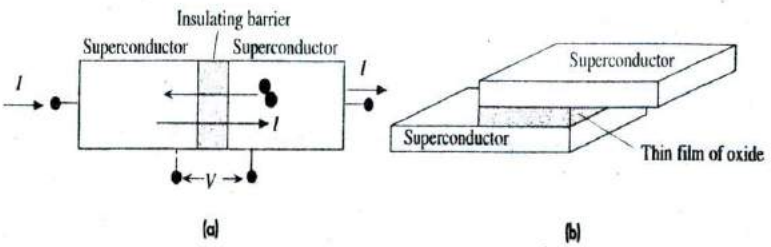


Figure 8.66

(a) A Josephson junction is a junction between two superconductors separated by a thin insulator.

(b) In practice, thin-film technology is used to fabricate a Josephson junction.

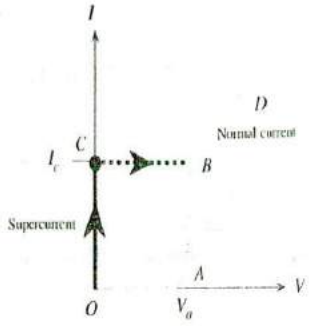
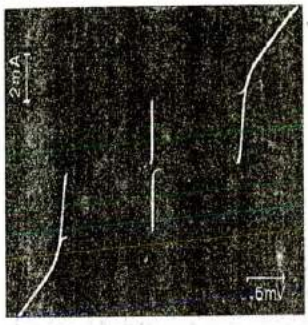


Figure 8.67 I–V characteristics of a Josephson junction for positive currents when the current is controlled by an external circuit.



I–V characteristics of a Sn-SnO-Sn Josephson junction at T = 1.52 K.

SOURCE: E. P. Balsano, G. Paterno, A. Barone, P. Rissman, and M. Russo, "Temperature dependence of the maximum (dc) Josephson current" Phys. Rev. B, vol. 10, p. 1882, Figure 2. © 1974 American Physical Society.

develops across the junction and increases with the current. The normal tunneling current in the range *OA* is negligible and rises suddenly when the voltage exceeds V_a. The reason is that a certain amount of voltage (corresponding to a potential energy eV_a) is needed to provide the necessary energy to disassociate the tunneling single electron from its Cooper pair. It is apparent that the thin insulation acts as a weak superconductor or as a **weak link** in the superconductor; weak with regard to the currents that can flow in the superconductor itself. The *I–V* characteristic in Figure 8.67 is symmetric about *O* (as in the photograph for an actual device), and is called the **dc characteristic of the Josephson junction.** In addition, the *I–V* behavior exhibits hysteresis; that is, if we were to decrease the current, the behavior does not follow *DBC* down to *O*, but follows the *DBA* curve. When the current is decreased nearly to zero, the normal tunneling current switches to the supercurrent. The Josephson junction is bistable; that is, it has two states corresponding to the superconducting state *OC* and normal state *ABD*. Thus, the device behaves as an electronic switch whose switching time, in theory, is determined by tunneling times, in the picoseconds range. In practice the switching time (~10 ps) is limited by the junction capacitance.

If, on the other hand, a dc voltage is applied across the Josephson junction, then the phase change θ is modulated by the applied voltage. The most interesting and surprising aspect is that the voltage modulates the rate of change of the phase through the barrier, that is,

Applied voltage modulates phase

$$\frac{d\theta}{dt} = \frac{2eV}{h}$$

When we integrate this, we find that θ is time and voltage dependent, so, according to Equation 8.27, the current is a sinusoidal function of time and voltage, that is,

$$I = I_c \sin\left(\theta_o - \frac{2\pi(2eV)t}{h}\right)$$

or

$$I = I_o \sin(2\pi ft)$$

where I_o is a new constant incorporating θ_o and the frequency of the oscillations of the current is given by

ac Josephson effect

$$f = \frac{2eV}{h} \qquad [8.28]$$

The Josephson junction therefore generates an oscillating current of frequency f when there is a dc voltage V across it. This is called the **ac Josephson effect,** a remarkable phenomenon originally predicted by Josephson as a graduate student at Cambridge (1962). According to the ac Josephson effect, the junction generates an ac current at a frequency of $2e/h$ Hz per volt or 483.6 MHz per microvolt. Furthermore, the frequency of the current has nothing to do with the material properties of the junction but is only determined by the applied voltage through e and h. The ac Josephson effect has been adopted to define the voltage standard: One volt is the voltage that, when applied to a Josephson junction, will generate an ac current and hence an electromagnetic radiation of frequency 483,597.9 GHz.

8.15 FLUX QUANTIZATION

Consider a ring of a superconducting material above its T_c. Suppose that the ring is immersed in magnetic flux lines from a magnet placed above it as shown in Figure 8.68a. When we cool the ring to below T_c, the magnetic flux lines are excluded from the ring itself, due to the Meissner effect, but they go through the hole, as shown in Figure 8.68b. If we now remove the magnet, we may think that the magnetic flux lines simply disappear, but this is not the case. A persistent current is set up on the inside surface of the superconducting ring that flows to maintain the flux constant in the hollow. This supercurrent generates flux lines in the hollow as if to replace those taken away by the removal of the magnet, as depicted in Figure 8.68c. Since the current can flow indefinitely in the ring, the overall effect is that the magnetic flux is *trapped* within the ring. Indeed, if we were to bring back the magnet, the current in the ring would disappear to ensure that the magnetic flux in the hollow remains unchanged. The origin of

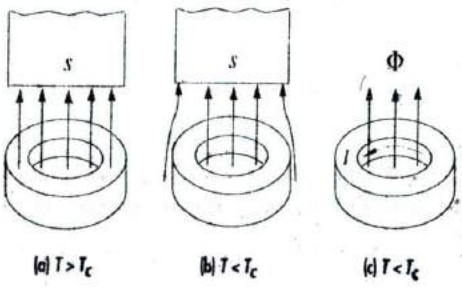


Figure 8.68
(a) Above T_c, the flux lines enter the ring.
(b) The ring and magnet are cooled through T_c. The flux lines do not enter the superconducting ring but stay in the hole.
(c) Removing the magnet does not change the flux in the hole.

flux trapping can be appreciated by considering what would happen if the flux were allowed to change, that is, $d\Phi/dt \neq 0$. A changing flux would induce a voltage $V = -d\Phi/dt$ around the ring that would drive an infinite current $I = V/R$ where $R = 0$. This is not possible, and hence the flux cannot change, which means we must have $d\Phi/dt = 0$. One should also note that there can be no electric field inside a superconductor because

$$\mathcal{E} = \frac{J}{\sigma} = 0$$

since the conductivity σ is infinite.

What would happen if we have a superconducting ring (below T_c) that initially had no flux in the hole? If we were to bring a magnet to it, then the flux lines would now be excluded from both the ring itself and also the hole since the trapped flux within the ring is zero.

It turns out that the trapped flux Φ inside the ring is quantized by virtue of superconductivity being a quantum phenomenon. The smallest quantized amount of flux is called the **magnetic flux quantum** and is given by $h/2e$ or 2.0679×10^{-15} Wb. The flux Φ in the ring is an integer multiple n of this quantum,

$$\Phi = n\frac{h}{2e} \qquad [8.29]$$

Trapped flux is quantized

CD Selected Topics and Solved Problems

Selected Topics	Solved Problems
Atomic Diamagnetism	Diamagnetism: Examples
Atomic Paramagnetism	
Ferrimagnetism and Ferrites	

DEFINING TERMS

Antiferromagnetic materials have crystals in which alternating permanent atomic spin magnetic moments are equal in magnitude but point in opposite directions (antiparallel), which leads to no net magnetization of the crystal.

Bloch wall is a magnetic domain wall.

Bohr magneton (β) is a useful elementary unit of magnetic moment on the atomic scale. It is equal to the magnetic moment of one electron spin along an applied magnetic field $\beta = e\hbar/2m_e$.

Coercivity or **coercive field** (H_c) measures the ability of a magnetized material to resist demagnetization. It is the required reverse applied field that would remove any remanent magnetization, that is, demagnetize the material.

Cooper pair is a quasi-particle formed by the mutual attraction of two electrons with opposite spins and opposite linear momenta below a critical temperature. It has a charge of $-2e$ and a mass of $2m_e$ but no net spin. It does *not* obey Fermi-Dirac statistics. The

electrons are held together by the induced distortions and vibrations of the lattice of positive metal ions with which the electrons interact.

Critical magnetic field (B_c) is the maximum field that can be applied to a superconductor without destroying the superconducting behavior. B_c decreases from its maximum value at absolute zero to zero at T_c.

Critical temperature (T_c) is a temperature that separates the superconducting state from the normal state. Above T_c, the substance is in the normal state with a finite resistivity, but below T_c, it is in the superconducting state with zero resistivity.

Curie temperature (T_C) is the critical temperature at which the ferromagnetic and ferrimagnetic properties are lost. Above the Curie temperature, the material behaves as if it were paramagnetic.

Diamagnetic material has a negative magnetic susceptibility and reduces or repels applied magnetic fields. Superconductors are perfect diamagnets that repel the applied field. Many substances possess weak diamagnetism, so the applied field is slightly decreased within the material.

Domain wall is a region between two neighboring magnetic domains of differing orientations of magnetization.

Domain wall energy is the excess energy in the domain wall as a result of the gradual orientations of the neighboring spin magnetic moments of atoms through the wall region. It is the excess energy due to the excess exchange interaction energy, magnetocrystalline anisotropy energy, and magnetostrictive energy in the wall region.

Easy direction is the crystal direction along which the atomic magnetic moments (due to spin) are spontaneously and most easily aligned. Exchange interaction energy is lowest (hence favorable) when the alignment of atomic spin magnetic moments is in this direction in the crystal. For the iron crystal, it is one of the six [100] directions (cube edge).

Eddy current loss is the Joule energy loss (I^2R) in a ferromagnetic material subjected to changing magnetic fields (in ac fields). The varying magnetic field induces voltages in the ferromagnetic material that drive currents (called eddy currents) that generate Joule heating due to I^2R.

Eddy currents are the induced conduction currents flowing in a ferromagnetic material as a result of varying (ac) magnetic fields.

Exchange interaction energy (E_{ex}) is a kind of Coulombic interaction energy between two neighboring electrons and positive metal ions that depends on the relative spin orientations of the electrons as a consequence of the Pauli exclusion principle. Its exact origin is quantum mechanical. Qualitatively, different spins lead to different electron wavefunctions, different negative charge distributions, and hence different Coulombic interactions. In ferromagnetic crystals, E_{ex} is negative when the neighboring electron spins are parallel.

Ferrimagnetic materials possess crystals that contain two sets of atomic magnetic moments that oppose each other, but one set has greater strength and therefore there is a net magnetization of the crystal. An unmagnetized ferrimagnetic substance normally has many magnetic domains whose magnetization vectors add to give no overall magnetization.

Ferrites are ferrimagnetic materials that are ceramics with insulating properties. They are therefore used in HF applications where eddy current losses are significant. Their general composition is $(MO)(Fe_2O_3)$, where M is typically a divalent metal. For magnetically soft ferrites, M is typically Fe, Mn, Zn, or Ni, whereas for magnetically hard ferrites, M is typically Sr or Ba. Hard ferrites such as $BaOFe_2O_3$ have the hexagonal crystal structure with a high degree of magnetocrystalline anisotropy and therefore possess high coercivity (difficult to demagnetize).

Ferromagnetic materials have the ability to possess large permanent magnetizations even in the absence of an applied field. An unmagnetized ferromagnetic material normally has many magnetic domains whose magnetization vectors add to give no overall magnetization. However, in a sufficiently strong magnetizing field, the whole ferromagnetic substance becomes one magnetic domain in which all the atomic spin magnetic moments are aligned to give a large magnetization along the field. Some of this magnetization is retained even after the removal of the field.

Giant magnetoresistance (GMR) is the large change in the resistance of a special multilayer structure when a magnetic field is applied; the simplest structure usually

consists of two thin ferromagnetic layers (*e.g.*, Fe) sandwiching an even thinner nonmagnetic metal (*e.g.*, Cu).

Hard direction is the crystal direction along which it is hardest to align the atomic spin magnetic moments relative to the easy direction. Exchange interaction energy E_{ex} favors the easy direction most (E_{ex} is most negative) and favors the hard direction least (E_{ex} is least negative).

Hard magnetic materials characteristically have high remanent magnetizations (B_r) and high coercivities (H_c), so once magnetized, they are difficult to demagnetize. They are suitable for permanent magnet applications. They have broad B–H hysteresis loops.

Hard magnetic particles are small particles of various shapes that have high coercivity due to having a single magnetic domain with high magnetocrystalline anisotropy energy, or possessing substantial shape anisotropy (aspect ratio—length-to-width ratio).

Hysteresis loop is the magnetization (M) versus applied magnetic field intensity (H) or B versus H behavior of a ferromagnetic (or ferrimagnetic) substance through one cycle as it is repeatedly magnetized and demagnetized.

Hysteresis loss is the energy loss involved in magnetizing and demagnetizing a ferromagnetic (or ferrimagnetic) substance. It arises from various energy losses involved in the irreversible motions of the domain walls. Hysteresis loss per unit volume of specimen is the area of the B–H hysteresis loop.

Initial permeability (μ_i, μ_{ri}) is the initial slope of the B versus H characteristic of an unmagnetized ferromagnetic (or ferrimagnetic) material and typically represents the magnetic permeability under very small applied magnetic fields. Initial relative permeability (μ_{ri}) is the relative permeability of an unmagnetized ferromagnetic (or ferrimagnetic) material under very small applied fields.

Magnetic dipole moment ($\boldsymbol{\mu}_m$) is defined as $IA\mathbf{u}_n$, where I is the current flowing in a circuit loop of area A and $\mathbf{u}_n$ is the unit vector in the direction of an advance of a screw when it is turned in the direction of the circulating current. Qualitatively, it measures the strength of the magnetic field created by a current loop and also the extent of interaction of the current loop with an externally applied magnetic field. $\boldsymbol{\mu}_m$ is normal to the surface of the loop. Magnetic moment in a magnetic field experiences a torque that tries to rotate $\boldsymbol{\mu}_m$ to align it with the field. In a nonuniform field, the magnetic moment experiences a force that attracts it to a greater field.

Magnetic domain is a region of a ferromagnetic (or ferrimagnetic) crystal that has spontaneous magnetization, that is, magnetization in the absence of an applied field, due to the alignment of all magnetic moments in that region.

Magnetic field, magnetic induction, or magnetic flux density (B) is a field that is generated by a current-carrying conductor that produces a force on a current-carrying conductor elsewhere. Equivalently, we can define it as the field generated by a moving charge that acts to produce a force on a moving charge elsewhere. The force is called the Lorentz force and is given by $\mathbf{F} = q\mathbf{v} \times \mathbf{B}$ where $\mathbf{v}$ is the velocity of the particle with charge q. The magnetic field $\mathbf{B}$ in a material is the sum of the applied field $\mu_o\mathbf{H}$, and that due to the magnetization of the material $\mu_o\mathbf{M}$, that is, $\mathbf{B} = \mu_o(\mathbf{H} + \mathbf{M})$.

Magnetic field intensity or magnetizing field (H) gauges the magnetic strength of external conduction currents (*e.g.*, currents flowing in the windings) in the absence of a material medium. It excludes the magnetization currents that become induced on the surfaces of any material placed in a magnetic field. $\mu_o H$ is the magnetic field in free space and is considered to be the *applied magnetic field*. The terms *intensity* or *strength* distinguish $\mathbf{H}$ from $\mathbf{B}$, which is simply called the magnetic field.

Magnetic flux (Φ) represents to what extent magnetic field lines are flowing through a given area perpendicular to the field lines. If δA is a small area perpendicular to the magnetic field B and B is constant over δA, then the flux $\delta\Phi$ through δA is defined by $\delta\Phi = B\,\delta A$. Total flux through any closed surface is zero.

Magnetic permeability (μ) is the magnetic field generated per unit magnetizing field, that is, $\mu = B/H$. Permeability gauges the effectiveness of a medium in generating as much magnetic field as possible per unit magnetizing field. Permeability of free space is the absolute permeability μ_o, which is the magnetic field generated in a vacuum per unit magnetizing field.

Magnetic susceptibility (χ_m) indicates the ease with which the material becomes magnetized under an applied magnetic field. It is the magnetization induced in the material per unit magnetizing field, $\chi_m = M/H$.

Magnetization or **magnetization vector (M)** represents the net magnetic moment per unit volume of material. In the presence of a magnetic field, individual atomic moments tend to align with the field, which results in a net magnetization. Magnetization of a specimen can be represented by the flow of currents on the surface over a unit length of the specimen; $M = I_m$, where I_m is the surface magnetization current per unit length.

Magnetization current (I_m) is a bound current per unit length that exists on the surface of a substance due to its magnetization. It is not, however, due to the flow of free charges but arises in the presence of an applied magnetic field as a result of the orientations of the electronic motions in the constituent atoms. In the bulk, these electronic motions cancel each other and there is no net bulk current, but on the surface, they add to give a bound surface current I_m per unit length, which is equal to the magnetization M of the substance.

Magnetocrystalline anisotropy is the anisotropy associated with magnetic properties such as the magnetization in different directions in a ferromagnetic (or ferrimagnetic) crystal. Atomic spins prefer to align along certain directions in the crystal, called easy directions. The direction along which it is most difficult to align the spins is called the hard direction. For example, in the iron crystal, all atomic spins prefer to align along one of the [100] directions (easy directions) and it is most difficult to align the spins along one of the [111] directions (hard directions).

Magnetocrystalline anisotropy energy (K) is the energy needed to rotate the magnetization of a ferromagnetic (or ferrimagnetic) crystal from its natural easy direction to a hard direction. For example, it takes an energy of about 48 mJ cm^{-3} to rotate the magnetization of an iron crystal from the easy direction [100] to the hard direction [111].

Magnetoresistance generally refers to the change in the resistance of a magnetic material when it is placed in a magnetic field. The change in the resistance of a nonmagnetic metal, such as copper, is usually very small. In a magnetic metal, the change in the resistivity due to the applied magnetic field is *anisotropic*; that is, it depends on the direction of current flow with respect to the applied field and is called **anisotropic magnetoresistance (AMR)**.

Magnetostatic energy is the potential energy stored in an external magnetic field. It takes external work to establish a magnetic field, and this energy is said to be stored in the magnetic field. Magnetic energy per unit volume at a point in free space is given by

$$E_{vol}(\text{air}) = \frac{1}{2}\mu_o H^2 = \frac{B^2}{2\mu_o}$$

Magnetostriction is the change in the length of a ferromagnetic (or ferrimagnetic) crystal as a result of its magnetization. An iron crystal placed in a magnetic field along an easy direction becomes longer along this direction but contracts in the transverse direction.

Magnetostrictive energy is the strain energy in the crystal due to magnetostriction, that is, the work done in straining the crystal when it becomes magnetized.

Maximum relative permeability ($\mu_{r,\max}$) is the maximum relative permeability of a ferromagnetic (or ferrimagnetic) material.

Meissner effect is the repulsion of all magnetic flux from the interior of a superconductor. The superconductor behaves as if it were a perfect diamagnet with $\chi_m = -1$.

Paramagnetic materials have a small and positive magnetic susceptibility. In an applied field, they develop a small amount of magnetization in the direction of the applied field, so the magnetic field in the material is slightly greater. They are attracted to a higher magnetic field.

Relative permeability (μ_r) measures the magnetic field in a medium with respect to that in a vacuum, $\mu_r = B/\mu_o H$. Since B depends on the magnetization of the medium, μ_r measures the ease with which the material becomes magnetized.

Remanence or **remanent magnetization** (M_r) is the magnetization that remains in a magnetic material after it has been fully magnetized and the magnetizing field has been removed. It measures the ability of a magnetic

material to retain a portion of its magnetization after the removal of the applied field. The corresponding magnetic field ($\mu_o \mathbf{M}_r$) is the remanent magnetic field $\mathbf{B}_r$.

Saturation magnetization is the maximum magnetization that can be obtained in a ferromagnetic crystal at a given temperature when all the magnetic moments have been aligned in the direction of the applied field, when there is a single magnetic domain with its magnetization **M** along the applied field.

Shape anisotropy is the anisotropy in magnetic properties associated with the shape of the ferromagnetic (or ferrimagnetic) substance. A crystal rod that is thin and long prefers to have its magnetization **M** along the length (long axis) of the rod because this direction of magnetization creates less external magnetic fields and leads to less external magnetostatic energy compared with the case when **M** is along the width (short axis) of the rod. Reversing the magnetization involves rotating **M** through the width of the rod, where the external magnetic field and hence magnetostatic energy are large, and requires large substantial work. It is therefore difficult to rotate magnetization around from the long axis to the short axis.

Soft magnetic materials characteristically have high saturation magnetizations (B_{sat}) but low saturation magnetizing fields (H_{sat}) and low coercivities (H_c), so they can be magnetized and demagnetized easily. They have tall and narrow B–H hysteresis loops.

Superconductivity is a phenomenon in which a substance loses all resistance to current flow (acquires zero resistivity) and also exhibits the Meissner effect (becomes a perfect diamagnet).

Type I superconductors have a single critical field (B_c) above which the superconducting behavior is totally lost.

Type II superconductors have a lower (B_{c1}) and an upper (B_{c2}) critical field. Below B_{c1}, the substance is in the superconducting phase with Meissner effect; all magnetic flux is excluded from the interior. Between B_{c1} and B_{c2}, magnetic flux lines pierce through local filamentary regions of the superconductor, which behave normally. Above B_{c2}, the superconductor reverts to normal behavior.

QUESTIONS AND PROBLEMS

8.1 **Inductance of a long solenoid** Consider the very long (ideally infinitely long) solenoid shown in Figure 8.69. If r is the radius of the core and ℓ is the length of the solenoid, then $\ell \gg r$. The total number of turns is N and the number of turns per unit length is $n = N/\ell$. The current through the coil wires is I. Apply Ampere's law around C, which is the rectangular circuit $PQRS$, and show that

$$B \approx \mu_o \mu_r n I$$

Further, show that the inductance is

$$L \approx \mu_o \mu_r n^2 V_{core}$$

Inductance of a long solenoid

where V_{core} is the volume of the core. How would you increase the inductance of a long solenoid?

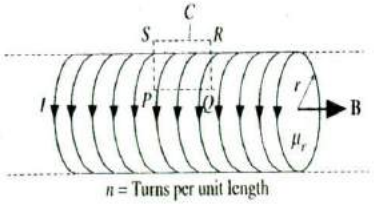


Figure 8.69

What is the approximate inductance of an air-cored solenoid with a diameter of 1 cm, length of 20 cm, and 500 turns? What is the magnetic field inside the solenoid and the energy stored in the whole solenoid when the current is 1 A? What happens to these values if the core medium has a relative permeability μ_r of 600?

8.2 **Magnetization** Consider a long solenoid with a core that is an iron alloy (see Problem 8.1 for the relevant formulas). Suppose that the diameter of the solenoid is 2 cm and the length of the solenoid is 20 cm. The number of turns on the solenoid is 200. The current is increased until the core is magnetized to saturation at about $I = 2$ A and the saturated magnetic field is 1.5 T.

a. What is the magnetic field intensity at the center of the solenoid and the applied magnetic field, $\mu_o H$, for saturation?

b. What is the saturation magnetization M_{sat} of this iron alloy?

c. What is the total magnetization current on the surface of the magnetized iron alloy specimen?

d. If we were to remove the iron-alloy core and attempt to obtain the same magnetic field of 1.5 T inside the solenoid, how much current would we need? Is there a practical way of doing this?

8.3 **Paramagnetic and diamagnetic materials** Consider bismuth with $\chi_m = -16.6 \times 10^{-5}$ and aluminum with $\chi_m = 2.3 \times 10^{-5}$. Suppose that we subject each sample to an applied magnetic field B_a of 1 T applied in the $+x$ direction. What is the magnetization **M** and the equivalent magnetic field $\mu_o M$ in each sample? Which is paramagnetic and which is diamagnetic?

8.4 **Mass and molar susceptibilities** Sometimes magnetic susceptibilities are reported as molar or mass susceptibilities. **Mass susceptibility** (in $m^3 kg^{-1}$) is χ_m/ρ where ρ is the density. **Molar susceptibility** (in $m^3 mol^{-1}$) is $\chi_m (M_{at}/\rho)$ where M_{at} is the atomic mass. Terbium (Tb) has a magnetic molar susceptibility of 2.0 $cm^3 mol^{-1}$. Tb has a density of 8.2 g cm^{-3} and an atomic mass of 158.93 g mol^{-1}. What is its susceptibility, mass susceptibility and relative permeability? What is the magnetization in the sample in an applied magnetic field of 2 T?

8.5 **Pauli spin paramagnetism** Paramagnetism in metals depends on the number of conduction electrons that can flip their spins and align with the applied magnetic field. These electrons are near the Fermi level E_F, and their number is determined by the density of states $g(E_F)$ at E_F. Since each electron has a spin magnetic moment of β, paramagnetic susceptibility can be shown to be given by

Pauli spin paramagnetism

$$\chi_{para} \approx \mu_o \beta^2 g(E_F)$$

where the density of states is given by Equation 4.10. The Fermi energy of calcium E_F is 4.68 eV. Evaluate the paramagnetic susceptibility of calcium and compare with the experimental value of 1.9×10^{-5}.

8.6 **Ferromagnetism and the exchange interaction** Consider dysprosium (Dy), which is a rare earth metal with a density of 8.54 g cm^{-3} and atomic mass of 162.50 g mol^{-1}. The isolated atom has the electron structure $[Xe]4f^{10}6s^2$. What is the spin magnetic moment in the isolated atom in terms of number of Bohr magnetons? If the saturation magnetization of Dy near absolute zero of temperature is 2.4 MA m^{-1}, what is the effective number of spins per atom in the ferromagnetic state? How does this compare with the number of spins in the isolated atom? What is the order of magnitude for the exchange interaction in eV per atom in Dy if the Curie temperature is 85 K?

8.7 **Magnetic domain wall energy and thickness** The energy of a Bloch wall depends on two main factors: the exchange energy E_{ex} (J/atom) and magnetocrystalline energy K (J m^{-3}). If a is the interatomic distance and δ is the wall thickness, then it can be shown that the potential energy per unit area of the wall is

Potential energy of a Bloch wall

$$U_{wall} = \frac{\pi^2 E_{ex}}{2a\delta} + K\delta$$

Show that the minimum energy occurs when the wall has the thickness

Bloch wall thickness

$$\delta' = \left(\frac{\pi^2 E_{ex}}{2aK}\right)^{1/2}$$

and show that when $\delta = \delta'$, the exchange and anisotropy energy contributions are *equal*. Using reasonable values for various parameters, estimate the Bloch energy and wall thickness for Ni. (See Example 8.4.)

***8.8 Toroidal inductor and radio engineers toroidal inductance equation**

a. Consider a toroidal coil (Figure 8.10) whose mean circumference is ℓ and that has N tightly wound turns around it. Suppose that the diameter of the core is $2a$ and $\ell \gg a$. By applying Ampere's law, show that if the current through the coil is I, then the magnetic field in the core is

$$B = \frac{\mu_0 \mu_r N I}{\ell} \qquad [8.30]$$

where μ_r is the relative permeability of the medium. Why do you need $\ell \gg a$ for this to be valid? Does this equation remain valid if the core cross section is not circular but rectangular, $a \times b$, and $\ell \gg a$ and b?

b. Show that the inductance of the toroidal coil is

$$L = \frac{\mu_0 \mu_r N^2 A}{\ell} \qquad [8.31]$$

Toroidal coil inductance

where A is the cross-sectional area of the core.

c. Consider a toroidal inductor used in electronics that has a ferrite core size FT-37, that is, round but with a rectangular cross section. The outer diameter is 0.375 in (9.52 mm), the inner diameter is 0.187 in (4.75 mm), and the height of the core is 0.125 in (3.175 mm). The initial relative permeability of the ferrite core is 2000, which corresponds to a ferrite called the 77 Mix. If the inductor has 50 turns, then using Equation 8.31, calculate the approximate inductance of the coil.

d. Radio engineers use the following equation to calculate the inductances of toroidal coils,

$$L(\text{mH}) = \frac{A_L N^2}{10^6} \qquad [8.32]$$

Radio engineers inductance equation

where L is the inductance in millihenries (mH) and A_L is an inductance parameter, called an **inductance index**, that characterizes the core of the inductor. A_L is supplied by the manufacturers of ferrite cores and is typically quoted as millihenries (mH) per 1000 turns. In using Equation 8.32, one simply substitutes the numerical value of A_L to find L in millihenries. For the FT-37 ferrite toroid with the 77 Mix as the ferrite core, A_L is specified as 884 mH/1000 turns. What is the inductance of the toroidal inductor in part (*c*) from the radio engineers equation in Equation 8.32? What is the percentage difference in values calculated by Equations 8.32 and 8.31? What is your conclusion? (*Comment*: The agreement is not always this close.)

***8.9 A toroidal inductor**

a. Equations 8.31 and 8.32 allow the inductance of a toroidal coil in electronics to be calculated. Equation 8.32 is the equation that is used in practice. Consider a toroidal inductor used in electronics that has a ferrite core of size FT-23 that is round but with a rectangular cross section. The outer diameter is 0.230 in (5.842 mm), the inner diameter is 0.120 in (3.05 mm), and the height of the core is 0.06 in (1.5 mm). The ferrite core is a 43-Mix that has an initial relative permeability of 850 and a maximum relative permeability of 3000. The inductance index for this 43-Mix ferrite core of size FT-23 is $A_L = 188$ (mH/1000 turns). If the inductor has 25 turns, then using Equations 8.31 and 8.32, calculate the inductance of the coil under small-signal conditions and comment on the two values.

b. The saturation field B_{sat} of the 43-Mix ferrite is 0.2750 T. What will be typical dc currents that will saturate the ferrite core (an estimate calculation is required)? It is not unusual to find such an inductor in an electronic circuit also carrying a dc current. Will your calculation of the inductance remain valid in these circumstances?

c. Suppose that the toroidal inductor discussed in parts (a) and (b) is in the vicinity of a very strong magnet that saturates the magnetic field inside the ferrite core. What will be the inductance of the coil?

*8.10 The transformer

a. Consider the transformer shown in Figure 8.70a whose primary is excited by an ac (sinusoidal) voltage of frequency f. The current flowing into the primary coil sets up a magnetic flux in the transformer core. By virtue of Faraday's law of induction and Lenz's law, the flux generated in the core is the flux necessary to induce a voltage nearly equal and opposite to the applied voltage. Thus,

$$v = \frac{d(\text{Total flux linked})}{dt} = \frac{NA\,dB}{dt}$$

where A is the cross-sectional area, assumed constant, and N is the number of turns in the primary. Show that if V_{rms} is the rms voltage at the primary ($V_{max} = V_{rms}\sqrt{2}$) and B_m is the maximum magnetic field in the core, then

Transformer equation

$$V_{rms} = 4.44\,NAfB_m \qquad [8.33]$$

Transformers are typically operated with B_m at the "knee" of the B–H curve, which corresponds roughly to maximum permeability. For transformer irons, $B_m \approx 1.2$ T. Taking $V_{rms} = 120$ V and a transformer core with $A = 10$ cm $\times$ 10 cm, what should N be for the primary winding? If the secondary winding is to generate 240 V, what should be the number of turns for the secondary coil?

b. The transformer core will exhibit hysteresis and eddy current losses. The **hysteresis loss** per unit second, as power loss in watts, is given by

Hysteresis loss

$$P_h = Kf B_m^n V_{core} \qquad [8.34]$$

where $K = 150.7$, f is the ac frequency (Hz), B_m is the maximum magnetic field (T) in the core (assumed to be in the range 0.2–1.5 T), $n = 1.6$, and V_{core} is the volume of the core. The eddy current losses are reduced by laminating the transformer core, as shown in Figure 8.70b. The **eddy current loss** is given by

Eddy current loss

$$P_e = 1.65\,f^2 B_m^2 \left(\frac{d^2}{\rho}\right) V_{core} \qquad [8.35]$$

where d is the thickness of the laminated iron sheet in meters (Figure 8.70b) and ρ is its resistivity (Ω m).

Suppose that the transformer core has a volume of 0.0108 m^3 (corresponds to a mean circumference of 1.08 m). If the core is laminated into sheets of thickness 1 mm and the resistivity of the transformer iron is 6×10^{-7} Ω m, calculate both the hysteresis and eddy current losses at $f = 60$ Hz, and comment on their relative magnitudes. How would you reduce each loss?

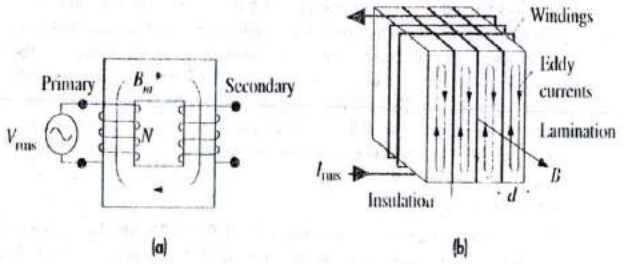


Figure 8.70

(a) A transformer with N turns in the primary.

(b) Laminated core reduces eddy current losses.

8.11 **Losses in a magnetic recording head** Consider eddy current losses in a permalloy magnetic head for audio recording up to 10 kHz. We will use Equation 8.35 for the eddy current losses. Consider a magnetic head weighing 30 g and made from a permalloy with density 8.8 g cm^{-3} and resistivity 6×10^{-7} Ω m. The head is to operate at B_m of 0.5 T. If the eddy current losses are not to exceed 1 mW, estimate the thickness of laminations needed. How would you achieve this?

***8.12** **Design of a ferrite antenna for an AM receiver** We consider an AM radio receiver that is to operate over the frequency range 530–1600 kHz. Suppose that the receiving antenna is to be a coil with a ferrite rod as core, as depicted in Figure 8.71. The coil has N turns, its length is ℓ, and the cross-sectional area is A. The inductance L of this coil is tuned with a variable capacitor C. The maximum value of C is 265 pF, which with L should correspond to tuning in the lowest frequency at 530 kHz. The coil with the ferrite core receives the EM waves, and the magnetic field of the EM wave permeates the ferrite core and induces a voltage across the coil. This voltage is detected by a sensitive amplifier, and in subsequent electronics it is suitably demodulated. The coil with the ferrite core therefore acts as the antenna of the receiver (ferrite antenna). We will try to find a suitable design for the ferrite coil by carrying out approximate calculations—in practice some trial and error experimentation would also be necessary. We will assume that the inductance of a finite solenoid is

$$L = \frac{\gamma \mu_{ri} \mu_0 A N^2}{\ell} \qquad [8.36]$$

Inductance of a solenoid

where A is the cross-sectional area of the core, ℓ is the coil length, N is the number of turns, and γ is a geometric factor that accounts for the solenoid coil being of finite length. Assume $\gamma \approx 0.75$. The resonant frequency f of an LC circuit is given by

$$f = \frac{1}{2\pi (LC)^{1/2}} \qquad [8.37]$$

LC circuit resonant frequency

a. If d is the diameter of the enameled wire to be used as the coil winding, then the length $\ell \approx Nd$. If we use an enameled wire of diameter 1 mm, what is the number of coil turns N we need for a ferrite rod given that its diameter is 1 cm and its initial relative permeability is 100?

b. Suppose that the magnetic field intensity H of the signal in free space is varying sinusoidally, that is,

$$H = H_m \sin(2\pi f t) \,. \qquad [8.38]$$

where H_m is the maximum magnetic field intensity. H is related to the electric field $\mathcal{E}$ at a point by $H = \mathcal{E}/Z_{space}$, where Z_{space} is the impedance of free space given by 377 Ω. Show that the induced voltage at the antenna coil is

$$V_m = \frac{\mathcal{E}_m d}{2\pi 377 C f \gamma} \qquad [8.39]$$

Induced voltage across a ferrite antenna

where f is the frequency of the AM wave and $\mathcal{E}_m$ is the electric field intensity of the AM station at the receiver point. Suppose that the electric field of a local AM station at the receiver is 10 mV m^{-1}. What is the voltage induced across the ferrite antenna and can this voltage be detected by an amplifier? Would you use a ferrite rod antenna at short-wave frequencies, given the same C but less N?

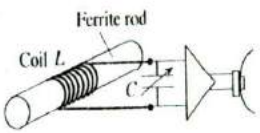


Figure 8.71 A ferrite antenna of an AM receiver.

***8.13** **A permanent magnet with an air gap** The magnetic field energy in the gap of a permanent magnet is available to do work. Suppose that B_m and B_g are the magnetic field in the magnet and the gap, H_m and H_g are the field intensities in the magnet and the gap, and V_m and V_g are the volumes of the magnet

and gap; show that, in terms of magnitudes,

Magnet and gap relationship

$$B_g H_g V_g \approx B_m H_m V_m$$

What is the significance of this result?

8.14 A permanent magnet with an air gap

a. Show that the maximum energy stored in the air gap of a permanent magnet can be written very roughly as

Energy in gap of a magnet

$$E_{gap} \approx \frac{1}{8} B_r H_c V_m$$

where V_m is the volume of the magnet, which is much greater than that of the gap; B_r is the remanent magnetic field; and H_c is the coercivity of the magnet.

b. Using Table 8.6, compare the $(BH)_{max}$ with the product $(\frac{1}{2} H_c)(\frac{1}{2} B_r)$ and comment on the closeness of agreement.

c. Calculate the energy in the gap of a rare earth cobalt magnet that has a volume of 0.1 m^3. Give an example of typical work (*e.g.*, raising so many apples, each 100 g, by so many meters) that could be done if all this energy could be converted to mechanical work.

8.15 Weight, cost, and energy of a permanent magnet with an air gap For a certain application, an energy of 1 kJ is required in the gap of a permanent magnet. There are three candidates, as shown in Table 8.11. Which material will give the lightest magnet? Which will give the cheapest magnet?

Table 8.11 Three permanent magnet candidates

Magnet	$(BH)_{max}$ (kJ m^{-3})	Density (g cm^{-3})	Yesterday's Relative Price (per unit mass)
Alnico	50	7.3	1
Rare earth	200	8.2	2
Ferrite	30	4.8	0.5

***8.16 Permanent magnet with yoke and air gap** Consider a permanent magnet bar that has L-shaped ferromagnetic (high permeability) pieces attached to its ends to direct the magnetic field to an air gap as depicted in Figure 8.72. The L-shaped high μ_r pieces for directing the magnetic field are called **yokes**. Suppose that A_m, A_y, and A_g are the cross-sectional areas of the magnet, yoke, and gap as indicated in the figure. The lengths of the magnet, yoke, and air gap are ℓ_m, y, and g, respectively. The magnet, the two yokes, and the gap can be considered to be all connected end-to-end or in series. Applying Ampere's circuital law for H we can write,

$$H_m \ell_m + 2H_y \ell_y + H_g \ell_g = 0$$

Since all four components, magnet, yokes, and gap, are in series, we can assume that the magnetic flux Φ through each of them is the same,

$$\Phi = B_m A_m = B_y A_y = B_g A_g$$

a. Show that

$$H_m = -\frac{A_m}{\ell_m}\left[\frac{\ell_g}{\mu_0 A_g} + \frac{2\ell_y}{\mu_0 \mu_{ry} A_y}\right] B_m$$

b. What does the equation in part (*a*) represent? Given that B_m and H_m in the magnet must obey the equation in part (*a*), and also the B–H characteristic of the magnet material itself, what is your conclusion?

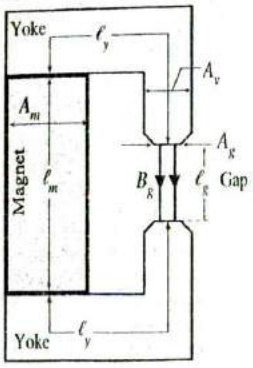


Figure 8.72 A permanent magnet with two pieces of yoke and an air gap.

c. Should the yokes be magnetically hard or soft? Justify your decision.

d. Show that if μ_{rv} is very large ($\mu_{rv} \approx \infty$),

$$H_m = -\frac{1}{\mu_v}\left[\frac{A_m \ell_g}{A_g \ell_m}\right] B_m$$

e. If $V_m = A_m \ell_m$ and $V_g = A_g \ell_g$ are the volumes of the magnet and gap, respectively, show that

$$B_g H_g V_g = B_m H_m V_m$$

What is your conclusion (consider the magnetic energy stored in the gap)?

f. Consider a rare earth permanent magnet, with a density of 8.2 g cm^{-3}, that has a $(BH)_{max}$ of about 200 kJ m^{-3}. Suppose that $(BH)_{max}$ occurs very roughly at $B_m \approx \frac{1}{2} B_r$ where for this rare earth magnet $B_r \approx 1$ T. Suppose that $A_m \approx A_g$. What is the volume, effective length (ℓ_m), and mass of the magnet that is needed to store the maximum energy in the gap if $\ell_g = 1$ cm and $A_g = 1000$ cm^2? What is the maximum energy in the gap?

8.17 **Superconductivity and critical current density** Consider two superconducting wires, tin (Sn; Type I) and Nb_3Sn (Type II), each 1 mm in thickness. The magnetic field on the surface of a current-carrying conductor is given by

$$B = \frac{\mu_0 I}{2\pi r}$$

a. Assuming that Sn wire loses its superconductivity when the field at the surface reaches the critical field (0.2 T), calculate the maximum current and hence the critical current density that can be passed through the Sn wire near absolute zero of temperature.

b. Calculate the maximum current and critical current density for the Nb_3Sn wire using the same assumption as in part (*a*) but taking the critical field to be the upper critical field B_{c2}, which is 24.5 T at 0 K. How does your calculation of J_c compare with the critical density of about 10^{11} A m^{-2} for Nb_3Sn at 0 K?

8.18 **Magnetic pressure in a solenoid** Consider a long solenoid with an air core. Diametrically opposite windings have oppositely directed currents and, due to the magnetic force, they repel each other. This means that the solenoid experiences a *radial force* F_r that is trying to open up the solenoid, *i.e.*, stretch out the windings as depicted in Figure 8.53. Suppose that A is the surface area of the core (on to which wires are wound). If we decrease the core diameter by dx, the volume changes by dV. We have to do work dW against the radial magnetic forces F_r.

$$dW = F_r\, dx = \left(\frac{F_r}{A}\right) A\, dx = P_r\, dV$$

where $P_r = F_r/A$ is the *radial pressure*, called the **magnetic pressure**, acting on the windings of the solenoid. (This pressure acts to tear apart the solenoid.) Using the fact that the work done against the magnetic forces in changing the volume changes the magnetic energy in the core, show that

Radial magnetic pressure in a solenoid

$$P_r = \frac{B^2}{2\mu_o}$$

What is the radial pressure on a solenoid that has a field of 35 T in the core? How many atmospheres is this? What is the equivalent ocean depth that gives the same pressure? What happens to this pressure at 100 T?

***8.19 Enterprising engineers in the high arctic building a superconducting inductor** A current-carrying inductor has energy stored in its magnetic field that can be converted to electrical work. A group of enterprising engineers and scientists living in Resolute in Nunavut (Canada) have decided to build a toroidal inductor to store energy so that this energy can be used to supply a small community of 10 houses each consuming on average 3 kW of energy during the night (6 months). They have discovered a superconductor (Type II) that has a $B_{c2} = 100$ T and a critical current density of $J_c = 5 \times 10^{10}$ A m^{-2} at night temperatures (it is obviously a novel high-T_c superconductor of some sort). Their superconducting wire has a diameter of 5 mm and is available in any desirable length. All the wiring in the community is done by superconductors except where energy needs to be converted to other forms (mechanical, heat, etc.). They have decided on the following design specification for their toroid:

> The mean diameter D_{toroid} of the toroid, $(\frac{1}{2})$ (Outside diameter + Inside diameter), is 10 times longer than the core diameter D_{core}. The field inside the toroid is therefore reasonably uniform to within 10 percent.
>
> The maximum operating magnetic field in the core is 35 T. Fields larger than this can result in mechanical fracture and failure.
>
> Assume that J_c decreases linearly with the magnetic field and that the mechanical engineers in the group can take care of the forces trying to blow open the toroid by building a proper support structure.

Find the size of the toroid (mean diameter and circumference), the number of turns and the length of the superconducting wire they need, the current in the coil, and whether this current is sufficiently below the critical current at that field. Is it feasible?

8.20 Magnetic storage media

a. Consider the storage of video information (FM signal) on a video tape. Suppose that the maximum signal frequency to be recorded as a spatial magnetic pattern is 10 MHz. The heads helically scan the tape, and the relative velocity of the tape to head is about 10 m s^{-1}. What is the minimum spatial wavelength of the stored magnetic pattern (information) on the tape?

b. Suppose that the speed of an audio cassette tape in a cassette player is 5 cm s^{-1}. If the maximum frequency that needs to be recorded is 20 kHz, what is the minimum spatial wavelength on the tape?

Note: An excellent quantitive description of magnetic recording may be found in R. L. Comstock, *Introduction to Magnetism and Magnetic Recording*, New York: John Wiley & Sons, 1999.

***8.21 Magnetic recording principles** In this "back of an envelope" calculation we consider the principle of operation of a recording head for writing on a magnetic tape (perhaps an audio or a video tape). The recording head has a small gap, of size g (about 1 μm or less), which is much smaller than the mean circumference of the head ℓ (perhaps a few millimeters) as shown in Figure 8.73. The coil of this head has N turns and is energized by the signal current i. The fringe field intensity H_f at the gap magnetizes the magnetic tape passing under the head. H_f must be greater than the coercivity H_c of the storage medium (tape) to be able to magnetize that region of the tape under the head. Suppose that H_m = magnetic field intensity in the core of the head; H_g = magnetic field intensity in the gap; H_f = fringing field intensity below the gap; $B_m = \mu_r \mu_o H_m$ = magnetic field in the core of the head; $B_g = \mu_o H_g$ = magnetic field in the gap.

The magnetic flux must be continuous through the small gap. Thus, if A is the cross-sectional area,

$$\text{Flux in the core} = AB_m = \text{Flux in the gap} = AB_g \quad \text{or} \quad B_g = B_m$$

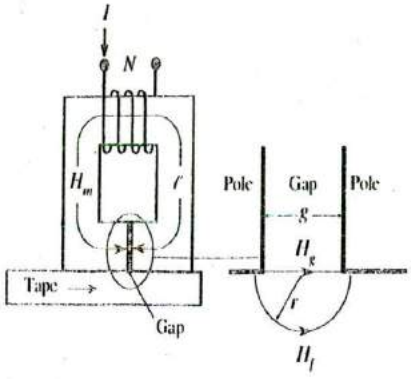


Figure 8.73 The gap of a recording head and the fringing field for magnetizing the tape.

a. Applying Ampere's law for H around the mean circumference, $\ell + g$, show that

$$H_g = \frac{1}{g + \ell/\mu_r} NI$$

Field in the gap

b. If we apply Ampere's law for H around the semicircle of radius r coming out from the gap into the tape as shown in Figure 8.73 we get

$$H_g g - H_f(\pi r) \approx 0$$

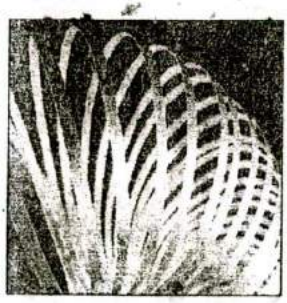
Show that,

$$H_f \approx \frac{\mu_r g}{\pi r(\mu_r g + \ell)} NI$$

Fringing field for recording on storage media

c. The fringing field must overcome the coercivity of the storage medium. Suppose that the storage medium has $H_c = 50$ kA m^{-1} and we have to determine Ni given the head material. Suppose that $\mu_r \approx 10^4$, $g = 1\ \mu\text{m} = 10^{-6}$ m, $\ell \approx 5$ mm $= 5 \times 10^{-3}$ m, and $r = 1\ \mu\text{m} = 10^{-6}$ m to record into a depth of 1 μm. What is the minimum NI? If the minimum signal current (after amplification) is 5 mA, how many turns do you need for the coil?

d. What is the magnetic field B_m in the core? Can you use a ferrite head?





Left: These high-temperature superconductor (HTS) flat tapes are based on $(Bi_2Pb_x)Sr_2Ca_2Cu_3O_{10+\delta}$ (Bi-2223). The tape has an outer surrounding protective metallic sheath. Right: HTS tapes have a major advantage over equivalent-sized metal conductors, in being able to transmit considerably higher power loads. Coils made from HTS tape can be used to create more compact and efficient motors, generators, magnets, transformers, and energy storage devices.

SOURCE: Courtesy of Australian Superconductors.



Augustin Jean Fresnel (1788–1827) was a French physicist, and a civil engineer for the French government, who was one of the principal proponents of the wave theory of light. He made a number of distinct contributions to optics including the well known Fresnel lens that was used in lighthouses in the nineteenth century. He fell out with Napoleon in 1815 and was subsequently put into house arrest until the end of Napoleon's reign. During his enforced leisure time he formulated his wave ideas of light into a mathematical theory.

SOURCE: Smithsonian Institution, courtesy of AIP Emilio Segrè Visual Archives



Christiaan Huygens (1629–1695), a Dutch physicist, explained double refraction of light in calcite in terms of ordinary and extraordinary waves. Christiaan Huygens made many contributions to optics and wrote prolifically on the subject.

SOURCE: Courtesy of Emilo Segrè Visual Archives, American Institute of Physics.

CHAPTER

9

Optical Properties of Materials

The way electromagnetic (EM) radiation interacts with matter depends very much on the wavelength of the EM wave. Many familiar types of EM radiation have wavelengths that range over many orders of magnitude. Although radio waves and X-rays are both EM waves, the two interact in a distinctly different way with matter. We tend to think of "light" as the electromagnetic radiation that we can see, that is, wavelengths in the visible range, typically 400 to 700 nm. However, in many applications, light is also used to describe EM waves that can have somewhat shorter or longer wavelengths such as ultraviolet (UV) and infrared (IR) light. For many practical purposes, it is useful to (arbitrarily) define light as EM waves that have wavelengths shorter than very roughly 100 μm but longer than long-wavelength X-rays, roughly 10 nm. Today's *light wave communications* use EM waves with wavelengths of 1300 and 1550 nm; in the infrared. *Optical properties* of materials are those characteristic properties that determine the interaction of light with matter; the best example being the refractive index n that determines the speed of light in a medium through $v = c/n$, where v is the speed of light in the medium and c is the speed of light in free space. The present chapter examines the key optical properties of matter and how these depend on the material and on the characteristics of the EM wave. The refractive index n, for example, depends on the dielectric polarization mechanisms as well as the wavelength λ. The material's n–λ behavior is called the **dispersion relation** and is one of the most important characteristics in many optical device applications.

We know from Chapter 3 that, depending on the experiment, we can treat light either as an EM wave, exhibiting typical wave-like properties, or as photons, exhibiting particle-like behavior. In this chapter we will primarily use the wave nature of light, though for absorption of light, the photon interpretation is more appropriate as the photons interact with electrons in the material.

9.1 LIGHT WAVES IN A HOMOGENEOUS MEDIUM

We know from well-established experiments that light exhibits typical wave-like properties such as interference and diffraction. We can treat light as an EM wave with time-varying electric and magnetic fields E_x and B_y, respectively, which propagate through space in such a way that they are always perpendicular to each other and the direction of propagation z is as depicted in Figure 9.1. The simplest traveling wave is a sinusoidal wave, which, for propagation along z, has the general mathematical form,[1]

Traveling wave along z

$$E_x = E_o \cos(\omega t - kz + \phi_o) \qquad [9.1]$$

where E_x is the electric field at position z at time t; k is the **propagation constant**, or **wavenumber**, given by $2\pi/\lambda$, where λ is the wavelength; ω is the angular frequency; E_o is the amplitude of the wave; and ϕ_o is a phase constant which accounts for the fact that at $t = 0$ and $z = 0$, E_x may or may not necessarily be zero depending on the choice of origin. The argument $(\omega t - kz + \phi_o)$ is called the **phase** of the wave and denoted by ϕ. Equation 9.1 describes a **monochromatic plane wave** of infinite extent traveling in the positive z direction as depicted in Figure 9.2. In any plane perpendicular to the direction of propagation (along z), the phase of the wave, according to Equation 9.1, is constant which means that the field in this plane is also constant. A surface over which the phase of a wave is constant is referred to as a **wavefront.** A wavefront of a plane wave is obviously a plane perpendicular to the direction of propagation as shown in Figure 9.2.

We know from electromagnetism that time-varying magnetic fields result in time-varying electric fields (Faraday's law) and vice versa. A time-varying electric field would set up a time-varying magnetic field with the same frequency. According to electromagnetic principles,[2] a traveling electric field E_x as represented by Equation 9.1 would always be accompanied by a traveling magnetic field B_y with the same wave frequency and propagation constant (ω and k) but the directions of the two fields would be orthogonal as in Figure 9.1. Thus, there is a similar traveling wave equation for the magnetic field component B_y. We generally describe the interaction of a light wave with a nonconducting matter (conductivity, $\sigma = 0$) through the electric field component E_x rather than B_y because it is the electric field that displaces the electrons in molecules or ions in the crystal and thereby gives rise to the polarization of matter. However, the two fields are linked, as in Figure 9.1, and there is an intimate relationship between the two fields. The **optical field** refers to the electric field E_x.

We can also represent a traveling wave using the exponential notation since $\cos\phi = \text{Re}[\exp(j\phi)]$ where Re refers to the real part. We then need to take the real

[1] This chapter uses E for the electric field which was reserved for energy in previous chapters. There should be no confusion with E_g that represents the energy bandgap. In addition, n is used to represent the refractive index rather than the electron concentration.

[2] Maxwell's equations formulate electromagnetic phenomena and provide relationships between the electric and magnetic fields and their space and time derivatives. We only need to use a few selected results from Maxwell's equations without delving into their derivations. The *magnetic field B* is also called the magnetic induction or magnetic flux density.

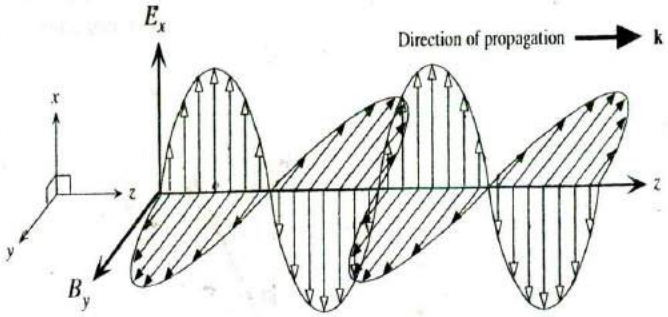


Figure 9.1 An electromagnetic wave is a traveling wave that has time-varying electric and magnetic fields that are perpendicular to each other and the direction of propagation z.

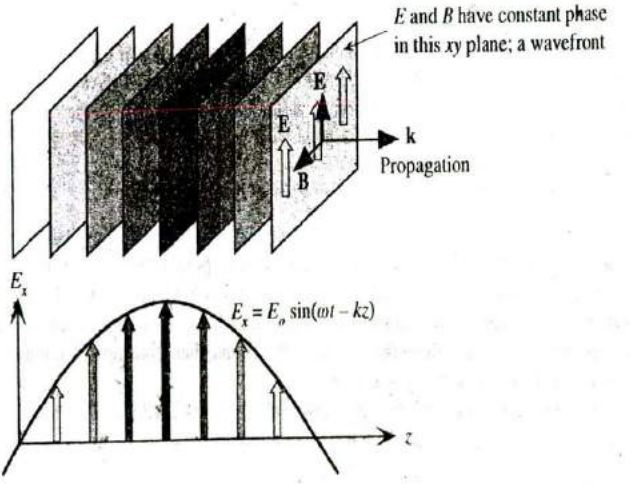


Figure 9.2 A plane EM wave traveling along z, has the same E_x (or B_y) at any point in a given xy plane.
All electric field vectors in a given xy plane are therefore in phase. The xy planes are of infinite extent in the x and y directions.

part of any complex result at the end of calculations. Thus, we can write Equation 9.1 as

$$E_x(z,t) = \text{Re}[E_o \exp(j\phi_o) \exp j(\omega t - kz)]$$

or

$$E_x(z,t) = \text{Re}[E_c \exp j(\omega t - kz)] \qquad [9.2]$$

Traveling wave along z

where $E_c = E_o \exp(j\phi_o)$ is a complex number that represents the amplitude of the wave and includes the constant phase information ϕ_o.

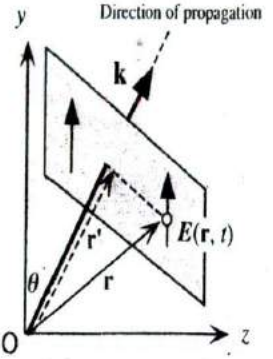


Figure 9.3 A traveling plane EM wave along a direction k.

We indicate the direction of propagation with a vector k, called the **wavevector**, whose magnitude is the propagation constant $k = 2\pi/\lambda$. It is clear that k is perpendicular to constant phase planes as indicated in Figure 9.2. When the EM wave is propagating along some arbitrary direction k, as indicated in Figure 9.3, then the electric field $E(\mathbf{r}, t)$ at a point r on a plane perpendicular to k is

Light wave in three dimensions

$$E(\mathbf{r}, t) = E_o \cos(\omega t - \mathbf{k} \cdot \mathbf{r} + \phi_o) \qquad [9.3]$$

because the dot product $\mathbf{k} \cdot \mathbf{r}$ is along the direction of propagation similar to kz. The dot product is the product of k and the projection of r onto k which is r′ in Figure 9.3, so $\mathbf{k} \cdot \mathbf{r} = kr'$. Indeed, if propagation is along z, $\mathbf{k} \cdot \mathbf{r}$ becomes kz. In general, if k has components k_x, k_y, and k_z along the x, y, and z directions, then from the definition of the dot product, $\mathbf{k} \cdot \mathbf{r} = k_x x + k_y y + k_z z$.

The time and space evolution of a given phase ϕ, for example, the phase corresponding to a maximum field, according to Equation 9.1 is described by

$$\phi = \omega t - kz + \phi_o = \text{constant}$$

During a time interval δt, this constant phase (and hence the maximum field) moves a distance δz. The phase velocity of this wave is therefore $\delta z/\delta t$. Thus the **phase velocity** v is

Phase velocity

$$v = \frac{dz}{dt} = \frac{\omega}{k} = \nu\lambda \qquad [9.4]$$

where ν is the frequency ($\omega = 2\pi\nu$).

We are frequently interested in the phase difference $\Delta\phi$ at a given time between two points on a wave (Figure 9.1) that are separated by a certain distance. If the wave is traveling along z with a wavevector k, as in Equation 9.1, then the phase difference between two points separated by Δz is simply $k\,\Delta z$ since ωt is the same for each point. If this phase difference is 0 or multiples of 2π, then the two points are in phase. Thus, the phase difference $\Delta\phi$ can be expressed as $k\,\Delta z$ or $2\pi\,\Delta z/\lambda$.

9.2 REFRACTIVE INDEX

When an EM wave is traveling in a dielectric medium, the oscillating electric field polarizes the molecules of the medium at the frequency of the wave. Intuitively, the EM wave propagation can be considered to be the propagation of this polarization in the medium. The field and the induced molecular dipoles become coupled. The net effect is that the polarization mechanism delays the propagation of the EM wave. The stronger the interaction between the field and the dipoles, the slower is the propagation of the wave. The relative permittivity ε_r measures the ease with which the medium becomes polarized, and hence it indicates the extent of interaction between the field and the induced dipoles. For an EM wave traveling in a nonmagnetic dielectric medium of relative permittivity ε_r, the phase velocity v is given by

Phase velocity in a medium with ε_r

$$v = \frac{1}{\sqrt{\varepsilon_r \varepsilon_o \mu_o}} \qquad [9.5]$$

If the frequency ν is in the optical frequency range, then ε_r will be due to electronic polarization as ionic polarization will be too sluggish to respond to the field. However, at the infrared frequencies or below, the relative permittivity also includes a significant contribution from ionic polarization and the phase velocity is slower. For an EM wave traveling in free space, $\varepsilon_r = 1$ and $v_{\text{vacuum}} = 1/\sqrt{\varepsilon_o \mu_o} = c = 3 \times 10^8 \text{ m s}^{-1}$, the velocity of light in a vacuum. The ratio of the speed of light in free space to its speed in a medium is called the **refractive index** n of the medium,

Definition of refractive index

$$n = \frac{c}{v} = \sqrt{\varepsilon_r} \qquad [9.6]$$

Suppose that in free space k_o is the wavevector ($k_o = 2\pi/\lambda_o$) and λ_o is the wavelength, then the wavevector k in the medium will be nk_o and the wavelength λ will be λ_o/n. Indeed, we can also define the refractive index in terms of the wavevector k in the medium with respect to that in a vacuum k_o,

Definition of refractive index

$$n = \frac{k}{k_o} \qquad [9.7]$$

Equation 9.6 is in agreement with our intuition that light propagates more slowly in a denser medium which has a higher refractive index. We should note that the frequency ν remains the same. The refractive index of a medium is not necessarily the same in all directions. In noncrystalline materials such as glasses and liquids, the material structure is the same in all directions and n does not depend on the direction. The refractive index is then **isotropic.** In crystals, however, the atomic arrangements and interatomic bonding are different along different directions. Crystals, in general, have nonisotropic, or *anisotropic*, properties. Depending on the crystal structure, the relative permittivity ε_r is different along different crystal directions. This means that, in general, the refractive index n seen by a propagating EM wave in a crystal will depend on the value of ε_r along the direction of the oscillating electric field (that is, along the direction of polarization). For example, suppose that the wave in Figure 9.1 is traveling along the z direction in a particular crystal with its electric field oscillating

along the x direction. If the relative permittivity along this x direction is ε_{r_x}, then $n_x = \sqrt{\varepsilon_{r_x}}$. The wave therefore propagates with a phase velocity that is c/n_x. The variation of n with direction of propagation and the direction of the electric field depends on the particular crystal structure. With the exception of cubic crystals (such as diamond) all crystals exhibit a degree of optical anisotropy which leads to a number of important applications. Typically noncrystalline solids, such as glasses and liquids, and cubic crystals are **optically isotropic;** they possess only one refractive index for all directions.

EXAMPLE 9.1 **RELATIVE PERMITTIVITY AND REFRACTIVE INDEX** Relative permittivity ε_r, or the dielectric constant, of materials is frequency dependent and further it depends on crystallographic direction since it is easier to polarize the medium along certain directions in the crystal. Glass has no crystal structure; it is amorphous. The relative permittivity is therefore isotropic but nonetheless frequency dependent.

The relationship $n = \sqrt{\varepsilon_r}$ between the refractive index n and ε_r must be applied at the same frequency for both n and ε_r. The relative permittivity for many materials can be vastly different at high and low frequencies because different polarization mechanisms operate at these frequencies. At low frequencies all polarization mechanisms present can contribute to ε_r, whereas at optical frequencies only the electronic polarization can respond to the oscillating field. Table 9.1 lists the relative permittivity ε_r(LF) at low frequencies (*e.g.*, 60 Hz or 1 kHz as would be measured for example using a capacitance bridge in the laboratory) for various materials. It then compares $\sqrt{\varepsilon_r(\text{LF})}$ with n.

For diamond and silicon there is an excellent agreement between $\sqrt{\varepsilon_r(\text{LF})}$ and n. Both are covalent solids in which electronic polarization (electronic bond polarization) is the only polarization mechanism at low and high frequencies. Electronic polarization involves the displacement of light electrons with respect to positive ions of the crystal. This process can readily respond to the field oscillations up to optical or even ultraviolet frequencies.

For AgCl and SiO_2, $\sqrt{\varepsilon_r(\text{LF})}$ is larger than n because at low frequencies both of these solids possess a degree of ionic polarization. The bonding has a substantial degree of ionic character which contributes to polarization at frequencies below far-infrared wavelengths. (The AgCl crystal has almost all ionic bonding.) In the case of water, the ε_r(LF) is dominated by orientational or

Table 9.1 Low-frequency (LF) relative permittivity ε_r(LF) and refractive index n

Material	ε_r (LF)	$\sqrt{\varepsilon_r(\text{LF})}$	n (optical)	Comments
Diamond	5.7	2.39	2.41 (at 590 nm)	Electronic bond polarization up to UV light
Si	11.9	3.44	3.45 (at 2.15 μm)	Electronic bond polarization up to optical frequencies
AgCl	11.14	3.33	2.00 (at 1–2 μm)	Ionic polarization contributes to ε_r(LF)
SiO_2	3.84	2.00	1.46 (at 600 nm)	Ionic polarization contributes to ε_r(LF)
Water	80	8.9	1.33 (at 600 nm)	Dipolar polarization contributes to ε_r(LF), which is large

dipolar polarization which is far too sluggish to respond to high-frequency oscillations of the field at optical frequencies.

It is instructive to consider what factors affect n. The simplest (and approximate) expression for the relative permittivity is

Relative permittivity and polarizability

$$\varepsilon_r \approx 1 + \frac{N\alpha}{\varepsilon_o} \qquad [9.8]$$

where N is the number of molecules per unit volume and α is the polarizability per molecule. Both atomic concentration, or density, and polarizability therefore increase n. For example, glasses of given type but with greater density tend to have higher n.

9.3 DISPERSION: REFRACTIVE INDEX–WAVELENGTH BEHAVIOR

The refractive index of materials in general depends on the frequency, or the wavelength. This wavelength dependence follows directly from the frequency dependence of the relative permittivity ε_r. Figure 9.4 shows what happens to an atom in the presence of an oscillating electric field E which is due to a light wave passing through this location; it may also be due to an applied external field.

In the absence of an electric field and in equilibrium, the center of mass C of the orbital motions of the electrons coincides with the positively charged nucleus at O and the net electric dipole moment is zero as indicated in Figure 9.4a. Suppose that the atom has Z number of electrons orbiting the nucleus and all the electrons are contained

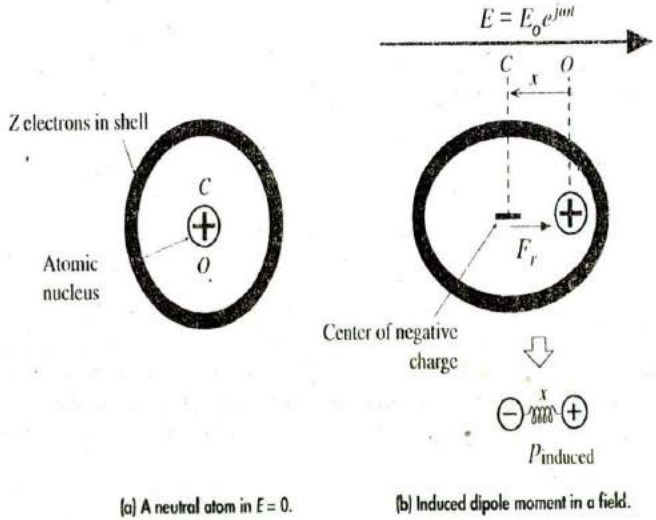


Figure 9.4 Electronic polarization of an atom. In the presence of a field in the $+x$ direction, the electrons are displaced in the $-x$ direction (from O), and the restoring force is in the $+x$ direction.

within a given shell. In the presence of the electric field E, however, the light electrons become displaced in the opposite direction to the field, so their center of mass C is shifted by some distance x with respect to the nucleus O which we take to be the origin as shown in Figure 9.4b. As the electrons are "pushed" away by the applied field, the Coulombic attraction between the electrons and nuclear charge "pulls in" the electrons. The force on the electrons, due to E, trying to separate them away from the nuclear charge is ZeE. The restoring force F_r, which is the Coulombic attractive force between the electrons and the nucleus, can be taken to be proportional to the displacement x provided that the latter is small. The reason is that $F_r = F_r(x)$ can be expanded in powers of x, and for small x only the linear term matters. The restoring force F_r is obviously zero when C coincides with O ($x = 0$). We can write $F_r = -\beta x$ where β is a constant and the negative sign indicates that F_r is always directed toward the nucleus O.

First consider applying a dc field. In equilibrium, the *net* force on the negative charge is zero or $ZeE = \beta x$ from which x is known. Therefore the *magnitude* of the induced electronic dipole moment is given by

Induced electronic dc dipole moment

$$p_{\text{induced}} = (Ze)x = \frac{Z^2e^2}{\beta}E \qquad [9.9]$$

As expected p_{induced} is proportional to the applied field. The electronic dipole moment in Equation 9.9 is valid under static conditions, *i.e.*, when the electric field is a dc field. Suppose that we suddenly remove the applied electric field polarizing the atom. There is then only the restoring force $-\beta x$, which always acts to pull the electrons toward the nucleus O. The equation of motion of the negative charge center is then (force = mass × acceleration)

Simple harmonic motion

$$-\beta x = Zm_e\frac{d^2x}{dt^2}$$

By solving this differential equation we can show that the displacement at any time is a simple harmonic motion, that is,

$$x(t) = x_o\cos(\omega_o t)$$

where the angular frequency of oscillation ω_o is

Natural frequency of the atom

$$\omega_o = \left(\frac{\beta}{Zm_e}\right)^{1/2} \qquad [9.10]$$

In essence, this is the oscillation frequency of the center of mass of the electron cloud about the nucleus and x_o is the displacement before the removal of the field. After the removal of the field, the electronic charge cloud executes simple harmonic motion about the nucleus with a **natural frequency** ω_o determined by Equation 9.10; ω_o is also called the **resonance frequency.** The oscillations, of course, die out with time because there is an inevitable loss of energy from an oscillating charge cloud. An oscillating electron is like an oscillating current and loses energy by radiating EM waves; all accelerating charges emit radiation.

Consider now the presence of an oscillating electric field due to an EM wave passing through the location of this atom as in Figure 9.4b. The applied field oscillates

harmonically in the $+x$ and $-x$ directions, that is, $E = E_o \exp(j\omega t)$. This field will drive and oscillate the electrons about the nucleus. There is again a restoring force F_r acting on the displaced electrons trying to bring back the electron shell to its equilibrium placement around the nucleus. For simplicity we will again neglect energy losses. Newton's second law for Ze electrons with mass Zm_e driven by E is given by

Lorentz oscillator model

$$Zm_e \frac{d^2x}{dt^2} = -ZeE_o \exp(j\omega t) - \beta x \qquad [9.11]$$

The solution of this equation gives the instantaneous displacement $x(t)$ of the center of mass of electrons from the nucleus (C from O),

$$x = x(t) = -\frac{eE_o \exp(j\omega t)}{m_e(\omega_o^2 - \omega^2)}$$

The induced electronic dipole moment is then simply given by $p_{\text{induced}} = -(Ze)x$. The negative sign is needed because normally x is measured from negative to positive charge whereas in Figure 9.4b it is measured from the nucleus. By definition, the electronic polarizability α_e is the induced dipole moment per unit electric field,

Electronic polarizability

$$\alpha_e = \frac{p_{\text{induced}}}{E} = \frac{Ze^2}{m_e(\omega_o^2 - \omega^2)} \qquad [9.12]$$

Thus, the displacement x and hence electronic polarizability α_e increase as ω increases. Both become very large when ω approaches the natural frequency ω_o. In practice, charge separation x and hence polarizability α_e do not become infinite at $\omega = \omega_o$ because two factors impose a limit. First, at large x, the system is no longer linear and this analysis is not valid. Secondly, there is always some energy loss.

Given that the polarizability is frequency dependent as in Equation 9.12, the effect on the refractive index n is easy to predict. The simplest (and a very rough) relationship between the relative permittivity ε_r and polarizability α_e is

Relative permittivity and polarizability

$$\varepsilon_r = 1 + \frac{N}{\varepsilon_o}\alpha_e$$

where N is the number of atoms per unit volume. Given that the refractive index n is related to ε_r by $n^2 = \varepsilon_r$, it is clear that n must be frequency dependent, *i.e.*,

Dispersion relation

$$n^2 = 1 + \left(\frac{NZe^2}{\varepsilon_o m_e}\right)\frac{1}{\omega_o^2 - \omega^2} \qquad [9.13]$$

We can also express this in terms of the wavelength λ. If $\lambda_o = 2\pi c/\omega_o$ is the resonance wavelength, then Equation 9.13 is equivalent to

Dispersion relation

$$n^2 = 1 + \left(\frac{NZe^2}{\varepsilon_o m_e}\right)\left(\frac{\lambda_o}{2\pi c}\right)^2 \frac{\lambda^2}{\lambda^2 - \lambda_o^2} \qquad [9.14]$$

This type of relationship between n and the frequency ω, or wavelength λ, is called the **dispersion relation.** Although the above treatment is grossly simplified, it does nonetheless emphasize that n will always be wavelength dependent and will exhibit a

Table 9.2 Sellmeier and Cauchy coefficients

	Sellmeier					
	A_1	A_2	A_3	λ_1 (μm)	λ_2 (μm)	λ_3 (μm)
SiO_2 (fused silica)	0.696749	0.408218	0.890815	0.0690660	0.115662	9.900559
86.5% SiO_2–13.5% GeO_2	0.711040	0.451885	0.704048	0.0642700	0.129408	9.425478
GeO_2	0.80686642	0.71815848	0.85416831	0.068972606	0.15396605	11.841931
Sapphire	1.023798	1.058264	5.280792	0.0614482	0.110700	17.92656
Diamond	0.3306	4.3356	—	0.1750	0.1060	—

	Cauchy				
	Range of $h\nu$ (eV)	n_{-2} (eV^2)	n_0	n_2 (eV^{-2})	n_{-4} (eV^{-4})
Diamond	0.05–5.47	-1.07×10^{-5}	2.378	8.01×10^{-3}	1.04×10^{-4}
Silicon	0.002–1.08	-2.04×10^{-8}	3.4189	8.15×10^{-2}	1.25×10^{-2}
Germanium	0.002–0.75	-1.0×10^{-8}	4.003	2.2×10^{-1}	1.4×10^{-1}

SOURCE: Sellmeier coefficients combined from various sources. Cauchy coefficients from D. Y. Smith *et al., J. Phys. CM* **13**, 3883, 2001.

substantial increase as the frequency increases toward a natural frequency of the polarization mechanism. In the above example, we considered the electronic polarization of an isolated atom with a well-defined natural frequency ω_o. In the crystal, however, the atoms interact, and further we also have to consider the valence electrons in the bonds. The overall result is that n is a complicated function of the frequency or the wavelength. One possibility is to assume a number of resonant frequencies, that is, not just λ_o but a series of resonant frequencies, $\lambda_1, \lambda_2, \ldots$, and then sum the contributions arising from each with some weighing factor A_1, A_2, etc.,

Sellmeier equation

$$n^2 = 1 + \frac{A_1\lambda^2}{\lambda^2 - \lambda_1^2} + \frac{A_2\lambda^2}{\lambda^2 - \lambda_2^2} + \frac{A_3\lambda^2}{\lambda^2 - \lambda_3^2} + \cdots \qquad [9.15]$$

where A_1, A_2, A_3 and λ_1, λ_2, and λ_3 are constants, called **Sellmeier coefficients.**[3] Equation 9.15 turns out to be quite a useful semiempirical expression for calculating n at various wavelengths if the Sellmeier coefficients are known. Higher terms involving A_4 and higher A coefficients can generally be neglected in representing n versus λ behavior over typical wavelengths of interest. For example, for diamond, we only need the A_1 and A_2 terms. The Sellmeier coefficients are listed in various optical data handbooks.

There is another well-known useful n–λ dispersion relation due originally to Cauchy (1836), which has the short form given by

Cauchy short-form dispersion equation

$$n = A + \frac{B}{\lambda^2} + \frac{C}{\lambda^4} \qquad [9.16]$$

[3] This is also known as the Sellmeier-Herzberger formula.

where A, B, and C are material specific constants. Typically, the Cauchy equation is used in the visible spectrum for various optical glasses. A more general Cauchy dispersion relation is of the form[4]

Cauchy dispersion equation in photon energy

$$n = n_{-2}(h\nu)^{-2} + n_0 + n_2(h\nu)^2 + n_4(h\nu)^4 \qquad [9.17]$$

where $h\nu$ is the photon energy, and n_0, n_{-2}, n_2, and n_4 are constants; values for diamond, Si, and Ge are listed in Table 9.2. The general Cauchy equation is usually applicable over a wide photon energy range.

EXAMPLE 9.2

GaAs DISPERSION RELATION For GaAs, from $\lambda = 0.89$ to 4.1 μm, the refractive index is given by the following dispersion relation,

GaAs dispersion relation

$$n^2 = 7.10 + \frac{3.78\lambda^2}{\lambda^2 - 0.2767} \qquad [9.18]$$

where λ is in microns (μm). What is the refractive index of GaAs for light with a photon energy of 1 eV?

SOLUTION

At $h\nu = 1$ eV,

$$\lambda = \frac{hc}{h\nu} = \frac{(6.62 \times 10^{-34}\ \text{J s})(3 \times 10^{8}\ \text{m s}^{-1})}{(1\ \text{eV} \times 1.6 \times 10^{-19}\ \text{J eV}^{-1})} = 1.24\ \mu\text{m}$$

Thus,

$$n^2 = 7.10 + \frac{3.78\lambda^2}{\lambda^2 - 0.2767} = 7.10 + \frac{3.78(1.24)^2}{(1.24)^2 - 0.2767} = 11.71$$

so that $n = 3.42$

Note that the n versus λ expression for GaAs is actually a Sellmeier-type formula because when $\lambda^2 \gg \lambda_1^2$, then A_1 can be simply lumped with 1 to give $1 + A_1 = 7.10$.

EXAMPLE 9.3

SELLMEIER EQUATION AND DIAMOND The relevant Sellmeier coefficients for diamond are given in Table 9.2. Calculate its refractive index at 550 nm (green light) to three decimal places.

SOLUTION

The Sellmeier dispersion relation for diamond is

$$n^2 = 1 + \frac{0.3306\lambda^2}{\lambda^2 - (175\ \text{nm})^2} + \frac{4.3356\lambda^2}{\lambda^2 - (106\ \text{nm})^2}$$

$$= 1 + \frac{0.3306(550\ \text{nm})^2}{(550\ \text{nm})^2 - (175\ \text{nm})^2} + \frac{4.3356(550\ \text{nm})^2}{(550\ \text{nm})^2 - (106\ \text{nm})^2} = 5.8707$$

So that $n = 2.423$

which is about 0.1 percent different than the experimental value of 2.426.

---

[4] D. Y. Smith *et al.*, *J. Phys. CM* **13**, 3883, 2001.

EXAMPLE 9.4 **CAUCHY EQUATION AND DIAMOND** Using the Cauchy coefficients for diamond in Table 9.2, calculate the refractive index at 550 nm.

SOLUTION

At $\lambda = 550$ nm, the photon energy is

$$h\nu = \frac{hc}{\lambda} = \frac{(6.62 \times 10^{-34}\ \text{J s})(3 \times 10^{8}\ \text{m s}^{-1})}{550 \times 10^{-9}\ \text{m}} \times \frac{1}{1.6 \times 10^{-19}\ \text{J eV}^{-1}} = 2.254\ \text{eV}$$

Using the Cauchy dispersion relation for diamond with coefficients from Table 9.2,

$$\begin{aligned} n &= n_{-2}(h\nu)^{-2} + n_0 + n_2(h\nu)^2 + n_4(h\nu)^4 \\ &= (-1.07 \times 10^{-5})(2.254)^{-2} + 2.378 + (8.01 \times 10^{-3})(2.254)^2 + (1.04 \times 10^{-4})(2.254)^4 \\ &= 2.421 \end{aligned}$$

The difference in n from the value in Example 9.3 is 0.08 percent, and is due to the Cauchy coefficients quoted in Table 9.2 being applicable over a wider wavelength range at the expense of some accuracy.

9.4 GROUP VELOCITY AND GROUP INDEX

Since there are no perfect monochromatic waves in practice, we have to consider the way in which a group of waves differing slightly in wavelength will travel along the z direction as depicted in Figure 9.5. When two perfectly harmonic waves of frequencies $\omega - \delta\omega$ and $\omega + \delta\omega$ and wavevectors $k - \delta k$ and $k + \delta k$ interfere, as shown in Figure 9.5, they generate a **wavepacket** which contains an oscillating field at the mean frequency ω that is amplitude modulated by a slowly varying field of frequency $\delta\omega$. The maximum amplitude moves with a wavevector δk and thus with a **group velocity** that is given by $\delta\omega/\delta k$, that is,

Group velocity

$$v_g = \frac{d\omega}{dk} \qquad [9.19]$$

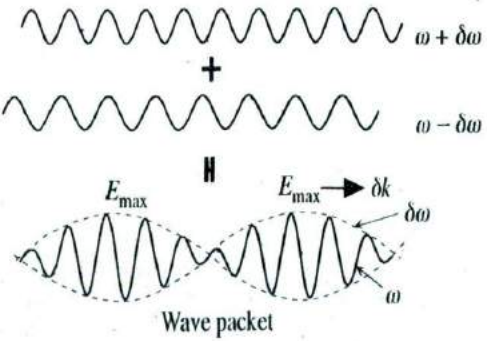


Figure 9.5 Two slightly different wavelength waves traveling in the same direction result in a wave packet that has an amplitude variation that travels at the group velocity.

The group velocity therefore defines the speed with which energy or information is propagated since it defines the speed of the envelope of the amplitude variation. The maximum electric field in Figure 9.5 advances with a velocity v_g, whereas the phase variations in the electric field are propagating at the phase velocity v.

Inasmuch as $\omega = vk$ and the phase velocity $v = c/n$, the group velocity in a medium can be readily evaluated from Equation 9.19. In a vacuum, obviously v is simply c and independent of the wavelength or k. Thus for waves traveling in a vacuum, $\omega = ck$ and the group velocity is

$$v_g(\text{vacuum}) = \frac{d\omega}{dk} = c = \text{Phase velocity} \qquad [9.20]$$

Group velocity in a vacuum

On the other hand, suppose that v depends on the wavelength or k by virtue of n being a function of the wavelength as in the case for glasses. Then,

$$\omega = vk = \left[\frac{c}{n(\lambda)}\right]\left(\frac{2\pi}{\lambda}\right) \qquad [9.21]$$

where $n = n(\lambda)$ is a function of the wavelength. The group velocity v_g in a medium, from differentiating Equation 9.21 in Equation 9.19, is approximately given by

$$v_g(\text{medium}) = \frac{d\omega}{dk} = \frac{c}{n - \lambda\dfrac{dn}{d\lambda}}$$

This can be written as

$$v_g(\text{medium}) = \frac{c}{N_g} \qquad [9.22]$$

Group velocity in a medium

where

$$N_g = n - \lambda\frac{dn}{d\lambda} \qquad [9.23]$$

Group index

is defined as the **group index** of the medium. Equation 9.23 defines the group refractive index N_g of a medium and determines the effect of the medium on the group velocity via Equation 9.22.

In general, for many materials the refractive index n and hence the group index N_g depend on the wavelength of light by virtue of the relative permittivity ε_r being frequency dependent. Then both the phase velocity v and the group velocity v_g depend on the wavelength and the medium is called a **dispersive medium.** The refractive index n and the group index N_g of pure SiO_2 (silica) glass are important parameters in optical fiber design in optical communications. Both of these parameters depend on the wavelength of light as shown in Figure 9.6. Around 1300 nm, N_g is at a minimum which means that for wavelengths close to 1300 nm, N_g is wavelength independent. Thus, light waves with wavelengths around 1300 nm travel with the same group velocity and do not experience dispersion. This phenomenon is significant in the propagation of light in glass fibers used in optical communications.

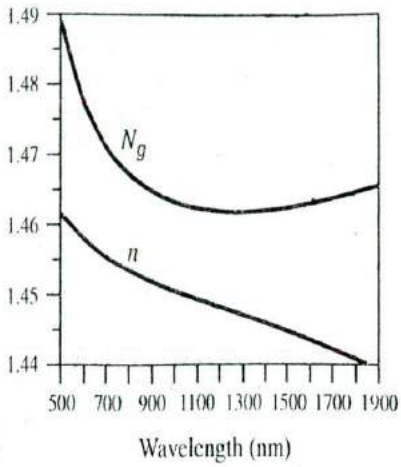


Figure 9.6 Refractive index n and the group index N_g of pure SiO_2 (silica) glass as a function of wavelength.

EXAMPLE 9.5

GROUP VELOCITY Consider two sinusoidal waves which are close in frequency, that is, waves of frequencies $\omega - \delta\omega$ and $\omega + \delta\omega$ as in Figure 9.5. Their wavevectors will be $k - \delta k$ and $k + \delta k$. The resultant wave will be

$$E_x(z, t) = E_o \cos[(\omega - \delta\omega)t - (k - \delta k)z] + E_o \cos[(\omega + \delta\omega)t - (k + \delta k)z]$$

By using the trigonometric identity $\cos A + \cos B = 2\cos\left[\frac{1}{2}(A - B)\right]\cos\left[\frac{1}{2}(A + B)\right]$ we arrive at

$$E_x(z, t) = 2E_o \cos[(\delta\omega)t - (\delta k)z]\cos[\omega t - kz]$$

As depicted in Figure 9.5, this represents a sinusoidal wave of frequency ω which is amplitude modulated by a very slowly varying sinusoidal of frequency $\delta\omega$. The system of waves, that is, the modulation, travels along z at a speed determined by the modulating term $\cos[(\delta\omega)t - (\delta k)z]$. The maximum in the field occurs when $[(\delta\omega)t - (\delta k)z] = 2m\pi = \text{constant}$ (m is an integer), which travels with a velocity

Group velocity

$$\frac{dz}{dt} = \frac{\delta\omega}{\delta k} \quad \text{or} \quad V_g = \frac{d\omega}{dk}$$

This is the group velocity of the waves, as stated in Equation 9.19, since it determines the speed of propagation of the maximum electric field along z.

EXAMPLE 9.6

GROUP AND PHASE VELOCITIES Consider a light wave traveling in a pure SiO_2 (silica) glass medium. If the wavelength of light is 1300 nm and the refractive index at this wavelength is 1.447, what is the phase velocity, group index (N_g), and group velocity (V_g)?

SOLUTION

The phase velocity is given by

$$V = \frac{c}{n} = \frac{3 \times 10^8 \text{ m s}^{-1}}{1.447} = 2.073 \times 10^8 \text{ m s}^{-1}$$

From Figure 9.6, at $\lambda = 1300$ nm, $N_g = 1.462$, so

$$v_g = \frac{c}{N_g} = \frac{3 \times 10^8 \text{ m s}^{-1}}{1.462} = 2.052 \times 10^8 \text{ m s}^{-1}$$

The group velocity is ~0.7 percent smaller than the phase velocity.

9.5 MAGNETIC FIELD: IRRADIANCE AND POYNTING VECTOR

Although we have considered the electric field component E_x of the EM wave, we should recall that the magnetic field (magnetic induction) component B_y always accompanies E_x in an EM wave propagation. In fact, if v is the phase velocity of an EM wave in an isotropic dielectric medium and n is the refractive index, then according to electromagnetism, at all times and anywhere in an EM wave,[5]

$$E_x = vB_y = \frac{c}{n} B_y \qquad [9.24]$$

Fields in an EM wave

where $v = (\varepsilon_o \varepsilon_r \mu_o)^{-1/2}$ and $n = \sqrt{\varepsilon_r}$. Thus, the two fields are simply and intimately related for an EM wave propagating in an isotropic medium. Any process that alters E_x also intimately changes B_y in accordance with Equation 9.24.

As the EM wave propagates in the direction of the wavevector $\mathbf{k}$ as shown in Figure 9.7, there is an energy flow in this direction. The wave brings with it electromagnetic energy. A small region of space where the electric field is E_x has an energy density, that is, energy per unit volume, given by $\frac{1}{2}\varepsilon_o \varepsilon_r E_x^2$. Similarly, a region of space where the magnetic field is B_y has an energy density $\frac{1}{2}B_y^2/\mu_o$. Since the two fields are related by Equation 9.24, the energy densities in the E_x and B_y fields are the same,

$$\frac{1}{2}\varepsilon_o \varepsilon_r E_x^2 = \frac{1}{2\mu_o} B_y^2 \qquad [9.25]$$

Energy densities in an EM wave

The total energy density in the wave is therefore $\varepsilon_o \varepsilon_r E_x^2$. Suppose that an ideal "energy meter" is placed in the path of the EM wave so that the receiving area A of this meter is perpendicular to the direction of propagation. In a time interval Δt, a portion of the wave of spatial length $v\,\Delta t$ crosses A as shown in Figure 9.7. Thus, a volume $Av\,\Delta t$ of the EM wave crosses A in time Δt. The energy in this volume consequently becomes received. If S is the EM power flow per unit area,

S = Energy flow per unit time per unit area

giving,

$$S = \frac{(Av\,\Delta t)\left(\varepsilon_o \varepsilon_r E_x^2\right)}{A\,\Delta t} = v\varepsilon_o \varepsilon_r E_x^2 = v^2 \varepsilon_o \varepsilon_r E_x B_y \qquad [9.26]$$

[5] This is actually a statement of Faraday's law for EM waves. In vector notation it is often expressed as $\omega \mathbf{B} = \mathbf{k} \times \mathbf{E}$

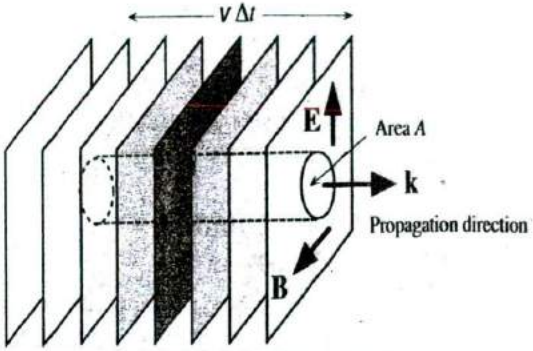


Figure 9.7 A plane EM wave traveling along **k** crosses an area A at right angles to the direction of propagation. In time Δt, the energy in the cylindrical volume $Av\,\Delta t$ (shown dashed) flows through A.

In an isotropic medium, the energy flow is in the direction of wave propagation. If we use the vectors **E** and **B** to represent the electric and magnetic fields in the EM wave, then the wave propagates in a direction **E** × **B**, because this direction is perpendicular to both **E** and **B**. The EM power flow per unit area in Equation 9.26 can be written as

Poynting vector

$$\mathbf{S} = v^2 \varepsilon_o \varepsilon_r \mathbf{E} \times \mathbf{B} \qquad [9.27]$$

where S, called the **Poynting vector**, represents the energy flow per unit time per unit area in a direction determined by **E** × **B** (direction of propagation). Its magnitude, power flow per unit area, is called the **irradiance.**[6]

The field E_x at the receiver location (say, $z = z_1$) varies sinusoidally which means that the energy flow also varies sinusoidally. The irradiance in Equation 9.26 is the **instantaneous irradiance.** If we write the field as $E_x = E_o \sin(\omega t)$ and then calculate the average irradiance by averaging S over one period, we would find the **average irradiance,**

Average irradiance (intensity)

$$I = S_{\text{average}} = \frac{1}{2} v \varepsilon_o \varepsilon_r E_o^2 \qquad [9.28]$$

Since $v = c/n$ and $\varepsilon_r = n^2$ we can write Equation 9.28 as

Average irradiance (intensity)

$$I = S_{\text{average}} = \frac{1}{2} c \varepsilon_o n E_o^2$$

$$= (1.33 \times 10^{-3}) n E_o^2 \qquad [9.29]$$

The instantaneous irradiance can only be measured if the power meter can respond more quickly than the oscillations of the electric field, and since this is in the

[6] The term *intensity* is widely used and interpreted by many engineers as power flow per unit area even though the strictly correct term is *irradiance*. Many optoelectronic data books simply use intensity to mean irradiance.

optical frequencies range, all practical measurements invariably yield the average irradiance because all detectors have a response rate much slower than the frequency of the wave.

9.6 SNELL'S LAW AND TOTAL INTERNAL REFLECTION (TIR)

We consider a traveling plane EM wave in a medium (1) of refractive index n_1 propagating toward a medium (2) with a refractive index n_2. Constant phase fronts are joined with broken lines, and the wavevector $\mathbf{k}_i$ is perpendicular to the wave fronts as shown in Figure 9.8. When the wave reaches the plane boundary between the two media, a transmitted wave in medium 2 and a reflected wave in medium 1 appear. The transmitted wave is called the **refracted light.** The angles, θ_i, θ_t, θ_r define the directions of the incident, transmitted, and reflected waves, respectively, with respect to the normal to the boundary plane as shown in Figure 9.8. The wavevectors of the reflected and transmitted waves are denoted as $\mathbf{k}_r$ and $\mathbf{k}_t$, respectively. Since both the incident and reflected waves are in the same medium, the magnitudes of $\mathbf{k}_r$ and $\mathbf{k}_i$ are the same, $k_r = k_i$.

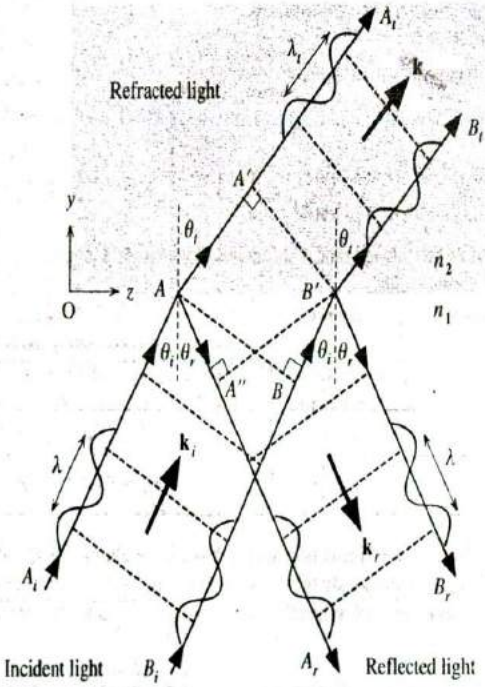


Figure 9.8 A light wave traveling in a medium with a greater refractive index ($n_1 > n_2$) suffers reflection and refraction at the boundary.

Simple arguments based on constructive interference can be used to show that there can only be one reflected wave that occurs at an angle equal to the incidence angle. The two waves along A_i and B_i are in phase. When these waves are reflected to become waves A_r and B_r, then they must still be in phase, otherwise they will interfere destructively and destroy each other. The only way the two waves can stay in phase is if $\theta_r = \theta_i$. All other angles lead to the waves A_r and B_r being out of phase and interfering destructively.

The refracted waves A_t and B_t are propagating in a medium of refracted index $n_2 (< n_1)$ that is different than n_1. Hence the waves A_t and B_t have different velocities than A_i and B_i. We consider what happens to a wavefront such as AB, corresponding perhaps to the maximum field, as it propagates from medium 1 to 2. We recall that the points A and B on this front are always in phase. During the time it takes for the phase B on wave B_i to reach B', phase A on wave A_t has progressed to A'. The wavefront AB thus becomes the front $A'B'$ in medium 2. Unless the two waves at A' and B' still have the same phase, there will be no transmitted wave. A' and B' points on the front are only in phase for one particular transmitted angle θ_t.

If it takes time t for the phase at B on wave B_i to reach B', then $BB' = v_1 t = ct/n_1$. During this time t, the phase A has progressed to A' where $AA' = v_2 t = ct/n_2$. A' and B' belong to the same front just like A and B, so AB is perpendicular to $\mathbf{k}_i$ in medium 1 and $A'B'$ is perpendicular to $\mathbf{k}_t$ in medium 2. From geometrical considerations, $AB' = BB'/\sin\theta_i$ and $AB' = AA'/\sin\theta_t$, so

$$AB' = \frac{v_1 t}{\sin\theta_i} = \frac{v_2 t}{\sin\theta_t}$$

or

Snell's law

$$\frac{\sin\theta_i}{\sin\theta_t} = \frac{v_1}{v_2} = \frac{n_2}{n_1} \qquad [9.30]$$

This is **Snell's law**[7] which relates the angles of incidence and refraction to the refractive indices of the media.

If we consider the reflected wave, the wave front AB becomes $A''B'$ in the reflected wave. In time t, phase B moves to B' and A moves to A''. Since they must still be in phase to constitute the reflected wave, BB' must be equal to AA''. Suppose it takes time t for the wavefront B to move to B' (or A to A''). Then, since $BB' = AA'' = v_1 t$, from geometrical considerations,

$$AB' = \frac{v_1 t}{\sin\theta_i} = \frac{v_1 t}{\sin\theta_r}$$

so that $\theta_i = \theta_r$. The angles of incidence and reflection are the same.

When $n_1 > n_2$, then obviously the transmitted angle is greater than the incidence angle as apparent in Figure 9.8. When the refraction angle θ_t reaches 90°, the incidence

[7] Willebrord van Roijen Snell (1581–1626), a Dutch physicist and mathematician, was born in Leiden and eventually became a professor at Leiden University. He obtained his refraction law in 1621 which was published by Rène Descartes in France in 1637; it is not known whether Descartes knew of Snell's law or formulated it independently.

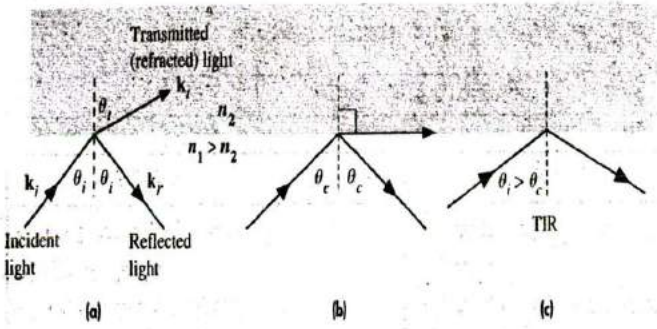


Figure 9.9 Light wave traveling in a more dense medium strikes a less dense medium. Depending on the incidence angle with respect to θ_c, determined by the ratio of the refractive indices, the wave may be transmitted (refracted) or reflected.
(a) $\theta_i < \theta_c$.
(b) $\theta_i = \theta_c$.
(c) $\theta_i > \theta_c$ and total internal reflection (TIR).

angle is called the **critical angle** θ_c which is given by

Critical angle for total internal reflection (TIR).

$$\sin\theta_c = \frac{n_2}{n_1} \qquad [9.31]$$

When the incidence angle θ_i exceeds θ_c, then there is no transmitted wave but only a reflected wave. The latter phenomenon is called **total internal reflection** (TIR). The effect of increasing the incidence angle is shown in Figure 9.9. It is the TIR phenomenon that leads to the propagation of waves in a dielectric medium surrounded by a medium of smaller refractive index as in optical waveguides (*e.g.*, optical fibers).

EXAMPLE 9.7

OPTICAL FIBERS IN COMMUNICATIONS Figure 9.10 shows a simplified view of a modern optical communications system. Information is converted into a digital signal (*e.g.*, current pulses) which drives a light emitter such as a semiconductor laser. The light pulses from the emitter are coupled into an **optical fiber**, which acts as a light guide. The optical fiber is a very thin glass fiber [made of silica (SiO_2)], almost as thin as your hair, that is able to optically guide the light pulses to their destination. The photodetector at the destination converts the light pulses into an electric signal, which is then decoded into the original information.

The **core** of the optical fiber has a higher refractive index than the surrounding region, which is called the **cladding** as shown in Figure 9.10. Optical fibers for short-distance applications (*e.g.*, communications in local area networks within a large building) usually have a core region that has a diameter of about 100 μm, and the whole fiber would be about 150–200 μm in diameter. The core and cladding refractive indices, n_1 and n_2, respectively, are normally only 1–3 percent different. The light propagates along the fiber core because light rays experience total internal reflections at the core-cladding interface as shown in Figure 9.10. Only those light rays that can exercise TIR travel along the fiber length and can reach the destination. Consider a fiber with n_1(core) = 1.455, and n_2(cladding) = 1.440. The critical angle for a ray traveling

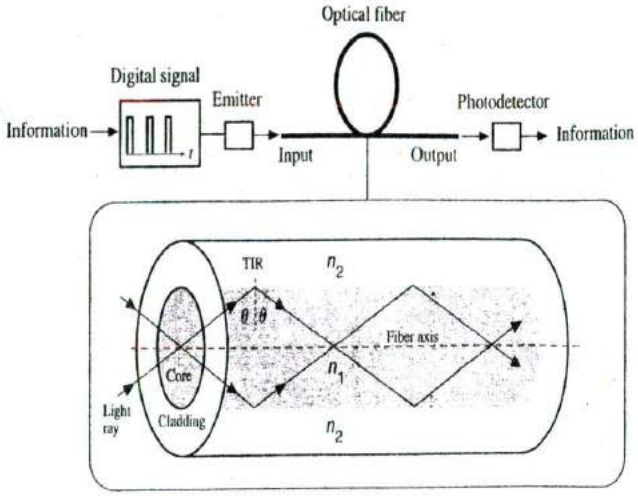
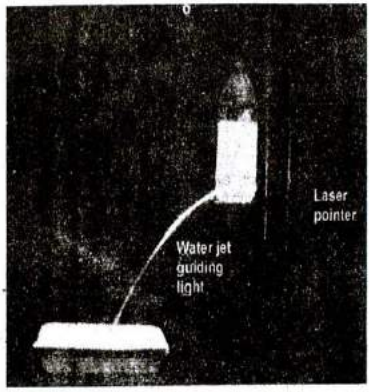


Figure 9.10 An optical fiber link for transmitting digital information in communications. The fiber core has a higher refractive index, so the light travels along the fiber inside the fiber core by total internal reflection at the core–cladding interface.



A small hole is made in a plastic bottle full of water to generate a water jet. When the hole is illuminated with a laser beam (from a green laser pointer), the light is guided by total internal reflections along the jet to the tray. Light guiding by a water jet was demonstrated by John Tyndall in 1854 to the Royal Institution. (Water with air bubbles was used to increase the visibility of light, since air bubbles scatter light.)

in the core is

$$\theta_c = \arcsin\left(\frac{n_2}{n_1}\right) = \arcsin\left(\frac{1.440}{1.455}\right) = 81.8°$$

Those light rays that have angles $\theta > \theta_c$ satisfy TIR and can propagate along the fiber.[8] Notice that the ray angles with respect to the fiber axis are less than 8.2°.

[8] The light propagation in an optical fiber is much more complicated than the simple zigzagging of light rays with TIRs at the core-cladding interface. The waves in the core have to satisfy not only TIR but also have to avoid destructive interference so that they are not destroyed as they travel along the guide; see for example, S. O. Kasap, *Optoelectronics and Photonics: Principles and Practices*, Upper Saddle River: Prentice Hall, 2001, chap. 2.

9.7 FRESNEL'S EQUATIONS

9.7.1 AMPLITUDE REFLECTION AND TRANSMISSION COEFFICIENTS

Although the ray picture with constant phase wave fronts is useful in understanding refraction and reflection, to obtain the magnitude of the reflected and refracted waves and their relative phases, we need to consider the electric field in the light wave. The electric field in the wave must be perpendicular to the direction of propagation as shown in Figure 9.11. We can resolve the field E_i of the incident wave into two components, one in the plane of incidence $E_{i,\parallel}$ and the other perpendicular to the plane of incidence $E_{i,\perp}$. The **plane of incidence** is defined as the plane containing the incident and the reflected rays which in Figure 9.11 corresponds to the plane of the paper.[9] Similarly for both the reflected and transmitted waves, we will have field components parallel and perpendicular to the plane of incidence, *i.e.*, $E_{r,\parallel}$, $E_{r,\perp}$ and $E_{t,\parallel}$, $E_{t,\perp}$.

As apparent from Figure 9.11, the incident, transmitted, and reflected waves all have a wavevector component along the z direction; that is, they have an effective velocity along z. The fields $E_{i,\perp}$, $E_{r,\perp}$, and $E_{t,\perp}$ are all perpendicular to the z direction. These waves are called **transverse electric field** (TE) waves. On the other hand, waves with $E_{i,\parallel}$, $E_{r,\parallel}$, and $E_{t,\parallel}$ only have their magnetic field components perpendicular to the z direction and these are called **transverse magnetic field** (TM) waves.

We will describe the incident, reflected, and refracted waves by the exponential representation of a traveling wave, *i.e.*,

Incident wave

$$E_i = E_{io} \exp j(\omega t - \mathbf{k}_i \cdot \mathbf{r}) \qquad [9.32]$$

Reflected wave

$$E_r = E_{ro} \exp j(\omega t - \mathbf{k}_r \cdot \mathbf{r}) \qquad [9.33]$$

Transmitted wave

$$E_t = E_{to} \exp j(\omega t - \mathbf{k}_t \cdot \mathbf{r}) \qquad [9.34]$$

where $\mathbf{r}$ is the position vector; the wavevectors $\mathbf{k}_i$, $\mathbf{k}_r$, and $\mathbf{k}_t$ describe, respectively, the directions of the incident, reflected, and transmitted waves; and E_{io}, E_{ro}, and E_{to} are the respective amplitudes. Any phase changes such as ϕ_r and ϕ_t in the reflected and transmitted waves with respect to the phase of the incident wave are incorporated into the complex amplitudes E_{ro} and E_{to}. Our objective is to find E_{ro} and E_{to} with respect to E_{io}.

We should note that similar equations can be stated for the magnetic field components in the incident, reflected, and transmitted waves, but these will be perpendicular to the corresponding electric fields. The electric and magnetic fields anywhere on the wave must be perpendicular to each other as a requirement of electromagnetic wave theory. This means that with $E_\parallel$ in the EM wave we have a magnetic field $B_\perp$ associated

[9] The definitions of the field components follow those of S. G. Lipson et al., *Optical Physics*, 3rd ed., Cambridge, MA, Cambridge University Press, 1995, and Grant Fowles, *Introduction to Modern Optics*, 2nd ed., New York, Dover Publications, Inc., 1975, whose clear treatments of this subject are highly recommended. The majority of the authors use a different convention which leads to different signs later in the equations; Fresnel's equations are related to the specific electric field directions from which they are derived.

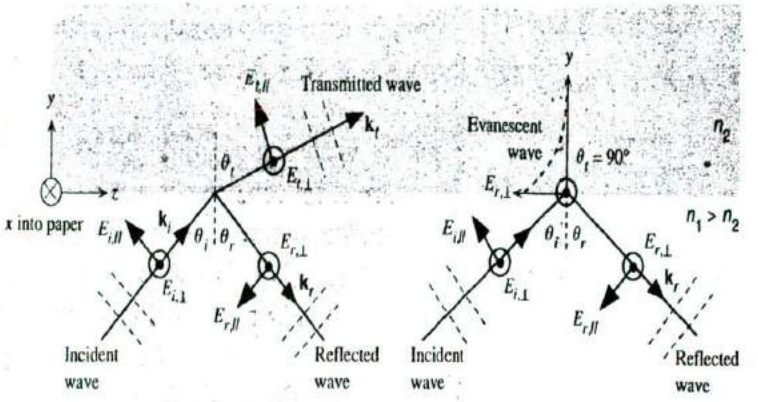


Figure 9.11 Light wave traveling in a more dense medium strikes a less dense medium.
The plane of incidence is the plane of the paper and is perpendicular to the flat interface between the two media. The electric field is normal to the direction of propagation. It can be resolved into perpendicular (⊥) and parallel (∥) components.

with it such that $B_\perp = (n/c)E_\parallel$. Similarly $E_\perp$ will have a magnetic field $B_\parallel$ associated with it such that $B_\parallel = (n/c)E_\perp$.

There are two useful fundamental rules in electromagnetism that govern the behavior of the electric and magnetic fields at a boundary between two dielectric media which we can arbitrarily label as 1 and 2. These rules are called boundary conditions. The first states that the electric field that is tangential to the boundary surface $E_{\text{tangential}}$ must be continuous across the boundary from medium 1 to 2, *i.e.*, at the boundary $y = 0$ in Figure 9.11,

Boundary condition

$$E_{\text{tangential}}(1) = E_{\text{tangential}}(2) \qquad [9.35]$$

The second rule is that the tangential component of the magnetic field $B_{\text{tangential}}$ to the boundary must be likewise continuous from medium 1 to 2 provided that the two media are nonmagnetic (relative permeability $\mu_r = 1$),

Boundary condition

$$B_{\text{tangential}}(1) = B_{\text{tangential}}(2) \qquad [9.36]$$

Using these boundary conditions for the fields at $y = 0$, and the relationship between the electric and magnetic fields, we can find the reflected and transmitted waves in terms of the incident wave. The boundary conditions can only be satisfied if the reflection and incidence angles are equal, $\theta_r = \theta_i$, and the angles for the transmitted and incident waves obey Snell's law, $n_1 \sin\theta_i = n_2 \sin\theta_t$.

Applying the boundary conditions to the EM wave going from medium 1 to 2, the amplitudes of the reflected and transmitted waves can be readily obtained in terms of

n_1, n_2, and the incidence angle θ_i alone.[10] These relationships are called **Fresnel's equations.** If we define $n = n_2/n_1$, as the relative refractive index of medium 2 to that of 1, then the **reflection** and **transmission coefficients** for $E_\perp$ are

$$r_\perp = \frac{E_{r0,\perp}}{E_{i0,\perp}} = \frac{\cos\theta_i - (n^2 - \sin^2\theta_i)^{1/2}}{\cos\theta_i + (n^2 - \sin^2\theta_i)^{1/2}} \qquad [9.37]$$

Reflection coefficient

and

$$t_\perp = \frac{E_{r0,\perp}}{E_{i0,\perp}} = \frac{2\cos\theta_i}{\cos\theta_i + (n^2 - \sin^2\theta_i)^{1/2}} \qquad [9.38]$$

Transmission coefficient

There are corresponding coefficients for the $E_\parallel$ fields with corresponding **reflection** and **transmission coefficients** $r_\parallel$ and $t_\parallel$:

$$r_\parallel = \frac{E_{r0,\parallel}}{E_{i0,\parallel}} = \frac{(n^2 - \sin^2\theta_i)^{1/2} - n^2\cos\theta_i}{(n^2 - \sin^2\theta_i)^{1/2} + n^2\cos\theta_i} \qquad [9.39]$$

Reflection coefficient

$$t_\parallel = \frac{E_{r0,\parallel}}{E_{i0,\parallel}} = \frac{2n\cos\theta_i}{n^2\cos\theta_i + (n^2 - \sin^2\theta_i)^{1/2}} \qquad [9.40]$$

Transmission coefficient

Further, the reflection and transmission coefficients are related by

$$r_\parallel + n t_\parallel = 1 \quad \text{and} \quad r_\perp + 1 = t_\perp \qquad [9.41]$$

Transmission and reflection

The significance of these equations is that they allow the amplitudes and phases of the reflected and transmitted waves to be determined from the coefficients $r_\perp$, $r_\parallel$, $t_\parallel$, and $t_\perp$. For convenience we take E_{io} to be a real number so that the phase angles of $r_\perp$ and $t_\perp$ correspond to the **phase changes** measured with respect to the incident wave. For example, if $r_\perp$ is a complex quantity, then we can write this as $r_\perp = |r_\perp| \exp(-j\phi_\perp)$ where $|r_\perp|$ and $\phi_\perp$ represent the relative amplitude and phase of the reflected wave with respect to the incident wave for the field perpendicular to the plane of incidence. Of course, when $r_\perp$ is a real quantity, then a positive number represents no phase shift and a negative number is a phase shift of 180° (or π). As with all waves, a negative sign corresponds to a 180° phase shift. Complex coefficients can only be obtained from Fresnel's equations if the terms under the square roots become negative, and this can only happen when $n < 1$ (or $n_1 > n_2$), and also when $\theta_i > \theta_c$, the critical angle. Thus, phase changes other than 0 or 180° occur only when there is total internal reflection.

Figure 9.12a shows how the magnitudes of the reflection coefficients $|r_\perp|$ and $|r_\parallel|$ vary with the incidence angle θ_i for a light wave traveling from a more dense medium, $n_1 = 1.44$, to a less dense medium, $n_2 = 1.00$, as predicted by Fresnel's equations. Figure 9.12b shows the changes in the phase of the reflected wave, $\phi_\perp$ and $\phi_\parallel$, with θ_i. The critical angle θ_c as determined from $\sin\theta_c = n_2/n_1$ in this case is 44°. It is clear that for incidence close to normal (small θ_i), there is no phase change in the reflected wave. For

[10] These equations are readily available in any electromagnetism textbook. Their derivation from the two boundary conditions involves extensive algebraic manipulation which we will not carry out here. The electric and magnetic field components on both sides of the boundary are resolved tangentially to the boundary surface and the boundary conditions are then applied. We then use such relations as $\cos\theta_t = (1 - \sin\theta_t)^{1/2}$ and $\sin\theta_t$ as determined by Snell's law, etc.

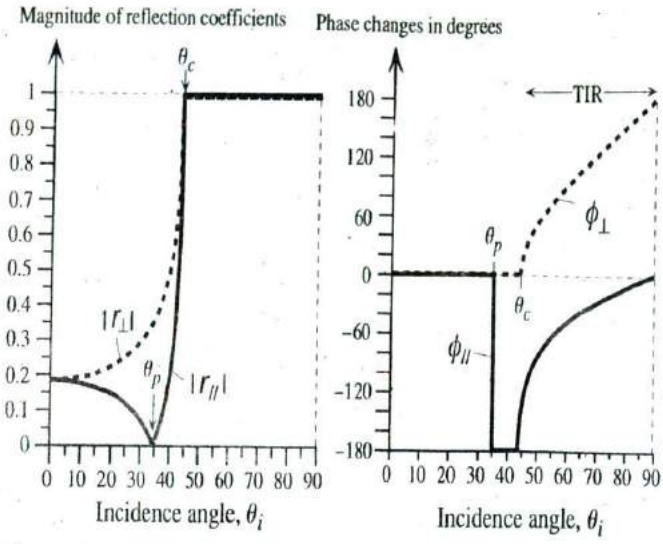


Figure 9.12 Internal reflection.

(a) Magnitude of the reflection coefficients $r_{\parallel}$ and $r_{\perp}$ versus the angle of incidence θ_i for $n_1 = 1.44$ and

$\phi_\perp$ versus incidence angle.

example, putting normal incidence ($\theta_i = 0$) into Fresnel's equations, we find

Normal incidence

$$r_{\parallel} = r_{\perp} = \frac{n_1 - n_2}{n_1 + n_2} \qquad [9.42]$$

This is a positive quantity for $n_1 > n_2$ which means that the reflected wave suffers no phase change. This is confirmed by $\phi_\perp$ and $\phi_{\parallel}$ in Figure 9.12b. As the incidence angle increases, eventually $r_{\parallel}$ becomes zero at an angle of about 35°. We can find this special incidence angle, labeled as θ_p, by solving the Fresnel equation, Equation 9.39, for $r_{\parallel} = 0$. The field in the reflected wave is then always perpendicular to the plane of incidence and hence well-defined. This special angle is called the **polarization angle** or **Brewster's angle** and from Equation 9.39 is given by

Brewster's polarization angle

$$\tan\theta_p = \frac{n_2}{n_1} \qquad [9.43]$$

The reflected wave is then said to be **linearly polarized** because it contains *electric field oscillations that are contained within a well-defined plane* which is perpendicular to the plane of incidence and also to the direction of propagation. Electric field oscillations in **unpolarized light,** on the other hand, can be in any one of an infinite number of directions that are perpendicular to the direction of propagation. In linearly polarized light, however, the field oscillations are contained within a well-defined plane. Light emitted from many light sources such as a tungsten light bulb or an LED diode is

unpolarized and the field is randomly oriented in a direction that is perpendicular to the direction of propagation.

For incidence angles greater than θ_p but smaller than θ_c, Fresnel's equation, Equation 9.39, gives a negative number for $r_\parallel$ which indicates a phase shift of 180° as shown in $\phi_\parallel$ in Figure 9.12b. The magnitudes of both $r_\parallel$ and $r_\perp$ increase with θ_i as apparent in Figure 9.12a. At the critical angle and beyond (past 44° in Figure 9.12), *i.e.*, when $\theta_i \geq \theta_c$, the magnitudes of both $r_\parallel$ and $r_\perp$ go to unity, so the reflected wave has the same amplitude as the incident wave. The incident wave has suffered **total internal reflection** (TIR). When $\theta_i > \theta_c$, in the presence of TIR, the Equations 9.37 to 9.40 are complex quantities because then $\sin\theta_i > n$ and the terms under the square roots become negative. The reflection coefficients become complex quantities of the type $r_\perp = 1 \cdot \exp(-j\phi_\perp)$ and $r_\parallel = 1 \cdot \exp(-j\phi_\parallel)$ with the phase angles $\phi_\perp$ and $\phi_\parallel$ being other than 0 or 180°. The reflected wave therefore suffers phase changes $\phi_\perp$ and $\phi_\parallel$ in the components $E_\perp$ and $E_\parallel$. These phase changes depend on the incidence angle, as apparent in Figure 9.12b, and on n_1 and n_2.

Examination of Equation 9.37 for $r_\perp$ shows that for $\theta_i > \theta_c$, we have $|r_\perp| = 1$, but the phase change $\phi_\perp$ is given by

$$\tan\left(\frac{1}{2}\phi_\perp\right) = \frac{(\sin^2\theta_i - n^2)^{1/2}}{\cos\theta_i} \qquad [9.44]$$

Phase change in TIR

For the $E_\parallel$ component, the phase change $\phi_\parallel$ is given by

$$\tan\left(\frac{1}{2}\phi_\parallel + \frac{1}{2}\pi\right) = \frac{(\sin^2\theta_i - n^2)^{1/2}}{n^2\cos\theta_i} \qquad [9.45]$$

Phase change in TIR

We can summarize that, in internal reflection ($n_1 > n_2$), the amplitude of the reflected wave from TIR is equal to the amplitude of the incident wave but its phase has shifted by an amount determined by Equations 9.44 and 9.45.[11] The fact that $\phi_\parallel$ has an additional π shift which makes $\phi_\parallel$ negative for $\theta_i > \theta_c$ is due to the choice for the direction of the reflected optical field $E_{r,\parallel}$ in Figure 9.11. (This π shift can be ignored if we simply invert $E_{r,\parallel}$.)

The reflection coefficients in Figure 9.12 considered the case in which $n_1 > n_2$. When light approaches the boundary from the higher index side, that is, $n_1 > n_2$, the reflection is said to be **internal reflection** and *at normal incidence there is no phase change*. On the other hand, if light approaches the boundary from the lower index side, that is, $n_1 < n_2$, then it is called **external reflection.** Thus in external reflection light becomes reflected by the surface of an optically denser (higher refractive index) medium. There is an important difference between the two. Figure 9.13 shows how the reflection coefficients $r_\perp$ and $r_\parallel$ depend on the incidence angle θ_i for external reflection ($n_1 = 1$ and $n_2 = 1.44$). At normal incidence, both coefficients are negative, which means that *in external reflection at normal incidence there is a phase shift of 180°*. Further, $r_\parallel$ goes through zero at the **Brewster angle** θ_p given by Equation 9.43. At this angle of incidence, the reflected wave is polarized in the $E_\perp$ component only. Transmitted light in both internal reflection (when $\theta_i < \theta_c$) and external reflection does not experience a phase shift.

[11] It should be apparent that the concepts and the resulting equations apply to a well-defined linearly polarized light wave.

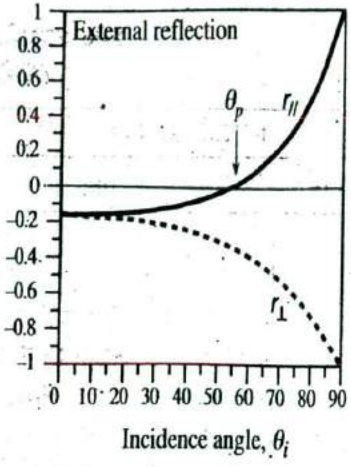


Figure 9.13 The reflection coefficients $r_{||}$ and $r_\perp$ versus angle of incidence θ_i for $n_1 = 1.00$ and $n_2 = 1.44$.

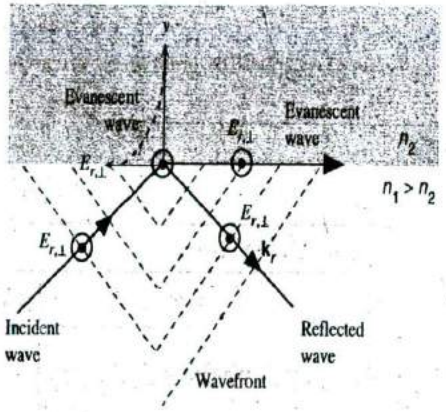


Figure 9.14 When $\theta_i > \theta_c$, for a plane wave that is reflected, there is an evanescent wave at the boundary propagating along z.

What happens to the transmitted wave when $\theta_i > \theta_c$? According to the boundary conditions, there must still be an electric field in medium 2; otherwise, the boundary conditions cannot be satisfied. When $\theta_i > \theta_c$, the field in medium 2 is a wave that travels near the surface of the boundary along the z direction as depicted in Figure 9.14. The wave is called an **evanescent wave** and advances along z with its field decreasing as we move into medium 2, *i.e.*,

Evanescent wave

$$E_{t,\perp}(y, z, t) \propto e^{-\alpha_2 y} \exp j(\omega t - k_{iz} z) \qquad [9.46]$$

where $k_{iz} = k_i \sin\theta_i$ is the wavevector of the incident wave along the z axis, and α_2 is an **attenuation coefficient** for the electric field penetrating into medium 2,

Attenuation evanescent wave

$$\alpha_2 = \frac{2\pi n_2}{\lambda}\left[\left(\frac{n_1}{n_2}\right)^2 \sin^2\theta_i - 1\right]^{1/2} \qquad [9.47]$$

where λ is the free-space wavelength. According to Equation 9.46, the evanescent wave travels along z and has an amplitude that decays exponentially as we move from the boundary into medium 2 (along y) as shown in Figure 9.11b. The field of the evanescent wave is e^{-1} in medium 2 when $y = 1/\alpha_2 = \delta$ which is called the **penetration depth.** It is not difficult to show that the evanescent wave is correctly predicted by Snell's law when $\theta_i > \theta_c$. The evanescent wave propagates along the boundary (along z) with the same speed as the z component velocity of the incident and reflected waves. In Equations 9.32 to 9.34 we had assumed that the incident and reflected waves were *plane waves*, that is, of infinite extent. If we were to extend the plane wavefronts on the reflected wave, these would cut the boundary as shown in Figure 9.14. The evanescent wave traveling along z can be thought of as arising from these plane wavefronts at the boundary as in Figure 9.14. (The evanescent wave is important in light propagation in optical waveguides such as in optical fibers.) If the incident wave is a narrow beam of light (*e.g.*, from a laser pointer), then the reflected beam would have the same cross section. There would still be an evanescent wave at the boundary, but it would exist only within the cross-sectional area of the reflected beam at the boundary.

9.7.2 INTENSITY, REFLECTANCE, AND TRANSMITTANCE

It is frequently necessary to calculate the intensity or irradiance of the reflected and transmitted waves when light traveling in a medium of index n_1 is incident at a boundary where the refractive index changes to n_2. In some cases we are simply interested in normal incidence where $\theta_i = 0°$. For example, in laser diodes light is reflected from the ends of an optical cavity where there is a change in the refractive index.

Reflectance R measures the intensity of the reflected light with respect to that of the incident light and can be defined separately for electric field components parallel and perpendicular to the plane of incidence. The reflectances $R_\perp$ and $R_\parallel$ are defined by

$$R_\perp = \frac{|E_{ro,\perp}|^2}{|E_{io,\perp}|^2} = |r_\perp|^2 \quad \text{and} \quad R_\parallel = \frac{|E_{ro,\parallel}|^2}{|E_{io,\parallel}|^2} = |r_\parallel|^2 \qquad [9.48]$$

From Equations 9.37 to 9.40 with normal incidence, these are simply given by

Reflectance at normal incidence

$$R = R_\perp = R_\parallel = \left(\frac{n_1 - n_2}{n_1 + n_2}\right)^2 \qquad [9.49]$$

Since a glass medium has a refractive index of around 1.5, this means that typically 4 percent of the incident radiation on an air–glass surface will be reflected back.

Transmittance T relates the intensity of the transmitted wave to that of the incident wave in a similar fashion to the reflectance. We must, however, consider that the transmitted wave is in a different medium and further its direction with respect to the boundary is also different by virtue of refraction. For normal incidence, the incident and transmitted beams are normal and the transmittances are defined and given by

$$T_\perp = \frac{n_2|E_{to,\perp}|^2}{n_1|E_{io,\perp}|^2} = \left(\frac{n_2}{n_1}\right)|t_\perp|^2 \quad \text{and} \quad T_\parallel = \frac{n_2|E_{to,\parallel}|^2}{n_1|E_{io,\parallel}|^2} = \left(\frac{n_2}{n_1}\right)|t_\parallel|^2 \qquad [9.50]$$

Transmittance at normal incidence

or

$$T = T_\perp = T_\parallel = \frac{4n_1n_2}{(n_1+n_2)^2} \qquad [9.51]$$

Further, the fraction of light reflected and fraction transmitted must add to unity. Thus $R + T = 1$.

EXAMPLE 9.8 **REFLECTION OF LIGHT FROM A LESS DENSE MEDIUM (INTERNAL REFLECTION)** A ray of light which is traveling in a glass medium of refractive index $n_1 = 1.460$ becomes incident on a less dense glass medium of refractive index $n_2 = 1.440$. Suppose that the free-space wavelength (λ) of the light ray is 1300 nm.

a. What should be the minimum incidence angle for TIR?

b. What is the phase change in the reflected wave when $\theta_i = 87°$ and when $\theta_i = 90°$?

c. What is the penetration depth of the evanescent wave into medium 2 when $\theta_i = 80°$ and when $\theta_i = 90°$?

SOLUTION

a. The critical angle θ_c for TIR is given by $\sin\theta_c = n_2/n_1 = 1.440/1.460$, so $\theta_c = 80.51°$.

b. Since the incidence angle $\theta_i > \theta_c$, there is a phase shift in the reflected wave. The phase change in $E_{r,\perp}$ is given by $\phi_\perp$. With $n_1 = 1.460$, $n_2 = 1.440$, and $\theta_i = 87°$,

$$\tan\left(\frac{1}{2}\phi_\perp\right) = \frac{(\sin^2\theta_i - n^2)^{1/2}}{\cos\theta_i} = \frac{\left[\sin^2(87°) - \left(\frac{1.440}{1.460}\right)^2\right]^{1/2}}{\cos(87°)}$$

$$= 2.989 = \tan\left[\tfrac{1}{2}(143.0°)\right]$$

so the phase change is 143°. For the $E_{r,\parallel}$ component, the phase change is

$$\tan\left(\frac{1}{2}\phi_\parallel + \frac{1}{2}\pi\right) = \frac{(\sin^2\theta_i - n^2)^{1/2}}{n^2\cos\theta_i} = \frac{1}{n^2}\tan\left(\frac{1}{2}\phi_\perp\right)$$

so

$$\tan\left(\frac{1}{2}\phi_\parallel + \frac{1}{2}\pi\right) = \left(\frac{n_1}{n_2}\right)^2 \tan\left(\frac{\phi_\perp}{2}\right) = \left(\frac{1.460}{1.440}\right)^2 \tan\left[\frac{1}{2}(143^\circ)\right]$$

which gives

$$\phi_\parallel = 143.95° - 180° = -36.05°$$

We can repeat the calculation with $\theta_i = 90°$ to find $\phi_\perp = 180°$ and $\phi_\parallel = 0°$.

Note that as long as $\theta_i > \theta_c$, the magnitude of the reflection coefficients are unity. Only the phase changes.

c. The amplitude of the evanescent wave as it penetrates into medium 2 is

$$E_{t,\perp}(y, t) \approx E_{to,\perp}\exp(-\alpha_2 y)$$

We ignore the z dependence, $\exp j(\omega t - k_z z)$, as this only gives a propagating property along z. The field strength drops to e^{-1} when $y = 1/\alpha_2 = \delta$, which is called the **penetration**

depth. The attenuation constant α_2 is

$$\alpha_2 = \frac{2\pi n_2}{\lambda}\left[\left(\frac{n_1}{n_2}\right)^2 \sin^2\theta_i - 1\right]^{1/2}$$

i.e.,

$$\alpha_2 = \frac{2\pi(1.440)}{(1300 \times 10^{-9}\ \text{m})}\left[\left(\frac{1.460}{1.440}\right)^2 \sin^2(87^\circ) - 1\right]^{1/2} = 1.104 \times 10^6\ \text{m}^{-1}$$

so the penetration depth is $\delta = 1/\alpha_2 = 1/(1.104 \times 10^6\ \text{m}) = 9.06 \times 10^{-7}$ m, or 0.906 μm. For 90°, repeating the calculation we find $\alpha_2 = 1.164 \times 10^6\ \text{m}^{-1}$, so $\delta = 1/\alpha_2 = 0.859$ μm. We see that the penetration is greater for smaller incidence angles. The values for the refractive indices and wavelength are typical of those values found in optical fiber communications.

EXAMPLE 9.9

REFLECTION AT NORMAL INCIDENCE. INTERNAL AND EXTERNAL REFLECTION Consider the reflection of light at normal incidence on a boundary between a glass medium of refractive index 1.5 and air of refractive index 1.

a. If light is traveling from air to glass, what is the reflection coefficient and the intensity of the reflected light with respect to that of the incident light?

b. If light is traveling from glass to air, what is the reflection coefficient and the intensity of the reflected light with respect to that of the incident light?

c. What is the polarization angle in the external reflection in part (*a*)? How would you make a polaroid device that polarizes light based on the polarization angle?

SOLUTION

a. The light travels in air and becomes partially reflected at the surface of the glass which corresponds to external reflection. Thus $n_1 = 1$ and $n_2 = 1.5$. Then,

$$r_\parallel = r_\perp = \frac{n_1 - n_2}{n_1 + n_2} = \frac{1 - 1.5}{1 + 1.5} = -0.2$$

This is negative which means that there is a 180° phase shift. The reflectance (R), which gives the fractional reflected power, is

$$R = r_\parallel^2 = 0.04 \quad \text{or} \quad 4\%$$

b. The light travels in glass and becomes partially reflected at the glass–air interface which corresponds to internal reflection. Thus $n_1 = 1.5$ and $n_2 = 1$. Then,

$$r_\parallel = r_\perp = \frac{n_1 - n_2}{n_1 + n_2} = \frac{1.5 - 1}{1.5 + 1} = 0.2$$

There is no phase shift. The reflectance is again 0.04 or 4 percent. In both cases (a) and (b), the amount of reflected light is the same.

c. Light is traveling in air and is incident on the glass surface at the polarization angle. Here $n_1 = 1$, $n_2 = 1.5$, and $\tan\theta_p = (n_2/n_1) = 1.5$, so $\theta_p = 56.3^\circ$.

If we were to reflect light from a glass plate keeping the angle of incidence at 56.3°, then the reflected light will be polarized with an electric field component perpendicular to the plane of incidence. The transmitted light will have the field greater in the plane of incidence; that is,

it will be partially polarized. By using a stack of glass plates one can increase the polarization of the transmitted light. (This type of *pile-of-plates polarizer* was invented by Dominique F. J. Arago in 1812.)

EXAMPLE 9.10 **ANTIREFLECTION COATINGS ON SOLAR CELLS** When light is incident on the surface of a semiconductor, it becomes partially reflected. Partial reflection is an important consideration in solar cells where transmitted light energy into the semiconductor device is converted to electric energy. The refractive index of Si is about 3.5 at wavelengths around 700–800 nm. Thus the reflectance with $n_1(\text{air}) = 1$ and $n_2(\text{Si}) \approx 3.5$ is

$$R = \left(\frac{n_1 - n_2}{n_1 + n_2}\right)^2 = \left(\frac{1 - 3.5}{1 + 3.5}\right)^2 = 0.309$$

This means that 30 percent of the light is reflected and is not available for conversion to electric energy, a considerable reduction in the efficiency of the solar cell.

However, we can coat the surface of the semiconductor device with a thin layer of a dielectric material such as Si_3N_4 (silicon nitride) that has an intermediate refractive index. Figure 9.15 illustrates how the thin dielectric coating reduces the reflected light intensity. In this case $n_1(\text{air}) = 1$, $n_2(\text{coating}) \approx 1.9$, and $n_3(\text{Si}) = 3.5$. Light is first incident on the air–coating surface, and some of it becomes reflected; this reflected wave is shown as A in Figure 9.15. Wave A has experienced a 180° phase change on reflection as this is an external reflection. The wave that enters and travels in the coating then becomes reflected at the coating–semiconductor surface. This wave, which is shown as B, also suffers a 180° phase change since $n_3 > n_2$. When wave B reaches A, it has suffered a total delay of traversing the thickness d of the coating twice. The phase difference is equivalent to $k_c(2d)$ where $k_c = 2\pi/\lambda_c$ is the wavevector in the coating and is given by $2\pi/\lambda_c$ where λ_c is the wavelength in the coating. Since $\lambda_c = \lambda/n_2$, where λ is the free-space wavelength, the phase difference $\Delta\phi$ between A and B is $(2\pi n_2/\lambda)(2d)$. To reduce the reflected light, A and B must interfere destructively, and this requires the phase difference to be π or odd multiples of π, $m\pi$ where $m = 1,3,5,\ldots$ is an odd integer. Thus

$$\left(\frac{2\pi n_2}{\lambda}\right)2d = m\pi \quad \text{or} \quad d = m\left(\frac{\lambda}{4n_2}\right)$$

Thus, the thickness of the coating must be multiples of the quarter wavelength in the coating and depends on the wavelength.

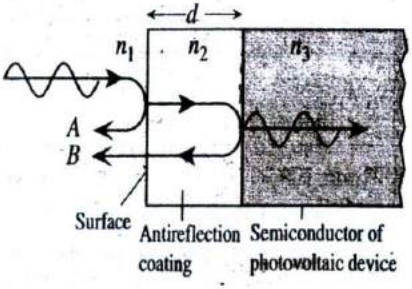


Figure 9.15 Illustration of how an antireflection coating reduces the reflected light intensity.

To obtain a good degree of destructive interference between waves A and B, the two amplitudes must be comparable. It turns out that we need $n_2 = \sqrt{n_1 n_3}$. When $n_2 = \sqrt{n_1 n_3}$, then the reflection coefficient between the air and coating is equal to that between the coating and the semiconductor. In this case we would need $\sqrt{3.5}$ or 1.87. Thus, Si_3N_4 is a good choice as an antireflection coating material on Si solar cells.

Taking the wavelength to be 700 nm, $d = (700\,\text{nm})/[4(1.9)] = 92.1\,\text{nm}$ or odd multiples of d.

EXAMPLE 9.11

DIELECTRIC MIRRORS A **dielectric mirror** consists of a stack of dielectric layers of alternating refractive indices as schematically illustrated in Figure 9.16 where n_1 is smaller than n_2. The thickness of each layer is a quarter wavelength or $\lambda_{layer}/4$, where λ_{layer} is the wavelength of light in that layer, or λ_o/n, where λ_o is the free-space wavelength at which the mirror is required to reflect the incident light and n is the refractive index of the layer. Reflected waves from the interfaces interfere constructively and give rise to a substantial reflected light. If there are a sufficient number of layers, the reflectance can approach unity at the wavelength λ_o. Figure 9.16 also shows schematically a typical reflectance versus wavelength behavior of a dielectric mirror with many layers.

The reflection coefficient r_{12} for light in layer 1 being reflected at the 1–2 boundary is $r_{12} = (n_1 - n_2)/(n_1 + n_2)$ and is a negative number indicating a π phase change. The reflection coefficient for light in layer 2 being reflected at the 2–1 boundary is $r_{21} = (n_2 - n_1)/(n_1 + n_2)$ which is $-r_{12}$ (positive) indicating no phase change. Thus the reflection coefficient alternates in sign through the mirror. Consider two arbitrary waves A and B which are reflected at two consecutive interfaces. The two waves are therefore already out of phase by π due to reflections at the different boundaries. Further, wave B travels an additional distance which is twice $(\lambda_2/4)$ before reaching wave A and therefore experiences a phase change equivalent to $2(\lambda_2/4)$ or $\lambda_2/2$, that is, π. The phase difference between A and B is then $\pi + \pi$ or 2π. Thus waves A and B are in phase and *interfere constructively*. We can similarly show that waves B and C also interfere constructively and so on, so all reflected waves from the consecutive boundaries interfere constructively. After several layers (depending on n_1 and n_2), the transmitted intensity will be very small and the reflected light intensity will be close to unity. Dielectric mirrors are widely used in modern vertical cavity surface emitting semiconductor lasers.

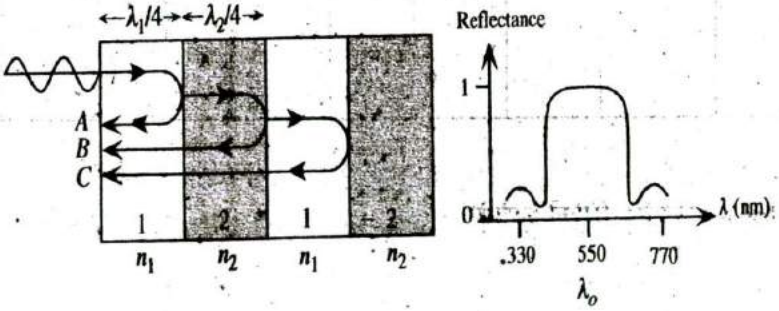


Figure 9.16 Schematic illustration of the principle of the dielectric mirror with many low and high refractive index layers and its reflectance.

9.8 COMPLEX REFRACTIVE INDEX AND LIGHT ABSORPTION

Generally when light propagates through a material, it becomes *attenuated* in the direction of propagation as illustrated in Figure 9.17. We distinguish between *absorption* and *scattering* both of which give rise to a loss of intensity in the regular direction of propagation. In **absorption,** the loss in the power in the propagating EM wave is due to the conversion of light energy to other forms of energy, *e.g.*, lattice vibrations (heat) during the polarization of the molecules of the medium, local vibrations of impurity ions, and excitation of electrons from the valence band to the conduction band. On the other hand, **scattering** is a process by which the energy from a propagating EM wave is redirected as secondary EM waves in various directions away from the original direction of propagation; this is discussed in Section 9.11.

Lossless propagation

It is instructive to consider what happens when a monochromatic light wave such as

$$E = E_o \exp j(\omega t - kz) \qquad [9.52]$$

is propagating in a dielectric medium. The electric field E in Equation 9.52 is either parallel to x or y since propagation is along z. As the wave travels through the medium, the molecules become polarized. This polarization effect is represented by the *relative permittivity* ε_r of the medium. If there were no losses in the polarization process, then the relative permittivity ε_r would be a real number and the corresponding refractive index $n = \sqrt{\varepsilon_r}$ would also be a real number. However, we know that there are always some losses in all polarization processes. For example, when the ions of an ionic crystal are displaced from their equilibrium positions by an alternating electric field and made to oscillate, some of the energy from the electric field is coupled and converted to lattice vibrations (intuitively, "sound" and heat). These losses are generally accounted for by describing the whole medium in terms of a **complex relative permittivity** (or **dielectric constant**) ε_r, that is,

Complex dielectric constant

$$\varepsilon_r = \varepsilon_r' - j\varepsilon_r'' \qquad [9.53]$$

where the real part ε_r' determines the polarization of the medium with losses ignored and the imaginary part ε_r'' describes the losses in the medium. For a lossless medium, obviously $\varepsilon_r = \varepsilon_r'$. The loss ε_r'' depends on the frequency of the wave and usually peaks at certain natural (resonant) frequencies. If the medium has a finite conductivity

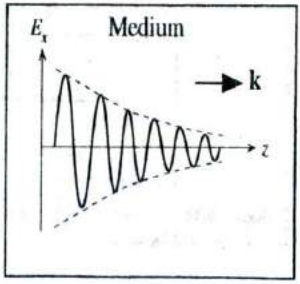


Figure 9.17 Attenuation of light in the direction of propagation.

(*e.g.*, due to a small number of conduction electrons), then there will be a Joule loss due to the electric field in the wave driving these conduction electrons. This type of light attenuation is called **free carrier absorption.** In such cases, ε_r'' and σ are related by

Conduction loss

$$\varepsilon_r'' = \frac{\sigma}{\varepsilon_o \omega} \qquad [9.54]$$

where ε_o is the absolute permittivity and σ is the conductivity at the frequency of the EM wave. Since ε_r is a complex quantity, we should also expect to have a *complex refractive index.*

An EM wave that is traveling in a medium and experiencing attenuation due to absorption can be generally described by a **complex propagation constant** k, that is,

Complex propagation constant

$$k = k' - jk'' \qquad [9.55]$$

where k' and k'' are the real and imaginary parts. If we put Equation 9.55 into Equation 9.52, we will find the following,

Attenuated propagation

$$E = E_o \exp(-k''z) \exp j(\omega t - k'z) \qquad [9.56]$$

The amplitude decays exponentially while the wave propagates along z. The real k' part of the complex propagation constant (wavevector) describes the propagation characteristics, *e.g.*, phase velocity $v = \omega/k'$. The **imaginary** k'' part describes the rate of attenuation along z. The intensity I at any point along z is

$$I \propto |E|^2 \propto \exp(-2k''z)$$

so the rate of change in the intensity with distance is

Imaginary part k''

$$\frac{dI}{dz} = -2k''I \qquad [9.57]$$

where the negative sign represents attenuation.

Suppose that k_o is the propagation constant in a vacuum. This is a real quantity as a plane wave suffers no loss in free space. The **complex refractive index** N with real part n and imaginary part K is defined as the ratio of the complex propagation constant in a medium to propagation constant in free space.

Complex refractive index

$$N = n - jK = \frac{k}{k_o} = \left(\frac{1}{k_o}\right)[k' - jk''] \qquad [9.58a]$$

i.e.,

Refractive index and extinction coefficient

$$n = \frac{k'}{k_o} \quad \text{and} \quad K = \frac{k''}{k_o} \qquad [9.58b]$$

The real part n is simply and generally called the **refractive index** and K is called the **extinction coefficient.** In the absence of attenuation,

$$k'' = 0 \qquad k = k' \qquad \text{and} \qquad N = n = \frac{k}{k_o} = \frac{k'}{k_o}$$

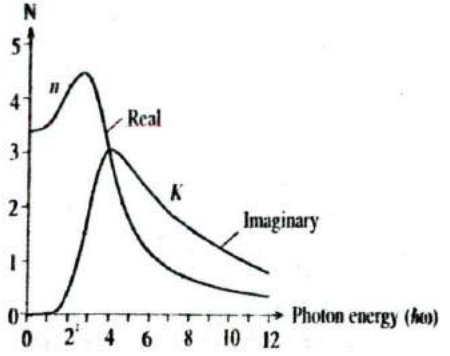


Figure 9.18 Optical properties of an amorphous silicon film in terms of real [n] and imaginary [K] parts of the complex refractive index.

We know that in the absence of loss, the relationship between the refractive index n and the relative permittivity ε_r is $n = \sqrt{\varepsilon_r}$. This relationship is also valid in the presence of loss except that we must use complex refractive index and complex relative permittivity, that is,

Complex refractive index

$$\mathrm{N} = n - jK = \sqrt{\varepsilon_r} = \sqrt{\varepsilon_r' - j\varepsilon_r''} \qquad [9.59]$$

By squaring both sides we can relate n and K directly to ε_r' and ε_r''. The final result is

Complex refractive index

$$n^2 - K^2 = \varepsilon_r' \quad \text{and} \quad 2nK = \varepsilon_r'' \qquad [9.60]$$

Optical properties of materials are typically reported either by showing the frequency dependences of n and K or ε_r' and ε_r''. Clearly we can use Equation 9.60 to obtain one set of properties from the other. Figure 9.18 shows the real (n) and imaginary (K) parts of the complex refractive index of amorphous silicon (noncrystalline form of Si) as a function of photon energy ($h\nu$). For photon energies below the bandgap energy, K is negligible and n is close to 3.5. Both n and K change strongly as the photon energy increases far beyond the bandgap energy.

If we know the frequency dependence of the real part ε_r' of the relative permittivity of a material, we can also determine the frequency dependence of the imaginary part ε_r'', and vice versa. This may seem remarkable, but it is true provided that we know the frequency dependence of either the real or imaginary part over as wide a range of frequencies as possible (ideally from dc to infinity) and the material is *linear, i.e.*, it has a relative permittivity that is independent of the applied field; the polarization response must be linearly proportional to the applied field.[12] The relationships that relate the real and imaginary parts of the relative permittivity are called **Kramers–Kronig relations.** If $\varepsilon_r'(\omega)$ and $\varepsilon_r''(\omega)$ represent the frequency dependences of the real and imaginary parts, respectively, then one can be determined from the other as depicted schematically in Figure 9.19.

The optical properties n and K can be determined by measuring the reflectance from the surface of a material as a function of polarization and the angle of incidence (based on Fresnel's equations).

[12] In addition the material system should be passive—contain no sources of energy.

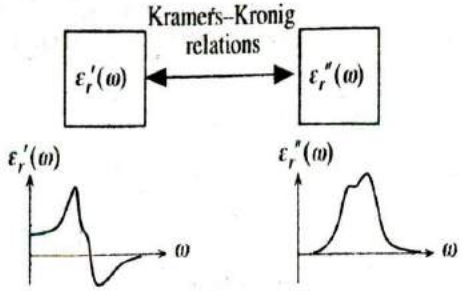


Figure 9.19 Kramers–Kronig relations allow frequency dependences of the real and imaginary parts of the relative permittivity to be related to each other. The material must be a linear system.

It is instructive to mention that the reflection and transmission coefficients that we derived in Section 9.7 were based on using a real refractive index, that is, neglecting losses. We can still use the reflection and transmission coefficients if we simply use the complex refractive index N instead of n. For example, consider a light wave traveling in free space incident on a material at normal incidence ($\theta_i = 90°$). The reflection coefficient is now

$$r = -\frac{\mathrm{N}-1}{\mathrm{N}+1} = -\frac{n-jK-1}{n-jK+1} \qquad [9.61]$$

Reflection coefficient

The reflectance is then

$$R = \left|\frac{n-jK-1}{n-jK+1}\right|^2 = \frac{(n-1)^2+K^2}{(n+1)^2+K^2} \qquad [9.62]$$

Reflectance

which reduce to the usual forms when the extinction coefficient $K = 0$.

EXAMPLE 9.12

COMPLEX REFRACTIVE INDEX Spectroscopic ellipsometry measurements on a silicon crystal at a wavelength of 826.6 nm show that the real and imaginary parts of the complex relative permittivity are 13.488 and 0.038, respectively. Find the complex refractive index, the reflectance and the absorption coefficient α at this wavelength, and the phase velocity.

SOLUTION

We know that $\varepsilon_r' = 13.488$ and $\varepsilon_r'' = 0.038$. Thus, from Equation 9.60, we have

$$n^2 - K^2 = 13.488 \quad \text{and} \quad 2nK = 0.038$$

We can take K from the second equation and substitute for it in the first equation,

$$n^2 - \left(\frac{0.038}{2n}\right)^2 = 13.488$$

This is a quadratic equation in n^2 that can be easily solved on a calculator to find $n = 3.67$. Once we know n, we can find $K = 0.038/2n = 0.00517$. If we simply take the square root of the real part of ε_r, we would still find $n = 3.67$, because the extinction coefficient K is small. The reflectance of the Si crystal is

$$R = \frac{(n-1)^2+K^2}{(n+1)^2+K^2} = \frac{(3.67-1)^2+0.00517^2}{(3.67+1)^2+0.00517^2} = 0.327$$

which is the same as simply using $(n-1)^2/(n+1)^2 = 0.327$, because K is small.

The absorption coefficient α describes the loss in the light intensity I via $I = I_o \exp(-\alpha z)$. By virtue of Equation 9.57,

$$\alpha = 2k'' = 2k_o K = 2\left(\frac{2\pi}{826.6 \times 10^{-9}}\right)(0.00517) = 7.9 \times 10^4 \text{ m}^{-1}$$

Almost all of this absorption is due to band-to-band absorption (photogeneration of electron–hole pairs).

The phase velocity is given by

$$v = \frac{c}{n} = \frac{3 \times 10^8 \text{ m s}^{-1}}{3.67} = 8.17 \times 10^7 \text{ m s}^{-1}$$

EXAMPLE 9.13 **COMPLEX REFRACTIVE INDEX OF InP** An InP crystal has a refractive index (real part) n of 3.549 at a wavelength of 620 nm (photon energy of 2 eV). The reflectance of the air–InP crystal surface at this wavelength is 0.317. Calculate the extinction coefficient K and the absorption coefficient α of InP at this wavelength.

SOLUTION

The reflectance R is given by

$$R = \frac{(n-1)^2 + K^2}{(n+1)^2 + K^2} \quad \text{or} \quad 0.317 = \frac{(3.549-1)^2 + K^2}{(3.549+1)^2 + K^2}$$

which on solving gives $K = 0.302$.

The absorption coefficient is

$$\alpha = 2k_o K = 2\left(\frac{2\pi}{620 \times 10^{-9}}\right)(0.302) = 6.1 \times 10^6 \text{ m}^{-1}$$

EXAMPLE 9.14 **FREE CARRIER ABSORPTION COEFFICIENT AND CONDUCTIVITY** Consider a semiconductor sample with a conductivity σ, and a refractive index n. Show that the absorption coefficient due to free carrier absorption (due to conductivity) is given by

$$\alpha = \left(\frac{1}{c\varepsilon_o}\right)\frac{\sigma}{n}$$

An n-type Ge has a resistivity of about $5 \times 10^{-3}\ \Omega$ m. Calculate the imaginary part ε_r'' of the relative permittivity at a wavelength of 10 μm where the refractive index is 4. Find the attenuation coefficient α due to free carrier absorption.

SOLUTION

Imaginary relative permittivity and conductivity

The relationship between the conductivity and the absorption coefficient is given by

$$\varepsilon_r'' = \frac{\sigma}{\varepsilon_o \omega} \qquad [9.63]$$

The relationship between the imaginary part ε_r'' of the relative permittivity and the extinction coefficient K is

$$2nK = \varepsilon_r''$$

where n is the refractive index (the real part of N). Since the absorption coefficient from Example 9.13 is

$$\alpha = 2k'' = 2k_o K = 2\left(\frac{2\pi}{\lambda}\right)\left(\frac{\varepsilon_r''}{2n}\right)$$

Absorption and imaginary relative permittivity

then

$$\alpha = \left(\frac{\omega}{c}\right)\frac{\varepsilon_r''}{n} \qquad [9.64]$$

where ω is the angular frequency of the EM radiation, $\omega = 2\pi c/\lambda$. Substituting for σ in terms of ε_r'' gives

Absorption and conductivity

$$\alpha = \left(\frac{1}{c\varepsilon_o}\right)\frac{\sigma}{n} \qquad [9.65]$$

The frequency ω is

$$\omega = \frac{2\pi c}{\lambda} = \left[\frac{2\pi(3 \times 10^8\ \text{m s}^{-1})}{10 \times 10^{-6}\ \text{m}}\right] = 1.88 \times 10^{14}\ \text{rad s}^{-1}$$

The relationship between the conductivity and ε_r'' is given by

$$\varepsilon_r'' = \frac{\sigma}{\varepsilon_o \omega} = \left[\frac{(5 \times 10^{-3}\ \Omega\ \text{m})^{-1}}{(8.85 \times 10^{-12}\ \text{F m}^{-1})(1.88 \times 10^{14}\ \text{rad s}^{-1})}\right]$$

i.e.,

$$\varepsilon_r'' = 0.120$$

The absorption coefficient due to free carriers is given by

$$\alpha = \left(\frac{1}{c\varepsilon_o}\right)\frac{\sigma}{n} = \left[\frac{1}{(3 \times 10^8\ \text{m s}^{-1})(8.85 \times 10^{-12}\ \text{F m}^{-1})}\right]\frac{(5 \times 10^{-3}\ \Omega\ \text{m})^{-1}}{4} = 1.9 \times 10^4\ \text{m}^{-1}$$

EXAMPLE 9.15

COMPLEX REFRACTIVE INDEX AND RESONANCE ABSORPTION Equation 9.12 is a simple expression for the electronic polarizability α_e due to an oscillating field. It is based on the *Lorentz model* in which there is a restoring force acting against polarization of the atom or the molecule. ω_o is a *resonant frequency*, or a natural frequency, associated with this type of electronic polarization. The same type of expression will also apply to *ionic polarization*, except that the resonant frequency ω_o will be lower, and the mass m_e has to be changed to an effective mass of the ions.[13] In practice there will be some loss mechanism that absorbs energy from the oscillating field and dissipates it. For example, in ionic polarization, this would involve energy transfer from light to lattice vibrations. In mechanics it is well known that the loss forces (frictional forces) are always proportional to the velocity dx/dt. If we include the energy loss in ac polarization, Equation 9.11 would have an additional term $-\gamma\, dx/dt$ on the right-hand side. If we then follow the same steps to obtain α_e, we would find

Electronic polarizability with loss

$$\alpha_e = \frac{Ze^2}{m_e(\omega_o^2 - \omega^2 + j\gamma\omega)} \qquad [9.66]$$

which is a complex number with real and imaginary parts ($\alpha_e = \alpha_e' - j\alpha_e''$).

[13] Both electronic and ionic polarizabilities have similar expressions. The ionic polarizability in an oscillating field was derived in Chapter 7, and looks almost exactly like Equation 9.66.

Since α_e is a complex quantity, so is ε_r, and hence the refractive index. Consider the simplest relationship between the relative permittivity ε_r and polarizability α_e.

Relative permittivity

$$\varepsilon_r = 1 + \frac{N}{\varepsilon_o}\alpha_e \qquad [9.67]$$

where N is the number of atoms per unit volume (or ion pairs per unit volume for ionic polarization). Thus, the relative permittivity is a *complex quantity*, that is $\varepsilon_r = \varepsilon_r' - j\varepsilon_r''$. We can substitute from Equation 9.66 into 9.67, and also use the fact that when $\omega = 0$, $\varepsilon_r = \varepsilon_{rdc}$, to obtain a simple expression for ε_r.

Complex relative permittivity

$$\varepsilon_r = 1 + \frac{\varepsilon_{rdc} - 1}{1 - \left(\frac{\omega}{\omega_o}\right)^2 + j\frac{\gamma\omega}{\omega_o^2}} \qquad [9.68]$$

The relationship between the complex refractive index N and the complex relative permittivity ε_r is

Complex refractive index

$$\mathrm{N} = n - jK = \varepsilon_r^{1/2} = (\varepsilon_r' - j\varepsilon_r'')^{1/2} \qquad [9.69]$$

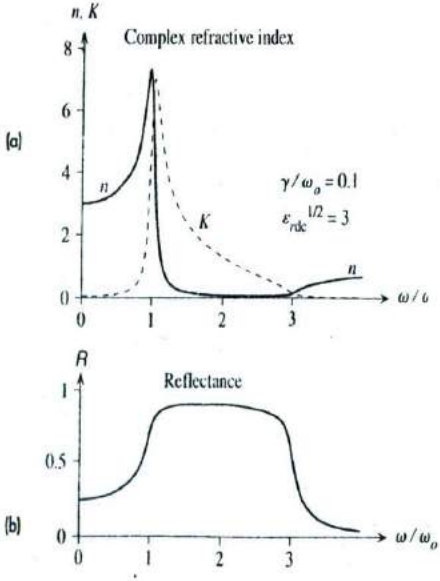
Suppose for simplicity we consider ionic polarization, and we set $\varepsilon_{rdc} = 9$ and $\gamma = 0.1\omega_o$ (reasonable values for ionic polarization). We can calculate ε_r from Equation 9.68 for any choice of ω/ω_o (or for ω by taking $\omega_o = 1$), and then calculate N, that is n and K. (Our calculator or the math program must be able to handle complex numbers.) Figure 9.20a shows the dependence of n and K on the frequency ω/ω_o for the simple Lorentz oscillator model in Equation 9.68. Notice how n and the extinction coefficient K peak close to $\omega = \omega_o$.

The reflectance from Equation 9.62 is plotted in Figure 9.20b as R versus ω/ω_o. It is apparent that R reaches its maximum value at a frequency slightly above $\omega = \omega_o$, and then remains high until ω reaches nearly $3\omega_o$; the *reflectance is substantial while absorption is strong*. It may

Figure 9.20

(a) Refractive index and extinction coefficient versus normalized frequency, ω/ω_o.

(b) Reflectance versus normalized frequency.



seem strange that the crystal is both highly reflecting and highly absorbing. The light that is incident is strongly reflected, and the light that is inside the crystal becomes strongly absorbed. This phenomenon is known as **infrared reflectance,** and occurs over a band of frequencies, called the **Reststrahlen band;** in the present case from ω_o to roughly $3\omega_o$.

9.9 LATTICE ABSORPTION

In optical absorption, some of the energy from the propagating EM wave is converted to other forms of energy, for example, to heat by the generation of lattice vibrations. There are a number of absorption processes that dissipate the energy from the wave. One important mechanism is called **lattice absorption (Reststrahlen absorption)** and involves the vibrations of the lattice atoms as illustrated in Figure 9.21. The crystal in this example consists of ions, and as an EM wave propagates it displaces the oppositely charged ions in opposite directions and forces them to vibrate at the frequency of the wave. In other words, the medium experiences *ionic polarization.* It is the displacements of these ions that give rise to ionic polarization and its contribution to the relative permittivity ε_r. As the ions and hence the lattice is made to vibrate by the passing EM wave, as shown in Figure 9.21, some energy is coupled into the natural lattice vibrations of the solid. This energy peaks when the frequency of the wave is close to the natural lattice vibration frequencies. Typically these frequencies are in the *infrared region.* Most of the energy is then absorbed from the EM wave and converted to lattice vibrational energy (heat). We associate this absorption with the resonance peak or relaxation peak of ionic polarization loss (imaginary part of the relative permittivity ε_r'').

Figure 9.22 shows the infrared resonance absorption peaks in the extinction coefficient K versus wavelength characteristics of GaAs and CdTe; both crystals have substantial ionic bonding. These absorption peaks in Figure 9.22 are usually called **Reststrahlen bands** because absorption occurs over a band of frequencies (even though the band may be narrow), and in some cases may even have identifiable features. Indeed, if we were to plot the reflectance (R) versus wavelength, it would be similar to that shown in Figure 9.20b, and the band would be identified with the high reflectance region.

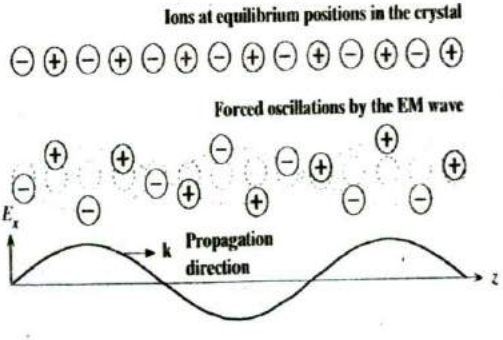


Figure 9.21 Lattice absorption through a crystal. The field in the EM wave oscillates the ions which consequently generate "mechanical" waves in the crystal; energy is thereby transferred from the wave to lattice vibrations.

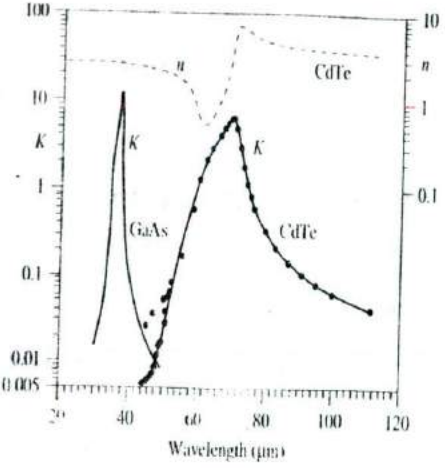


Figure 9.22 Lattice or Reststrahlen absorption in CdTe and GaAs in terms of the extinction coefficient versus wavelength. For reference, n versus λ for CdTe is also shown.

Although Figure 9.21 depicts an ionic solid to visualize absorption due to lattice waves, energy from a passing EM wave can also be absorbed by various ionic impurities in a medium as these charges can couple to the electric field and oscillate. Bonding between an oscillating ion and the neighboring atoms causes the mechanical oscillations of the ion to be coupled to neighboring atoms. This leads to a generation of lattice waves which takes away energy from the EM wave.

RESTSTRAHLEN ABSORPTION Figure 9.22 shows the infrared extinction coefficient K of GaAs and CdTe. Consider CdTe. Calculate the absorption coefficient α and the reflectance R of CdTe at the Reststrahlen peak, and also at 50 μm and at 100 μm. What is your conclusion?

SOLUTION

At the resonant peak, $\lambda \approx 72$ μm, $K \approx 6$, and $n \approx 5$, so the corresponding free-space wavevector is

$$k_o = \frac{2\pi}{\lambda} = \frac{2\pi}{72 \times 10^{-6}\ \text{m}} = 8.7 \times 10^4\ \text{m}^{-1}$$

The absorption coefficient α, by definition, is $2k''$ in Equation 9.57, so

$$\alpha = 2k'' = 2k_o K = 2(8.7 \times 10^4\ \text{m}^{-1})(6) = 1.0 \times 10^6\ \text{m}^{-1}$$

which corresponds to an *absorption depth* $1/\alpha$ of about 1 μm. The reflectance is

$$R = \frac{(n-1)^2 + K^2}{(n+1)^2 + K^2} = \frac{(5-1)^2 + 6^2}{(5+1)^2 + 6^2} = 0.72 \quad \text{or} \quad 72\%$$

Repeating the above calculations at $\lambda = 50$ μm, we get $\alpha = 8.3 \times 10^2\ \text{m}^{-1}$, and $R = 0.11$ or 11 percent. There is a sharp increase in the reflectance from 11 to 72 percent as we approach the resonant peak. At $\lambda = 100$ μm, $\alpha = 6.3 \times 10^3\ \text{m}^{-1}$ and $R = 0.31$ or 31 percent, which is again smaller than the peak reflectance. R is maximum around the Reststrahlen peak.

9.10 BAND-TO-BAND ABSORPTION

The photon absorption process for photogeneration, that is, the creation of electron–hole pairs (EHPs), requires the photon energy to be at least equal to the bandgap energy E_g of the semiconductor material to excite an electron from the valence band (VB) to the conduction band (CB). The **upper cut-off wavelength** (or the threshold wavelength) λ_g for photogenerative absorption is therefore determined by the bandgap energy E_g of the semiconductor, so $h(c/\lambda_g) = E_g$ or

Cut-off wavelength and bandgap

$$\lambda_g(\mu\text{m}) = \frac{1.24}{E_g(\text{eV})} \qquad [9.70]$$

For example, for Si, $E_g = 1.12$ eV and λ_g is 1.11 μm whereas for Ge, $E_g = 0.66$ eV and the corresponding $\lambda_g = 1.87$ μm. It is clear that Si photodiodes cannot be used for optical communications at 1.3 and 1.55 μm, whereas Ge photodiodes are commercially available for use at these wavelengths. Table 9.3 lists some typical bandgap energies and the corresponding cut-off wavelengths of various photodiode semiconductor materials.

Incident photons with wavelengths shorter than λ_g become absorbed as they travel in the semiconductor, and the light intensity, which is proportional to the number of photons, decays exponentially with distance into the semiconductor. The light intensity I at a distance x from the semiconductor surface is given by

Absorption coefficient

$$I(x) = I_o \exp(-\alpha x) \qquad [9.71]$$

where I_o is the intensity of the incident radiation and α is the **absorption coefficient** that depends on the photon energy or wavelength λ. The absorption coefficient α is a material property. Most of the photon absorption (63%) occurs over a distance $1/\alpha$, and $1/\alpha$ is called the **penetration depth** δ. Figure 9.23 shows the α versus λ characteristics of various semiconductors where it is apparent that the behavior of α with the wavelength λ depends on the semiconductor material.

Absorption in semiconductors can be understood in terms of the behavior of the electron energy (E) with the electron momentum ($\hbar k$) in the crystal, called the **crystal**

Table 9.3 Bandgap energy E_g at 300 K, cut-off wavelength λ_g, and type of bandgap [D = direct and I = indirect] for some photodetector materials

Semiconductor	E_g (eV)	λ_g (μm)	Type
InP	1.35	0.91	D
$GaAs_{0.88}Sb_{0.12}$	1.15	1.08	D
Si	1.12	1.11	I
$In_{0.7}Ga_{0.3}As_{0.64}P_{0.36}$	0.89	1.4	D
$In_{0.53}Ga_{0.47}As$	0.75	1.65	D
Ge	0.66	1.87	I
InAs	0.35	3.5	D
InSb	0.18	7	D

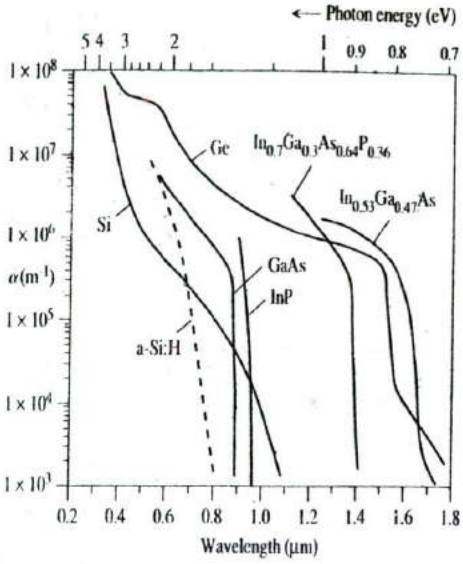


Figure 9.23 Absorption coefficient α versus wavelength λ for various semiconductors.
SOURCE: Data selectively collected and combined from various sources.

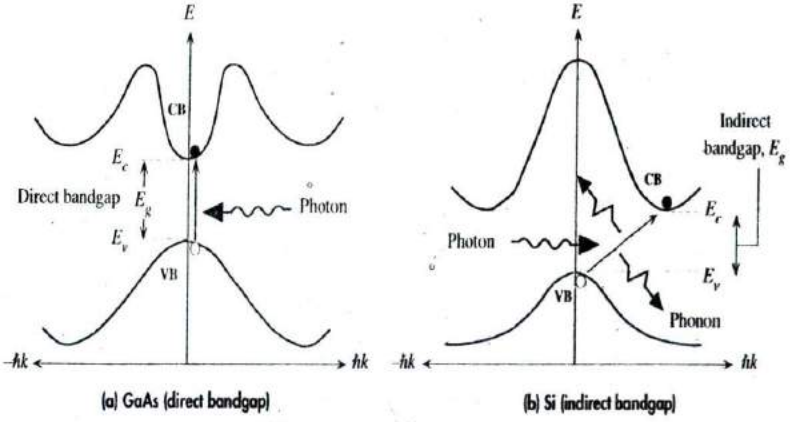


Figure 9.24 Electron energy E versus crystal momentum $\hbar k$ and photon absorption.
(a) Photon absorption in a direct bandgap semiconductor.
(b) Photon absorption in an indirect bandgap semiconductor (VB = valence band; CB = conduction band).

momentum. If k is the wavevector of the electron's wavefunction in the crystal, then the momentum of the electron within the crystal is $\hbar k$. E versus $\hbar k$ behaviors for electrons in the conduction and valence bands of direct and indirect bandgap semiconductors are shown in Figure 9.24a and b, respectively. In **direct bandgap** semiconductors such as III–V semiconductors (*e.g.*, GaAs, InAs, InP, GaP) and in many of their alloys (*e.g.*, InGaAs, GaAsSb) the photon absorption process is a direct process

which requires no assistance from lattice vibrations. The photon is absorbed and the electron is excited directly from the valence band to the conduction band without a change in its k-vector, or its crystal momentum $\hbar k$, inasmuch as the photon momentum is very small. The change in the electron momentum from the valence to the conduction band is

$$\hbar k_{CB} - \hbar k_{VB} = \text{Photon momentum} \approx 0$$

This process corresponds to a vertical transition on the electron energy (E) versus electron momentum ($\hbar k$) diagram as shown in Figure 9.24a. The absorption coefficient of these semiconductors rises sharply with decreasing wavelength from λ_g as apparent for GaAs and InP in Figure 9.23.

In **indirect bandgap** semiconductors such as Si and Ge, the photon absorption for photon energies near E_g requires the absorption and emission of lattice vibrations, that is, **phonons,**[14] during the absorption process as shown in Figure 9.24. If K is the wavevector of a lattice wave (lattice vibrations travel in the crystal), then $\hbar K$ represents the momentum associated with such a lattice vibration; that is, $\hbar K$ is a **phonon momentum.** When an electron in the valence band is excited to the conduction band, there is a change in its momentum in the crystal, and this change in the momentum cannot be supplied by the momentum of the incident photon which is very small. Thus, the momentum difference must be balanced by a phonon momentum,

$$\hbar k_{CB} - \hbar k_{VB} = \text{Phonon momentum} = \hbar K$$

The absorption process is said to be **indirect** as it depends on lattice vibrations which in turn depend on the temperature. Since the interaction of a photon with a valence electron needs a third body, a lattice vibration, the probability of photon absorption is not as high as in a direct transition. Furthermore, the cut-off wavelength is not as sharp as for direct bandgap semiconductors. During the absorption process, a phonon may be absorbed or emitted. If ϑ is the frequency of the lattice vibrations, then the phonon energy is $h\vartheta$. The photon energy is $h\nu$ where ν is the photon frequency. Conservation of energy requires that

$$h\nu = E_g \pm h\vartheta$$

Thus, the onset of absorption does not exactly coincide with E_g, but typically it is very close to E_g inasmuch as $h\vartheta$ is small (< 0.1 eV). The absorption coefficient initially rises slowly with decreasing wavelength from about λ_g as apparent in Figure 9.23 for Si and Ge.

EXAMPLE 9.17

FUNDAMENTAL ABSORPTION A GaAs infrared LED emits at about 860 nm. A Si photodetector is to be used to detect this radiation. What should be the thickness of the Si crystal that absorbs most of this radiation?

---

[14] As much as an electromagnetic radiation is quantized in terms of photons, lattice vibrations in the crystal are quantized in terms of phonons. A phonon is a quantum of lattice vibration. If K is the wavevector of a vibrational wave in a crystal lattice and ω is its angular frequency, then the momentum of the wave is $\hbar K$ and its energy is $\hbar\omega$.

SOLUTION

According to Figure 9.23, at $\lambda \approx 0.8$ μm, Si has $\alpha \approx 6 \times 10^4$ m^{-1}, so the absorption depth

$$\delta = \frac{1}{\alpha} = \frac{1}{6 \times 10^4 \text{ m}^{-1}} = 1.7 \times 10^{-5} \text{ m} \quad \text{or} \quad 17 \text{ μm}$$

If the crystal thickness is δ, then 63 percent of the radiation will be absorbed. If the thickness is 2δ, then the fraction of absorbed radiation, from Equation 9.71, will be

$$\text{Fraction of absorbed radiation} = 1 - \exp[-\alpha(2\delta)] = 0.86 \quad \text{or} \quad 86\%$$

9.11 LIGHT SCATTERING IN MATERIALS

Scattering of an EM wave implies that a portion of the energy in a light beam is directed away from the original direction of propagation as illustrated for a small dielectric particle scattering a light beam in Figure 9.25. There are various types of scattering processes.

Consider what happens when a propagating wave encounters a molecule, or a small dielectric particle (or region), which is smaller than the wavelength. The electric field in the wave polarizes the particle by displacing the lighter electrons with respect to the heavier positive nuclei. The electrons in the molecule couple and oscillate with the electric field in the wave (ac electronic polarization). The oscillation of charge "up" and "down," or the oscillation of the induced dipole, radiates EM waves all around the molecule as depicted in Figure 9.25. We should remember that an oscillating charge is like an alternating current which always radiates EM waves (like an antenna). The net effect is that the incident wave becomes partially reradiated in different directions and hence loses intensity in its original direction of propagation. We may think of the process as the particle absorbing some of the energy via electronic polarization and reradiating it in different directions. It may be thought that the scattered waves constitute a spherical wave emanating from the scattering molecule, but this is not generally the case as the reemitted radiation depends on the shape and polarizability of the molecule in different directions. We assumed a small particle so that at any time the field has no spatial variation through the particle, whose polarization then oscillates with the electric field oscillation. Whenever the size of the scattering region, whether an inhomogeneity or a small

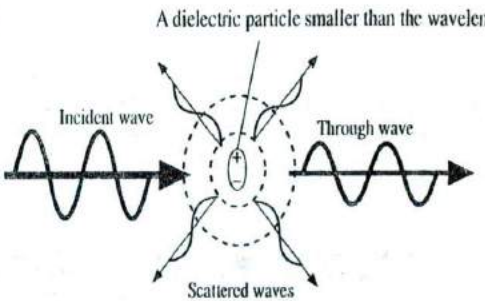


Figure 9.25 Rayleigh scattering involves the polarization of a small dielectric particle or a region that is much smaller than the light wavelength.
The field forces dipole oscillations in the particle (by polarizing it), which leads to the emission of EM waves in "many" directions so that a portion of the light energy is directed away from the incident beam.

particle or a molecule, is much smaller than the wavelength λ of the incident wave, the scattering process is generally termed **Rayleigh scattering.** In this type of scattering, typically the particle size is smaller than one-tenth of the wavelength.

Rayleigh scattering of waves in a medium arises whenever there are small inhomogeneous regions in which the refractive index is different than the medium (which has some average refractive index). This means a local change in the relative permittivity and polarizability. The result is that the small inhomogeneous region acts like a small dielectric particle and scatters the propagating wave in different directions. In the case of optical fibers, dielectric inhomogeneities arise from fluctuations in the relative permittivity that is part of the intrinsic glass structure. As the fiber is drawn by freezing a liquid-like flow, random thermodynamic fluctuations in the composition and structure that occur in the liquid state become frozen into the solid structure. Consequently, the glass fiber has small fluctuations in the relative permittivity which leads to Rayleigh scattering. Nothing can be done to eliminate Rayleigh scattering in glasses as it is part of their intrinsic structure.

It is apparent that the scattering process involves electronic polarization of the molecule or the dielectric particle. We know that this process couples most of the energy at ultraviolet frequencies where the dielectric loss due to electronic polarization is maximum and the loss is due to EM wave radiation. Therefore, as the frequency of light increases, the scattering becomes more severe. In other words, *scattering decreases with increasing wavelength.* For example, blue light which has a shorter wavelength than red light is scattered more strongly by air molecules. When we look at the sun directly, it appears yellow because the blue light has been scattered in the direct light more than the red light. When we look at the sky in any direction but the sun, our eyes receive scattered light which appears blue; hence the sky is blue. At sunrise and sunset, the rays from the sun have to traverse the longest distance through the atmosphere and have the most blue light scattered which gives the sun its red color at these times.

9.12 ATTENUATION IN OPTICAL FIBERS

As light propagates through an optical fiber, it becomes attenuated by a number of processes that depend on the wavelength of light. Figure 9.26 shows the attenuation coefficient, as dB per km, of a typical silica-glass-based optical fiber as a function of wavelength. The sharp increase in the attenuation at wavelengths beyond 1.6 μm in the *infrared* region is due to energy absorption by "lattice vibrations" of the constituent ions of the glass material. Fundamentally, energy absorption in this region corresponds to the stretching of the Si–O bonds in ionic polarization induced by the EM wave. Absorption increases with wavelength as we approach the resonance wavelength of the Si–O bond which is around 9 μm. In the case of Ge–O glasses, this is further away, around 11 μm. There is another intrinsic material absorption in the region below 500 nm, not shown in Figure 9.26, which is due to photons exciting electrons from the valence band to the conduction band of the glass.

There is a marked attenuation peak centered at 1.4 μm, and a barely discernible minor peak at about 1.24 μm. These attenuation regions arise from the presence of hydroxyl ions as impurities in the glass structure inasmuch as it is difficult to remove all

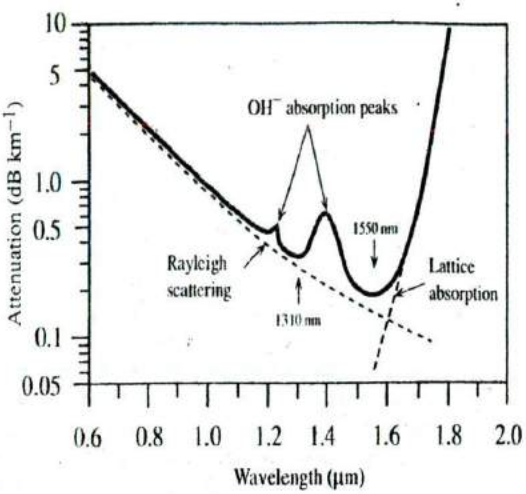


Figure 9.26 Illustration of typical attenuation versus wavelength characteristics of a silica-based optical fiber.
There are two communications channels at 1310 and 1550 nm.

traces of hydroxyl (water) products during fiber production. Further, hydrogen atoms can easily diffuse into the glass structure at high temperatures during production which leads to the formation of hydrogen bonds in the silica structure and OH ions. Energy is absorbed mainly by the stretching vibrations of the OH bonds within the silica structure which has a fundamental resonance in the infrared region (beyond 2.7 μm) but overtones or harmonics at lower wavelengths (or higher frequencies). The first overtone at around 1.4 μm is the most significant as can be seen in Figure 9.26. The second overtone is around 1 μm, and in high-quality fibers this is negligible. A combination of the first overtone of the OH vibration and the fundamental vibrational frequency of SiO_2 gives rise to a minor loss peak at around 1.24 μm. There are two important windows in the attenuation versus wavelength behavior where the attenuation exhibits minima. The window at around 1.3 μm is the region between two neighboring OH^- absorption peaks. This window is widely used in optical communications at 1310 nm. The window at around 1.55 μm is between the first harmonic absorption of OH^- and the infrared lattice absorption tail and represents the lowest attenuation. Current technological drive is to use this window for long-haul communications. It can be seen that it is important to keep the hydroxyl content in the fiber within tolerable levels.

There is a background attenuation process that decreases with wavelength and is due to the Rayleigh scattering of light by the local variations in the refractive index. Glass has a noncrystalline or an amorphous structure which means that there is no long-range order to the arrangement of the atoms but only a short-range order, typically a few bond lengths. The glass structure is as if the structure of the melt has been suddenly frozen. We can only define the number of bonds a given atom in the structure will have. Random variations in the bond angle from atom to atom lead to a disordered structure. There is therefore a random local variation in the density over a few bond lengths which leads to fluctuations in the refractive index over few atomic lengths. These random fluctuations in the refractive index give rise to light scattering and hence

light attenuation along the fiber. It should be apparent that since a degree of structural randomness is an intrinsic property of the glass structure, this scattering process is unavoidable and represents the lowest attenuation possible through a glass medium. As one may surmise, attenuation by scattering in a medium is minimum for light propagating through a "perfect" crystal. In this case the only scattering mechanisms will be due to thermodynamic defects (vacancies) and the random thermal vibrations of the lattice atoms.

As mentioned above, the Rayleigh scattering process decreases with wavelength and, according to Lord Rayleigh, it is inversely proportional to λ^4. The expression for the attenuation α_R in a single component glass due to Rayleigh scattering is approximately given by

Rayleigh scattering in silica

$$\alpha_R \approx \frac{8\pi^3}{3\lambda^4}(n^2 - 1)^2 \beta_T k T_f \qquad [9.72]$$

where λ is the free-space wavelength, n is the refractive index at the wavelength of interest, β_T is the isothermal compressibility (at T_f) of the glass, k is the Boltzmann constant, and T_f is a quantity called the *fictive temperature* (roughly the *softening temperature of glass*) where the liquid structure during the cooling of the fiber is frozen to become the glass structure. Fiber is drawn at high temperatures, and as the fiber cools eventually the temperature drops sufficiently for the atomic motions to be so sluggish that the structure becomes essentially "frozen-in" and remains like this even at room temperature. Thus T_f marks the temperature below which the liquid structure is frozen, and hence the density fluctuations are also frozen into the glass structure. It is apparent that Rayleigh scattering represents the lowest attenuation one can achieve using a glass structure. By proper design, the attenuation window at 1.5 μm may be lowered to approach the Rayleigh scattering limit.

EXAMPLE 9.18

RAYLEIGH SCATTERING LIMIT What is the attenuation due to Rayleigh scattering at around the $\lambda = 1.55$ μm window given that pure silica (SiO_2) has the following properties: $T_f = 1730$ °C (softening temperature), $\beta_T = 7 \times 10^{-11}$ m^2 N^{-1} (at high temperatures), $n = 1.4446$ at 1.5 μm?

SOLUTION

We simply calculate the Rayleigh scattering attenuation using

$$\alpha_R \approx \frac{8\pi^3}{3\lambda^4}(n^2 - 1)^2 \beta_T k T_f$$

so

$$\alpha_R \approx \frac{8\pi^3}{3(1.55 \times 10^{-6})^4}(1.4446^2 - 1)^2(7 \times 10^{-11})(1.38 \times 10^{-23})(1730 + 273)$$

$$= 3.27 \times 10^{-5}\ \text{m}^{-1} \quad \text{or} \quad 3.27 \times 10^{-2}\ \text{km}^{-1}$$

Attenuation in dB per km is then

$$\alpha_{dB} = 4.34\alpha_R = (4.34)(3.27 \times 10^{-2}\ \text{km}^{-1}) = 0.142\ \text{dB km}^{-1}$$

This represents the lowest possible attenuation for a silica glass fiber at 1.55 μm.

9.13 LUMINESCENCE, PHOSPHORS, AND WHITE LEDS

We know from our general experience that certain substances, known as *phosphors*, can absorb light and then reemit light even after the excitation light source has been turned off; this is an example of luminescence. In general, **luminescence** is the emission of light by a material, called a **phosphor,** due to the absorption and conversion of energy into electromagnetic radiation as illustrated in Figure 9.27a and b. The luminescent radiation emitted by the phosphor material is considered to be quite separate from the thermal radiation emitted by virtue of its temperature. Luminescence is light emitted by a nonthermal source when it is excited, in contrast to the emission of radiation from a heated object such as the tungsten filament of a light bulb; the latter is called **incandescence.** Typically the emission of light occurs from certain dopants, impurities, or even defects, called **luminescent** or **luminescence centers,** purposefully introduced into a **host matrix,** which may be a crystal or glass as shown in Figure 9.27c. The luminescent center is also called an **activator.** There are many examples of phosphors. For example, in ruby, the Cr^{3+} ions are the luminescent centers in the sapphire (Al_2O_3) crystal host. Cr^{3+} ions can absorb UV or violet light and then emit red light. This phosphor system is written as Al_2O_3:Cr^{3+}. The excitation and emission involves only the Cr^{3+} ion. In other cases, the activator excitation may also involve the host as discussed later.

Luminescence is normally categorized according to the source of excitation energy. **Photoluminescence** involves excitation by photons (light) as in Figure 9.27a. **X-ray luminescence** involves incident X-rays exciting a phosphor to emit light. **Cathodoluminescence,** as shown in Figure 9.27b, is light emission when the excitation is the bombardment of the phosphor with energetic electrons as in TV cathode ray tubes. **Electroluminescence** is light emission due to the passage of an electric current. Electroluminescence in semiconductive materials appears as a result of an excited electron transiting down to the ground energy level, which would correspond to the recombination of an electron and a hole; the excited electron is the conduction band (CB), and its ground state corresponds to a hole in the valence band (VB). The direct electron–hole recombination mechanism generally occurs very quickly. For example, typical minority carrier lifetimes are in the range of nanoseconds, so light emission from a semiconductor stops within nanoseconds after the removal of the excitation. Such quick luminescence processes occurring over a nanosecond time scale or shorter are normally identified as **fluorescence.** The emission of light from a fluorescent tube

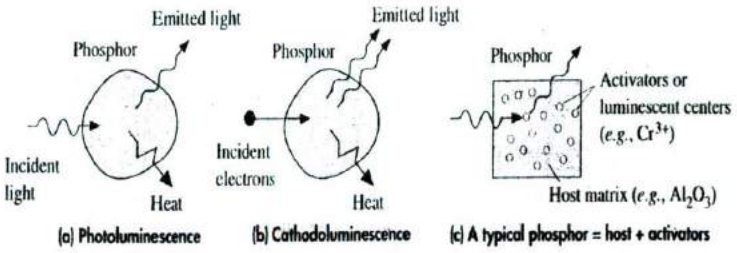


Figure 9.27 Photoluminescence, cathodoluminescence, and a typical phosphor.



This flashlight uses a white LED instead of an incandescent light bulb. The flashlight can operate continuously for 200 hours and can project an intense spot over 30 ft. White LEDs use a phosphor to generate yellow light from the blue light emitted from the LED's semiconductor chip. The mixture of blue and yellow light appears as white.

is actually a fluorescence process. The tube contains a gas mixture of argon and mercury. The Ar and Hg gas atoms become excited by the electrical discharge process and emit light mainly in the ultraviolet region. This UV light is absorbed by the fluorescent coating on the inside of the tube. The excited activators in the phosphor coating then emit radiation in the visible region. A number of phosphors are used to obtain "white" light from the tube.

There are also phosphors from which light emission may continue for milliseconds to hours after the cessation of excitation. These slow luminescence processes are normally referred to as **phosphorescence** (also known as *afterglow*).

Many phosphors are based on activators doped into a host matrix; for example, Eu^{3+} (europium ion) in a Y_2O_3 (yttrium oxide) matrix is a widely used modern phosphor. When excited by UV radiation, it provides an efficient luminescence emission in the red (around 613 nm). It is used as the red-emitting phosphor in color TV tubes and in modern tricolor fluorescent lamps. In very general terms, we can represent the energy of an activator in a host matrix by the highly simplified energy diagram in Figure 9.28.

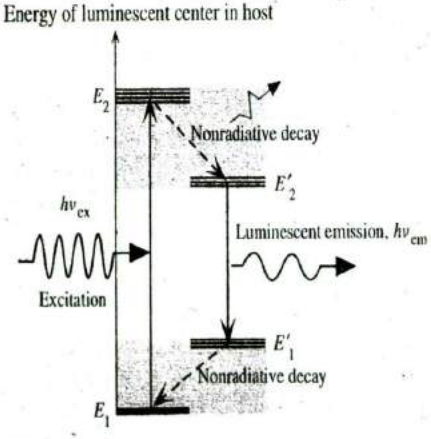


Figure 9.28 Photoluminescence: light absorption, excitation, nonradiative decay and light emission, and return to the ground state E_1.

The energy levels have been displaced horizontally for clarity.

The ground state of the activator is E_1. Upon excitation by an incident radiation of suitable energy $h\nu_{ex}$ the activator becomes excited to E_2. From this energy level, it decays, or *relaxes*, down relatively quickly (on a time scale of the order of picoseconds) to an energy level E_2' by emitting phonons or lattice vibrations. This type of decay is called *radiationless* or *nonradiative decay*. From E_2', the activator decays down to E_1' by emitting a photon (spontaneous emission), which is the emitted luminescent radiation. The emitted photon energy is $h\nu_{em}$, which is less than the excitation photon energy $h\nu_{ex}$. The return from E_1' to the ground state E_1 involves phonon emissions. Further, for some activators, E_1' is either very close to E_1, or it is E_1. The energy levels such as E_2, E_2', E_1', etc., are not well-defined single levels but involve finely spaced multilevels. The higher levels may form multilevel narrow energy "bands." In this example, the activator absorbed the incident radiation and was directly excited, which is known as **activator excitation.** The Cr^{3+} ions in Al_2O_3:Cr^{3+} can be excited directly by blue light and would then emit in the red. There are many phosphors in which the excitation involves the host. In **host excitation,** the host matrix absorbs the incident radiation and transfers the energy to the activator, which then becomes excited to E_2 in Figure 9.28, and so on. In X-ray phosphors, for example, the X-rays are absorbed by the host, which subsequently transfers the energy to the activators. It is apparent from Figure 9.28 that the emitted radiation ($h\nu_{em}$) has a *longer* wavelength than the exciting radiation ($h\nu_{ex}$), that is, $h\nu_{em} < h\nu_{ex}$. The downshift in the light frequency from absorbed to emitted radiation is called the **Stoke's shift.** It should be emphasized that the energy levels of the activator (as shown in Figure 9.28) also depend on the host, because the internal electric fields within the host crystal act on the activator and shift these levels up and down. The emission characteristics depend firstly on the activator, and secondly on the host.

There are a number of host excitation mechanisms. In one possible process, which involves a semiconductor host, as depicted in Figure 9.29, an incident photon initially excites a valence band (VB) electron to the conduction band (CB). The electron then thermalizes, *i.e.*, loses the excess energy as it collides with lattice vibrations, and falls close to E_c, and wanders around in the crystal. In one process, *a* in Figure 9.29, the electron can be captured into an excited state D of a luminescent center or an activator. The electron then falls down in energy to the ground state A of the activator releasing a photon, which is the luminescent emission. The electron at the ground state then recombines with a hole in the VB. Thus the activator acts as a **radiative recombination center.** In some cases D and A may be separate centers representing *donor* and *acceptor*-like centers, hence the labels D and A. In other cases, the radiative recombination center may simply be a single energy level in the bandgap, which is shown as R in Figure 9.29. The electron can emit a photon as it is captured into R, shown as process *b* in Figure 9.29, or emit the photon after it is captured by R, as it recombines with a hole, shown as process *c* in Figure 9.29. Processes *a* and *b* occur in various ZnS-based phosphors. For example, in ZnS:Cu^+ phosphors, the activator is Cu^+, which has an energy level at A in Figure 9.29. The luminescent emission is enhanced by using a coactivator, such as Al in ZnS:Cu^+. Al acts as a shallow donor D, and the luminescence is due to process *a* in Figure 9.29.

There may also be traps in the semiconductor because of various crystal defects, or there may be added impurities. The electron can become captured by a trap at a localized energy level E_t in the bandgap, but close to E_c. These electron traps temporarily

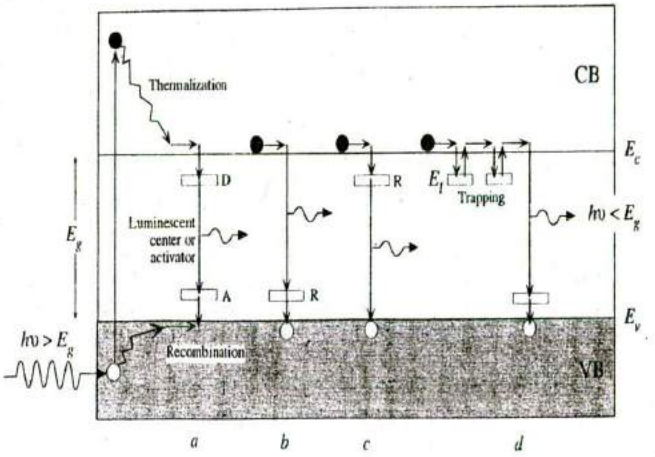


Figure 9.29 Optical absorption generates an EHP.
Both carriers thermalize. There are a number of recombination processes via a dopant that can result in a luminescent emission.

capture an electron from the conduction band and thereby immobilize it. The time the electron spends trapped at E_t depends on the energy depth of the trap from the conduction band, $E_c - E_t$. After a while a strong lattice vibration returns the electron back into the conduction band (by thermal excitation). The time interval between photogeneration and recombination can be relatively long if the electron remains captured at E_t for a considerable length of time. In fact, the electron may become trapped and detrapped many times before it finally recombines, so the emission of light can persist for a relatively long time after the cessation of excitation (*e.g.*, milliseconds or longer) as indicated by process *d* in Figure 9.29.

It is also possible to excite electrons into the CB by bombarding the material with a high-energy electron beam, which leads to cathodoluminescence. Color CRT displays are typically coated uniformly with three sets of phosphor dots which exhibit cathodoluminescence in the blue, red, and green wavelengths. In electroluminescence, an electric current, either ac or dc, is used to inject electrons into the CB which then recombine with holes and emit light. For example, passing a current through certain semiconducting phosphors such as ZnS doped with Mn causes light emission by electroluminescence. The emission of light from a light emitting diode (LED) is an example of **injection electroluminescence** in which the applied voltage causes charge carrier injection and recombination in a device (diode) that has a junction between a *p*-type and an *n*-type semiconductor.

Zinc sulfide with various activators has been one of the traditional phosphors. The $ZnS:Ag^+$ in which Ag^+ is the activator, is still used as a blue emitting phosphor, though in some cases Cd is substituted for some of the Zn. $ZnS:Cu^+$ emits in the green, which is also a useful phosphor. Most modern phosphors, on the other hand, have been based on using rare earth activators in various hosts. For example, $Y_2O_3:Eu^{3+}$ absorbs UV

Table 9.4 Selected phosphor examples

Phosphor	Activator	Useful Emission	Example Excitation	Comment or Application
$Y_2O_3:Eu^{3+}$	Eu^{3+}	Red	UV	Fluorescent lamp, color TV
$BaMgAl_{10}O_{17}:Eu^{2+}$	Eu^{2+}	Blue	UV	Fluorescent lamp
$CeMgAl_{11}O_{19}:Tb^{3+}$	Tb^{3+}	Green	UV	Fluorescent lamp
$Y_3Al_5O_{12}:Ce^{3+}$	Ce^{3+}	Yellow	Blue, violet	White LED
$Sr_2SiO_4:Eu^{3+}$	Eu^{3+}	Yellow	Violet	White LED (experimental)
$ZnS:Ag^+$	Ag^+	Blue	Electron beam	Color TV blue phosphor
$Zn_{0.68}Cd_{0.32}S:Ag^+$	Ag^+	Green	Electron beam	Color TV green phosphor
$ZnS:Cu^+$	Cu^+	Green	Electron beam	Color TV green phosphor

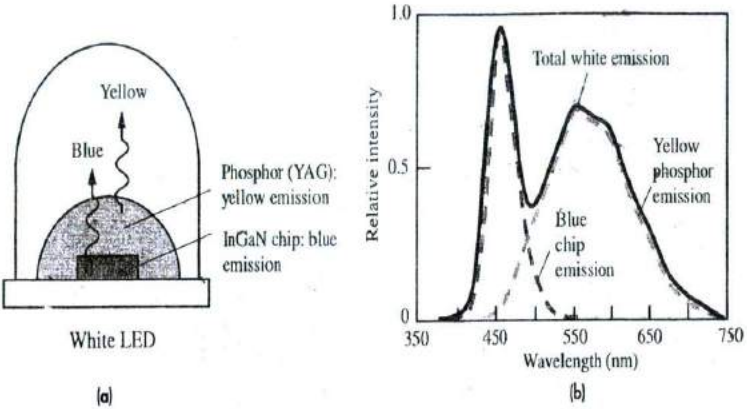


Figure 9.30

(a) A typical "white" LED structure.

(b) The spectral distribution of light emitted by a white LED. Blue luminescence is emitted by the GaInN chip and "yellow" phosphorescence or luminescence is produced by a phosphor. The combined spectrum looks "white."

radiation and emits in the red. $Y_3Al_5O_{12}:Ce^{3+}$ absorbs blue light and emits yellow light. Some of the most popular activators are Eu^{3+} for red, Eu^{2+} for blue, and Tb^{3+} for green. Table 9.4 summarizes a number of phosphors commonly used in various applications.

Recent inexpensive white LEDs that have appeared on the market seem to emit white light by emitting a mixture of blue and yellow light which are registered visually by the eye as appearing white. (Yellow consists of red and green mixed together, so mixing blue and yellow generates "white.") The production of white LEDs became possible due to development of bright blue-emitting LEDs based on gallium-indium-nitride (GaInN). The white LED uses a semiconductor chip emitting at a short wavelength (blue, violet, or ultraviolet) and a *phosphor* to convert some of the blue light to yellow light as depicted in Figure 9.30a. The phosphor absorbs light from the diode

and undergoes luminescent emission at a longer wavelength. Obviously, the quality and spectral characteristics of the combined emission vary with different designs; Figure 9.30b shows example spectra involved in the blue and yellow emissions and the overall "white" emission from a white LED. Typical phosphors have been based on yttrium-aluminum- ($Y_3Al_5O_{12}$) garnets (YAGs) as the host material. This host is doped with one of the rare earth elements for the activator. Cerium is a common dopant element in YAG phosphors; that is, the phosphor is $Y_3Al_5O_{12}:Ce^{3+}$, which is able to efficiently absorb the blue and emit the yellow. White LEDs are soon expected to challenge the existing incandescent sources for general lighting.

9.14 POLARIZATION

A propagating EM wave has its electric and magnetic fields at right angles to the direction of propagation. If we place a z axis along the direction of propagation, then the electric field can be in any direction in the plane perpendicular to the z axis. The term **polarization** of an EM wave describes the behavior of the electric field vector in the EM wave as it propagates through a medium. If the oscillations of the electric field at all times are contained within a well-defined line, then the EM wave is said to be **linearly polarized** as shown in Figure 9.31a. The field vibrations and the direction of propagation (z) define a plane of polarization (plane of vibration), so linear polarization implies a wave that is **plane-polarized.** By contrast, if a beam of light has waves with the E field in each in a random direction but perpendicular to z, then this light beam is *unpolarized*. A light beam can be linearly polarized by passing the beam

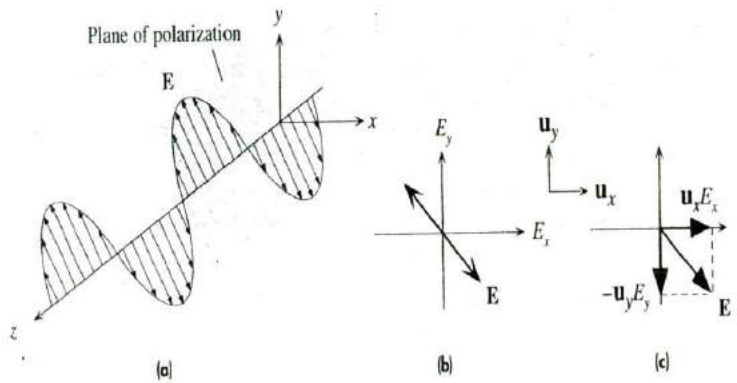


Figure 9.31

(a) A linearly polarized wave has its electric field oscillations defined along a line perpendicular to the direction of propagation z. The field vector **E** and z define a *plane of polarization*.

(b) The E-field oscillations are contained in the plane of polarization.

(c) A linearly polarized light at any instant can be represented by the superposition of two fields E_x and E_y with the right magnitude and phase.

through a *polarizer,* such as a polaroid sheet, a device that only passes electric field oscillations lying on a well-defined plane parallel to its transmission axis.

Suppose that we arbitrarily place the x and y axes and describe the electric field in terms of its components E_x and E_y along x and y (we are justified to do this because E_x and E_y are perpendicular to z). To find the electric field in the wave at any space and time location, we add E_x and E_y *vectorially.* Both E_x and E_y can individually be described by a wave equation which must have the same angular frequency ω and wavenumber k. However, we must include a phase difference ϕ between the two:

$$E_x = E_{xo}\cos(\omega t - kz) \qquad [9.73]$$

and

$$E_y = E_{yo}\cos(\omega t - kz + \phi) \qquad [9.74]$$

where ϕ is the phase difference between E_y and E_x; ϕ can arise if one of the components is delayed (retarded).

The linearly polarized wave in Figure 9.31a has the **E** oscillations at $-45°$ to the x axis as shown in Figure 9.31b. We can generate this field by choosing $E_{xo} = E_{yo}$ and $\phi = \pm 180° (\pm\pi)$ in Equations 9.73 and 9.74. Put differently, E_x and E_y have the same magnitude, but they are out of phase by 180°. If $\mathbf{u}_x$ and $\mathbf{u}_y$ are the unit vectors along x and y, using $\phi = \pi$ in Equation 9.74, the field in the wave is

$$\mathbf{E} = \mathbf{u}_x E_x + \mathbf{u}_y E_y = \mathbf{u}_x E_{xo}\cos(\omega t - kz) - \mathbf{u}_y E_{yo}\cos(\omega t - kz)$$

or

$$\mathbf{E} = \mathbf{E}_o \cos(\omega t - kz) \qquad [9.75]$$

where

$$\mathbf{E}_o = \mathbf{u}_x E_{xo} - \mathbf{u}_y E_{yo} \qquad [9.76]$$

Equations 9.75 and 9.76 state that the vector $\mathbf{E}_o$ is at $-45°$ to the x axis and propagates along the z direction.

There are many choices for the behavior of the electric field besides the simple linear polarization in Figure 9.31. For example, if the magnitude of the field vector **E** remains constant but its tip at a given location on z traces out a circle by rotating in a clockwise sense with time, as observed by the receiver of the wave, then the wave is said to be **right circularly polarized**[15] as in Figure 9.32. If the rotation of the tip of **E** is counterclockwise, the wave is said to be **left circularly polarized.** From Equations 9.73 and 9.74, it should be apparent that a right circularly polarized wave has $E_{xo} = E_{yo} = A$ (an amplitude) and $\phi = \pi/2$. This means that,

$$E_x = A\cos(\omega t - kz) \qquad [9.77]$$

and

$$E_y = -A\sin(\omega t - kz) \qquad [9.78]$$

[15] There is a difference in this definition in optics and engineering. The definition here follows that in optics which is more prevalent in optoelectronics.

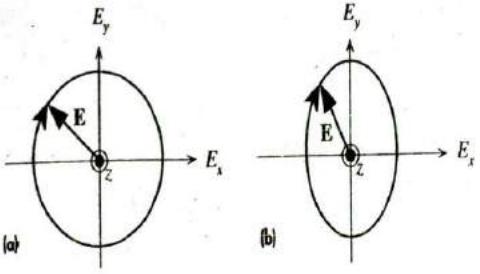


Figure 9.32
(a) A *right circularly polarized light* that is traveling along z (out of paper). The field vector **E** is always at right angles to z, rotates clockwise around z with time, and traces out a full circle over one wavelength of distance propagated.
(b) An elliptically polarized light.

It is relatively straightforward to show that Equations 9.77 and 9.78 represent a circle that is

$$E_x^2 + E_y^2 = A^2 \qquad [9.79]$$

as shown in Figure 9.32.

When the phase difference ϕ is other than 0, $\pm\pi$, or $\pm\pi/2$, the resultant wave is **elliptically polarized** and the tip of the vector in Figure 9.32 traces out an *ellipse*.

9.15 OPTICAL ANISOTROPY

An important characteristic of crystals is that many of their properties depend on the crystal direction; that is, crystals are generally anisotropic. The dielectric constant ε_r depends on electronic polarization which involves the displacement of electrons with respect to positive atomic nuclei. Electronic polarization depends on the crystal direction inasmuch as it is easier to displace electrons along certain crystal directions. *This means that the refractive index n of a crystal depends on the direction of the electric field in the propagating light beam.* Consequently, the velocity of light in a crystal depends on the direction of propagation and on the state of its polarization, *i.e.*, the direction of the electric field. Most noncrystalline materials, such as glasses and liquids, and all cubic crystals are **optically isotropic,** that is, the refractive index is the same in all directions. For all classes of crystals excluding cubic structures, the refractive index depends on the propagation direction and the state of polarization. The result of optical anisotropy is that, except along certain special directions, any unpolarized light ray entering such a crystal breaks into two different rays with different polarizations and phase velocities. When we view an image through a calcite crystal, an optically anisotropic crystal, we see two images, each constituted by light of different polarization passing through the crystal, whereas there is only one image through an optically isotropic crystal as depicted in Figure 9.33. Optically anisotropic crystals are called **birefringent** because an incident light beam may be doubly refracted.

Experiments and theories on "most anisotropic crystals," *i.e.*, those with the highest degree of anisotropy, show that we can describe light propagation in terms of *three* refractive indices, called **principal refractive indices** n_1, n_2, and n_3, along three mutually orthogonal directions in the crystal, say x, y, and z, called **principal axes.** These

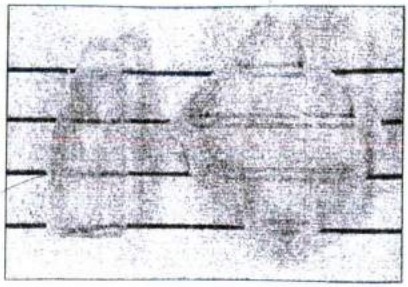


Figure 9.33 A line viewed through a cubic sodium chloride (halite) crystal (optically isotropic) and a calcite crystal (optically anisotropic).

indices correspond to the polarization state of the EM wave along these axes. In addition, anisotropic crystals may possess one or two optic axes. An **optic axis** is a special direction in the crystal along which the velocity of propagation does *not* depend on the state of polarization. The propagation velocity along the optic axis is the same whatever the polarization of the EM wave.

Crystals that have three distinct principal indices also have *two* optic axes and are called **biaxial crystals.** On the other hand, **uniaxial crystals** have two of their principal indices the same ($n_1 = n_2$) and have only *one* optic axis. Table 9.5 summarizes crystal classifications according to optical anisotropy. Uniaxial crystals, such as quartz, that have $n_3 > n_1$, are called **positive,** and those such as calcite that have $n_3 < n_1$ are called **negative** uniaxial crystals.

Table 9.5 Principal refractive indices of some optically isotropic and anisotropic crystals (near 589 nm, yellow Na-D line)

Optically Isotropic	$n = n_o$		
Glass (crown)	1.510		
Diamond	2.417		
Fluorite (CaF_2)	1.434		
Uniaxial—Positive	n_o	n_e	
Ice	1.309	1.3105	
Quartz	1.5442	1.5533	
Rutile (TiO_2)	2.616	2.903	
Uniaxial—Negative	n_o	n_e	
Calcite ($CaCO_3$)	1.658	1.486	
Tourmaline	1.669	1.638	
Lithium niobate ($LiNBO_3$)	2.29	2.20	
Biaxial	n_1	n_2	n_3
Mica (muscovite)	1.5601	1.5936	1.5977

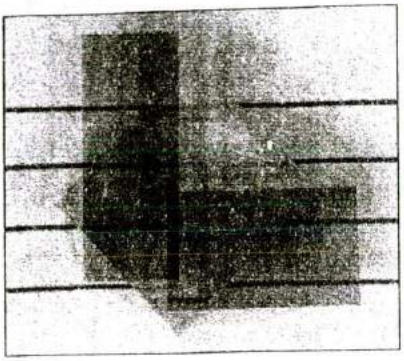


Figure 9.34 Two polaroid analyzers are placed with their transmission axes, along the long edges, at right angles to each other.
The ordinary ray, undeflected, goes through the left polarizer, whereas the extraordinary wave, deflected, goes through the right polarizer. The two waves therefore have orthogonal polarizations.

9.15.1 UNIAXIAL CRYSTALS AND FRESNEL'S OPTICAL INDICATRIX

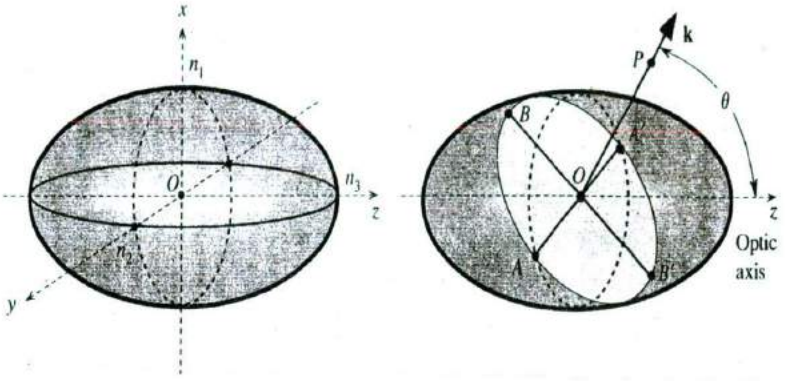
For our discussions of optical anisotropy, we will consider uniaxial crystals such as calcite and quartz. All experiments and theories lead to the following basic principles.[16]

Any EM wave entering an anisotropic crystal splits into two orthogonal linearly polarized waves that travel with different phase velocities; that is, they experience different refractive indices. These two orthogonally polarized waves in uniaxial crystals are called **ordinary** (o) and **extraordinary** (e) waves. The o-wave has the same phase velocity in all directions and behaves like an ordinary wave in which the field is perpendicular to the phase propagation direction. The e-wave has a phase velocity that depends on its direction of propagation and its state of polarization, and further the electric field in the e-wave is not necessarily perpendicular to the phase propagation direction. These two waves propagate with the same velocity only along a special direction called the **optic axis.** The o-wave is always perpendicularly polarized to the optic axis and obeys the usual Snell's law.

The two images observed through the calcite crystal in Figure 9.33 are due to o-waves and e-waves being refracted differently, so when they emerge from the crystal they have been separated. Each ray constitutes an image, but the field directions are **orthogonal.** The fact that this is so is easily demonstrated by using two polaroid analyzers with their transmission axes at right angles as in Figure 9.34. If we were to view an object along the optic axis of the crystal, we would not see two images because the two rays would experience the same refractive index.

As mentioned, we can represent the optical properties of a crystal in terms of three refractive indices along three orthogonal axes, the principal axes of the crystal, shown as x, y, and z in Figure 9.35a. These are special axes along which the polarization vector and the electric field are parallel. (Put differently, the electric displacement[17] $\mathbf{D}$ and the electric field $\mathbf{E}$ vectors are parallel.) The refractive indices along these x, y, and z axes are the principal indices n_1, n_2, and n_3, respectively, for electric

[16] These statements can be proved by solving Maxwell's equations in an anisotropic medium.
[17] Electric displacement $\mathbf{D}$ at any point is defined by $\mathbf{D} = \varepsilon_o\mathbf{E} + \mathbf{P}$ where $\mathbf{E}$ is the electric field and $\mathbf{P}$ is the polarization at that point.



(a) Fresnel's ellipsoid (for $n_1 = n_2 < n_3$; quartz)

(b) An EM wave propagating along *OP* at an angle θ to optic axis.

Figure 9.35

field oscillations along these directions (not to be confused with the wave propagation direction). For example, for a wave with a polarization parallel to the x axes, the refractive index is n_1.

The refractive index associated with a particular EM wave in a crystal can be determined by using Fresnel's *refractive index ellipsoid,* called the **optical indicatrix,**[18] which is a refractive index surface placed in the center of the principal axes, as shown in Figure 9.35a, where the x, y, and z axis intercepts are n_1, n_2, and n_3. If all three indices were the same, $n_1 = n_2 = n_3 = n_o$, we would have a spherical surface and all electric field polarization directions would experience the same refractive index n_o. Such a spherical surface would represent an optically isotropic crystal. For positive uniaxial crystals such as quartz, $n_1 = n_2 < n_3$, which is the ellipsoid example shown in Figure 9.35a.

Suppose that we wish to find the refractive indices experienced by a wave traveling with an arbitrary wavevector **k**, which represents the direction of phase propagation. This phase propagation direction is shown as *OP* in Figure 9.35b and is at an angle θ to the z axis. We place a plane perpendicular to *OP* and passing through the center *O* of the indicatrix. This plane intersects the ellipsoid surface in a curve *ABA'B'* which is an *ellipse*. The major (*BOB'*) and minor (*AOA'*) axes of this ellipse determine the field oscillation directions and the refractive indices associated with this wave. Put differently, the original wave is now represented by two orthogonally polarized EM waves.

The line *AOA'*, the *minor axis,* corresponds to the polarization of the ordinary wave, and its semiaxis *AA'* is the refractive index $n_o = n_2$ of this *o*-wave. The electric displacement and the electric field are in the same direction and parallel to *AOA'*. If

[18] There are various names in the literature with various subtle nuances: the Fresnel ellipsoid, optical indicatrix, index ellipsoid, reciprocal ellipsoid, Poinsot ellipsoid, ellipsoid of wave normals.

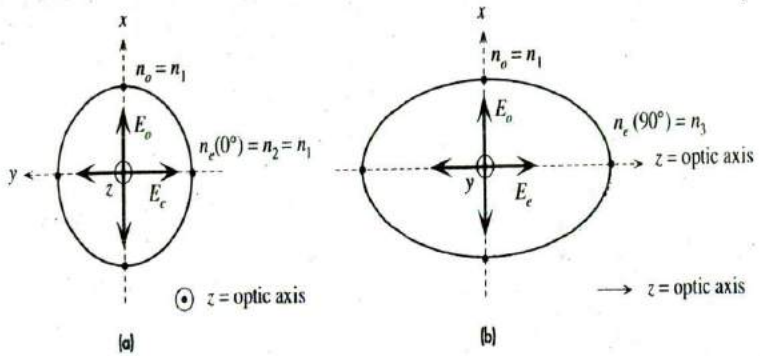


Figure 9.36 $E_o = E_{o\text{-wave}}$ and $E_e = E_{e\text{-wave}}$.
(a) Wave propagation along the optic axis.
(b) Wave propagation normal to the optic axis.

we were to change the direction of *OP*, we would always find the same minor axis, *i.e.*, n_o is either n_1 or n_2 whatever the orientation of *OP* (try orientating *OP* to be along *y* and along *x*). This means that the *o*-wave always experiences the same refractive index in all directions. (The *o*-wave behaves just like an ordinary wave, hence the name.)

The line *BOB'* in Figure 9.35b, the *major axis*, corresponds to the electric displacement field (**D**) oscillations in the extraordinary wave, and its semiaxis *OB* is the refractive index $n_e(\theta)$ of this *e*-wave. This refractive index is smaller than n_3 but greater than n_2 ($= n_o$). The *e*-wave therefore travels more slowly than the *o*-wave in this particular direction and in this crystal. If we change the direction of *OP*, we find that the length of the major axis changes with the *OP* direction. Thus, $n_e(\theta)$ depends on the wave direction θ. As apparent, $n_e = n_o$ when *OP* is along the *z* axis, that is, when the wave is traveling along *z* as in Figure 9.36a. This direction is the *optic axis*, and all waves traveling along the optic axis have the same phase velocity whatever their polarization. When the *e*-wave is traveling along the *y* axis, or along the *x* axis, $n_e(\theta) = n_3 = n_e$ and the *e*-wave has its slowest phase velocity as shown in Figure 9.36b. Along any *OP* direction that is at an angle θ to the optic axis, the *e*-wave has a refractive index $n_e(\theta)$ given by

$$\frac{1}{n_e(\theta)^2} = \frac{\cos^2\theta}{n_o^2} + \frac{\sin^2\theta}{n_e^2} \qquad [9.80]$$

Refractive index of the e-wave

Clearly, for $\theta = 0°$, $n_e(0°) = n_o$ and for $\theta = 90°$, $n_e(90°) = n_e$.

The major axis *BOB'* in Figure 9.35b determines the *e*-wave polarization by defining the direction of the displacement vector **D** and not **E**. Although **D** is perpendicular to **k**, this is not true for **E**. The electric field $\mathbf{E}_{e\text{-wave}}$ of the *e*-wave is orthogonal to that of the *o*-wave, and it is in the plane determined by **k** and the optic axis. $\mathbf{E}_{e\text{-wave}}$ is orthogonal to **k** only when the *e*-wave propagates along one of the principal axes. In birefringent crystals it is usual to take the *ray direction* as the direction of energy flow,

that is the direction of the Poynting vector (**S**). The $\mathbf{E}_{e\text{-wave}}$ is then orthogonal to the ray direction. For the *o*-wave, the wavefront propagation direction **k** is the same as the energy flow direction **S**. For the *e*-wave, however, the wavefront propagation direction **k** is not the same as the energy flow direction **S**.

9.15.2 BIREFRINGENCE OF CALCITE

Consider a calcite crystal ($CaCO_3$) which is a negative uniaxial crystal and also well known for its double refraction. When the surfaces of a calcite crystal have been cleaved, that is, cut along certain crystal planes, the crystal attains a shape that is called a *cleaved form* and the crystal faces are rhombohedrons (parallelogram with 78.08° and 101.92°). A cleaved form of the crystal is called a *calcite rhomb*. A plane of the calcite rhomb that contains the optical axis and is normal to a pair of opposite crystal surfaces is called a *principal section*.

Consider what happens when an unpolarized or natural light enters a calcite crystal at *normal* incidence and thus also normal to a principal section to this surface, but at an angle to the optic axis as shown in Figure 9.37. The ray breaks into ordinary (*o*) and extraordinary (*e*) waves with mutually orthogonal polarizations. The waves propagate in the plane of the principal section as this plane also contains the incident light. The *o*-wave has its field oscillations perpendicular to the optic axis. It obeys Snell's law which means that it enters the crystal undeflected. Thus the direction of *E*-field oscillations must come out of the paper so that it is normal to the optic axis and also to the direction of propagation. The field $E_\perp$ in the *o*-ray is shown as dots, oscillating into and out of the paper.

The *e*-wave has a polarization orthogonal to the *o*-wave and in the principal section. The *e*-wave polarization is in the plane of the paper, indicated as $E_\parallel$, in Figure 9.37. It travels with a different velocity and diverges from the *o*-wave. Clearly, the

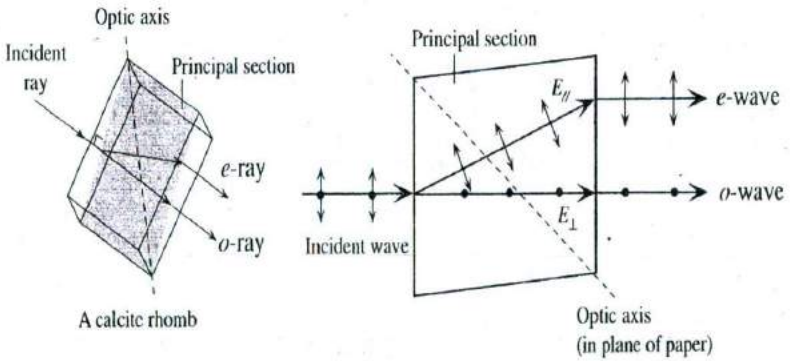


Figure 9.37 An EM wave that is off the optic axis of a calcite crystal splits into two waves called ordinary and extraordinary waves.

These waves have orthogonal polarizations and travel with different velocities. The o-wave has a polarization that is always perpendicular to the optical axis.

e-wave does not obey the usual Snell's law inasmuch as the angle of refraction is not zero. We can determine the e-ray direction by noting that the e-wave propagates sideways as in Figure 9.37b at right angles to $E_{\parallel}$.

9.15.3 DICHROISM

In addition to the variation in the refractive index, some anisotropic crystals also exhibit **dichroism,** a phenomenon in which the optical absorption in a substance depends on the direction of propagation and the state of polarization of the light beam. A dichroic crystal is an optically anisotropic crystal in which either the e-wave or the o-wave is heavily attenuated (absorbed). This means that a light wave of arbitrary polarization entering a dichroic crystal emerges with a well-defined polarization because the other orthogonal polarization would have been attenuated. Generally dichroism depends on the wavelength of light. For example, in a tourmaline (aluminum borosilicate) crystal, the o-wave is much more heavily absorbed with respect to the e-wave.

9.16 BIREFRINGENT RETARDING PLATES

Consider a positive uniaxial crystal such as a quartz ($n_e > n_o$) plate that has the optic axis (taken along z) parallel to the plate faces as in Figure 9.38. Suppose that a *linearly polarized* wave is normally incident on a plate face. If the field **E** is parallel to the optic axis (shown as $E_{\parallel}$), then this wave will travel through the crystal as an e-wave with a velocity c/n_e slower than the o-wave since $n_e > n_o$. Thus, the optic axis is the "slow axis" for waves polarized parallel to it. If **E** is at right angles to the optic axis (shown as $E_{\perp}$), then this wave will travel with a velocity c/n_o, which will be the fastest velocity in the crystal. Thus the axis perpendicular to the optic axis (say x) will be the "fast axis" for polarization along this direction. When a light ray enters a crystal at normal incidence to the optic axis and plate surface, then the o- and e-waves travel along the same direction as shown in Figure 9.38. We can of course resolve a linear polarization at an angle α to z into $E_{\perp}$ and $E_{\parallel}$. The o-wave corresponds to the propagation of $E_{\perp}$ and the e-wave to the propagation of $E_{\parallel}$ in the crystal. When the light comes out at the

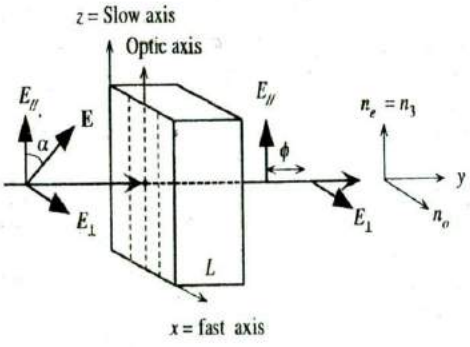


Figure 9.38 A retarder plate.
The optic axis is parallel to the plate face. The o- and e-waves travel in the same direction but at different speeds.

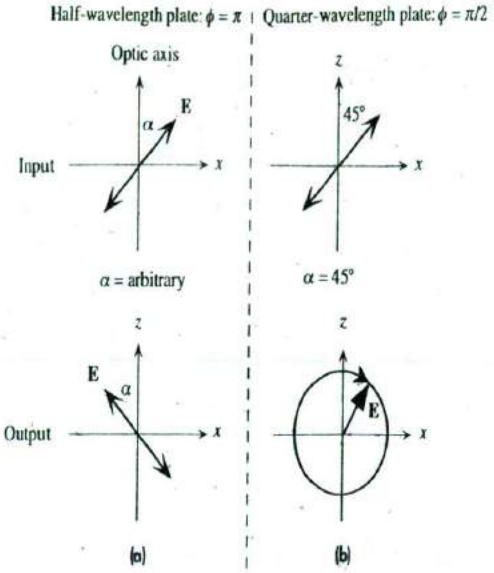


Figure 9.39 Input and output polarizations of light through (a) a half-wavelength plate and (b) through a quarter-wavelength plate.

opposite face, these two components $E_\perp$ and $E_\parallel$ would have been phase shifted by ϕ. Depending on the initial angle α of **E** and the length of the crystal, which determines the total phase shift ϕ through the plate, the emerging beam can have its initial linear polarization rotated, or changed into an elliptically or circularly polarized light as summarized in Figure 9.39.

If L is the thickness of the plate, then the o-wave experiences a phase change given by $k_{o\text{-wave}}L$ through the plate where $k_{o\text{-wave}}$ is the wavevector of the o-wave; $k_{o\text{-wave}} = (2\pi/\lambda)n_o$, where λ is the free-space wavelength. Similarly, the e-wave experiences a phase change $(2\pi/\lambda)n_eL$ through the plate. Thus, the phase difference ϕ between the orthogonal components $E_\perp$ and $E_\parallel$ of the emerging beam is

Relative phase through retarder plate

$$\phi = \frac{2\pi}{\lambda}(n_e - n_o)L \qquad [9.81]$$

The phase difference ϕ expressed in terms of full wavelengths is called the **retardation** of the plate. For example, a phase difference ϕ of 180° is a half-wavelength retardation.

The polarization of the exiting-beam depends on the crystal-type, $(n_e - n_o)$, and the plate thickness L. We know that depending on the phase difference ϕ between the orthogonal components of the field, the EM wave can be linearly, circularly, or elliptically polarized.

A **half-wave plate retarder** has a thickness L such that the phase difference ϕ is π or 180° corresponding to a half wavelength ($\lambda/2$) of retardation. The result is that $E_\parallel$ is delayed by 180° with respect to $E_\perp$. If we add the emerging $E_\perp$ and $E_\parallel$ with this phase shift ϕ, **E** would be at an angle $-\alpha$ to the optic axis and still linearly polarized. **E** has been rotated counterclockwise through 2α.

A quarter-wave plate retarder has a thickness L such that the phase difference ϕ is $\pi/2$ or 90°, corresponding to a quarter wavelength $\frac{1}{4}\lambda$. If we add the emerging $E_\perp$ and $E_\parallel$ with this phase shift ϕ, the emerging light will be elliptically polarized if $0 < \alpha < 45°$ and circularly polarized if $\alpha = 45°$.

EXAMPLE 9.19

QUARTZ HALF-WAVE PLATE What should be the thickness of a half-wave quartz plate for a wavelength $\lambda \approx 707$ nm given the extraordinary and ordinary refractive indices are $n_o = 1.541$ and $n_e = 1.549$?

SOLUTION

Half-wavelength retardation is a phase difference of π, so from Equation 9.81

$$\phi = \frac{2\pi}{\lambda}(n_e - n_o)L = \pi$$

giving

$$L = \frac{\frac{1}{2}\lambda}{(n_e - n_o)} = \frac{\frac{1}{2}(707 \times 10^{-9}\ \text{m})}{(1.549 - 1.541)} = 44.2\ \mu\text{m}$$

This is roughly the thickness of a sheet of paper.

9.17 OPTICAL ACTIVITY AND CIRCULAR BIREFRINGENCE

When a linearly polarized light wave is passed through a quartz crystal along its optic axis, it is observed that the emerging wave has its E-vector (plane of polarization) rotated, which is illustrated in Figure 9.40. This rotation increases continuously with the distance traveled through the crystal (about 21.7° per mm of quartz). The rotation of the plane of polarization by a substance is called **optical activity.** In very simple intuitive terms, optical activity occurs in materials in which the electron motions induced by the external electromagnetic field follows spiraling or helical paths (orbits).[19] Electrons flowing in helical paths resemble a current flowing in a coil and thus possess a magnetic moment. The optical field in light therefore induces oscillating magnetic moments which can be either parallel or antiparallel to the induced oscillating electric dipoles. Wavelets emitted from these oscillating induced magnetic and electric dipoles interfere to constitute a forward wave that has its optical field rotated either clockwise or counterclockwise.

If θ is the angle of rotation, then θ is proportional to the distance L propagated in the optically active medium as depicted in Figure 9.40. For an observer receiving the wave through quartz, the rotation of the plane of polarization may be *clockwise* (to the right) or *counterclockwise* (to the left) which are called *dextrorotatory* and *levorotatory* forms of optical activity. The structure of quartz is such that atomic arrangements spiral around the optic axis either in clockwise or counterclockwise sense. Quartz thus occurs in two distinct crystalline forms, right-handed and left-handed, which exhibit

[19] The explanation of optical activity involves examining both induced magnetic and electric dipole moments which will not be described here in detail.

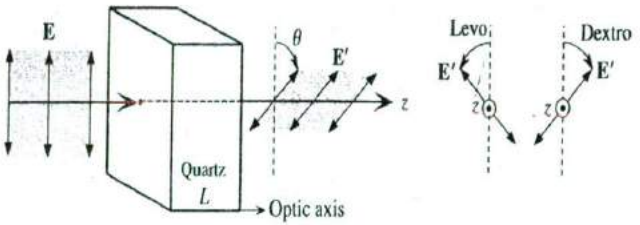


Figure 9.40 An optically active material such as quartz rotates the plane of polarization of the incident wave: The optical field **E** rotated to **E′**.
If we reflect the wave back into the material, **E′** rotates back to **E**.

dextrorotatory and levorotatory types of optical activity, respectively. Although we used quartz as an example, there are many substances that are optically active, including various biological substances and even some liquid solutions (*e.g.*, corn syrup) that contain various organic molecules with a rotatory power.

The **specific rotatory power** (θ/L) is defined as the extent of rotation per unit distance traveled in the optically active substance. Specific rotatory power depends on the wavelength. For example, for quartz this is 49° per mm at 400 nm but 17° per mm at 650 nm.

Optical activity can be understood in terms of left and right circularly polarized waves traveling at different velocities in the crystal, *i.e.*, experiencing different refractive indices. Due to the helical twisting of the molecular or atomic arrangements in the crystal, the velocity of a circularly polarized wave depends on whether the optical field rotates clockwise or counterclockwise. A vertically polarized light with a field **E** at the input can be thought of as two right- and left-handed circularly polarized waves $\mathbf{E}_R$ and $\mathbf{E}_L$ that are symmetrical with respect to the y axis, *i.e.*, at any instant $\alpha = \beta$, as shown in Figure 9.41. If they travel at the same velocity through the crystal, then they remain symmetrical with respect to the vertical ($\alpha = \beta$ remains the same) and the resultant is still a vertically polarized light. If, however, these travel at different velocities through a medium, then at the output $\mathbf{E}'_L$ and $\mathbf{E}'_R$ are no longer symmetrical with respect to the vertical, $\alpha' \neq \beta'$, and their resultant is a vector $\mathbf{E}'$ at an angle θ to the y axis.

Suppose that n_R and n_L are the refractive indices experienced by the right- and left-handed circularly polarized light, respectively. After traversing the crystal length L, the phase difference between the two optical fields $\mathbf{E}'_R$ and $\mathbf{E}'_L$ at the output leads to a new optical field $\mathbf{E}'$ that is **E** rotated by θ, given by

Optical activity

$$\theta = \frac{\pi}{\lambda}(n_L - n_R)L \qquad [9.82]$$

where λ is the free-space wavelength. For a left-handed quartz crystal, and for 589 nm light propagation along the optic axis, $n_R = 1.54427$ and $n_L = 1.54420$, which means θ is about 21.4° per mm of crystal.

In a **circularly birefringent** medium, the right- and left-handed circularly polarized waves propagate with different velocities and experience different refractive indices n_R and n_L. Since optically active materials naturally rotate the optical field, it is

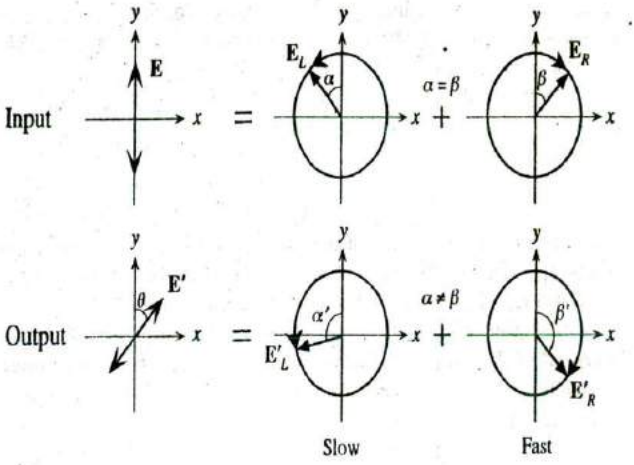


Figure 9.41 Vertically polarized wave at the input can be thought of as two right- and left-handed circularly polarized waves that are symmetrical; *i.e.*, at any instant $\alpha = \beta$. If these travel at different velocities through a medium, then at the output they are no longer symmetric with respect to y, $\alpha \neq \beta$, and the result is a vector $\mathbf{E}'$ at an angle θ to y.

not unreasonable to expect that a circularly polarized light with its optical field rotating in the same sense as the optical activity will find it easier to travel through the medium. Thus, an optically active medium possesses different refractive indices for right- and left-handed circularly polarized light and exhibits circular birefringence. It should be mentioned that if the direction of the light wave is reversed in Figure 9.40, the ray simply retraces itself and $\mathbf{E}'$ becomes $\mathbf{E}$.

ADDITIONAL TOPICS

9.18 ELECTRO-OPTIC EFFECTS[20]

Electro-optic effects refer to changes in the refractive index of a material induced by the application of an external electric field, which therefore "modulates" the optical properties. We can apply such an external field by placing electrodes on opposite faces of a crystal and connecting these electrodes to a battery. The presence of such a field distorts the electron motions in the atoms or molecules of the substance or distorts the crystal structure resulting in changes in the optical properties. For example, an applied external field can cause an optically isotropic crystal such as GaAs to become birefringent. In this case, the field induces principal axes and an optic axis. Typically changes in the refractive index are small. The frequency of the applied field has to be such that

[20] An extensive discussion and applications of the electro-optic effects may be found in S. O. Kasap, *Optoelectronics and Photonics: Principles and Practices*, Prentice Hall, 2001, Upper Saddle River, NJ, ch. 7.

the field appears static over the time scale it takes for the medium to change its properties, that is, respond, as well as for any light to cross the substance. The electro-optic effects are classified according to first- and second-order effects.

If we were to take the refractive index n to be a function of the applied electric field E, that is, $n = n(E)$, we can of course expand this as a Taylor series in E. The new refractive index n' is

Field induced refractive index

$$n' = n + a_1 E + a_2 E^2 + \cdots \quad [9.83]$$

where the coefficients a_1 and a_2 are called the *linear* electro-optic effect and *second-order* electro-optic effect coefficients. Although we would expect even higher terms in the expansion in Equation 9.83, these are generally very small and their effects negligible within the highest practical fields. The change in n due to the first E term is called the **Pockels effect.** The change in n due to the second E^2 term is called the **Kerr effect,**[21] and the coefficient a_2 is generally written as λK where K is called the Kerr coefficient. Thus, the two effects are

Pockels effect

$$\Delta n = a_1 E \quad [9.84]$$

and

Kerr effect

$$\Delta n = a_2 E^2 = (\lambda K) E^2 \quad [9.85]$$

All materials exhibit the Kerr effect. It may be thought that we will always find some (nonzero) value for a_1 for all materials, but this is not true and only certain crystalline materials exhibit the Pockels effect. If we apply a field **E** in one direction and then reverse the field and apply $-\mathbf{E}$, then according to Equation 9.84, Δn should change sign. If the refractive index increases for **E**, it must decrease for $-\mathbf{E}$. Reversing the field should *not* lead to an identical effect (the same Δn). The structure has to respond differently to **E** and $-\mathbf{E}$. There must therefore be some *asymmetry* in the structure to distinguish between **E** and $-\mathbf{E}$. In a noncrystalline material, Δn for **E** would be the same as Δn for $-\mathbf{E}$ as all directions are equivalent in terms of dielectric properties. Thus $a_1 = 0$ for all noncrystalline materials (such as glasses and liquids). Similarly, if the crystal structure has a center of symmetry, then reversing the field direction has an identical effect and a_1 is again zero. Only crystals that are **noncentrosymmetric**[22] exhibit the Pockels effect. For example, a NaCl crystal (centrosymmetric) exhibits no Pockels effect, but a GaAs crystal (noncentrosymmetric) does.

The Pockels effect expressed in Equation 9.84 is an oversimplification because in reality we have to consider the effect of an applied field along a particular crystal direction on the refractive index for light with a given propagation direction and polarization. For example, suppose that x, y, and z are the principal axes of a crystal with refractive indices n_1, n_2, and n_3 along these directions. For an optically isotropic crystal, these would be the same whereas for a uniaxial crystal such as $LiNbO_3$ $n_1 = n_2 \neq n_3$ as depicted in the xy cross section in Figure 9.42a. Suppose that we suitably apply a voltage across a crystal and thereby apply an external dc field E_a. In the Pockels effect, the field will modify the

[21] John Kerr (1824–1907) was a Scottish physicist who was a faculty member at Free Church Training College for Teachers, Glasgow (1857–1901) where he set up an optics laboratory and demonstrated the Kerr effect (1875).

[22] A crystal is a center of symmetry about a point O, if any atom (or point) with a position vector $\mathbf{r}$ from O also appears when we invert $\mathbf{r}$, that is, take $-\mathbf{r}$.

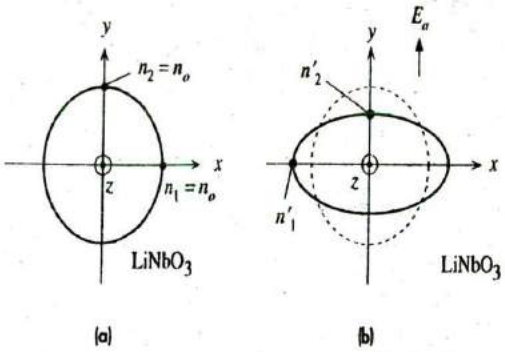


Figure 9.42
(a) Cross section of the optical indicatrix with no applied field, $n_1 = n_2 = n_o$.
(b) Applied field along y in $LiNbO_3$ modifies the indicatrix and changes n_1 and n_2 to n'_1 and n'_2.

optical indicatrix. The exact effect depends on the crystal structure. For example, a crystal like GaAs, optically isotropic with a spherical indicatrix, becomes *birefringent* with two different refractive indices. In the case of $LiNbO_3$ (lithium niobate), which is an optoelectronically important uniaxial crystal, a field E_a along the y direction changes the principal refractive indices n_1 and n_2 (both equal to n_o) to n'_1 and n'_2 as illustrated in Figure 9.42b. Moreover, in some crystals such as KDP (KH_2PO_4, potassium dihydrogen phosphate), the field E_a along z rotates the principal axes by 45° about z and changes the principal indices. Rotation of principal axes in $LiNbO_3$ is small and can be neglected.

As an example consider a wave propagating along the z direction (optic axis) in a $LiNbO_3$ crystal. This wave will experience the same refractive index ($n_1 = n_2 = n_o$) whatever the polarization as in Figure 9.42a. However, in the presence of an applied field E_a parallel to the principal y axis as in Figure 9.42b, the light propagates as two orthogonally polarized waves (parallel to x and y) experiencing different refractive indices n'_1 and n'_2. The applied field thus *induces a birefringence* for light traveling along the z axis. (The field induced rotation of the principal axes in this case, though present, is small and can be neglected.) Before the field E_a is applied, the refractive indices n_1 and n_2 are both equal to n_o. The Pockels effect then gives the new refractive indices n'_1 and n'_2 in the presence of E_a as

$$n'_1 \approx n_1 + \frac{1}{2}n_1^3 r_{22} E_a \quad \text{and} \quad n'_2 \approx n_2 - \frac{1}{2}n_2^3 r_{22} E_a \qquad [9.86]$$

Pockels effect

where r_{22} is a constant, called a **Pockels coefficient,** that depends on the crystal structure and the material. The reason for the seemingly unusual subscript notation is that there are more than one constant and these are elements of a tensor that represents the optical response of the crystal to an applied field along a particular direction with respect to the principal axes (the exact theory is more mathematical than intuitive). We therefore have to use the correct Pockels coefficients for the refractive index changes for a given crystal and a given field direction.[23] If the field were along z, the Pockels coefficient in Equation 9.86 would be r_{13}. Table 9.6 shows some typical values for Pockels coefficients of various crystals.

[23] The reader should not be too concerned with the subscripts but simply interpret them as identifying the right Pockels coefficient value for the particular electro-optic problem at hand.

Table 9.6 Pockels (r) and Kerr (K) coefficients in various materials

Material	Crystal	Indices	Pockels Coefficients $\times 10^{-12}$ m/V	Comment
$LiNbO_3$	Uniaxial	$n_o = 2.272$ $n_e = 2.187$	$r_{13} = 8.6$; $r_{33} = 30.8$ $r_{22} = 3.4$; $r_{51} = 28$	$\lambda \approx 500$ nm
KDP	Uniaxial	$n_o = 1.512$ $n_e = 1.470$	$r_{41} = 8.8$; $r_{63} = 10.5$	$\lambda \approx 546$ nm
GaAs	Isotropic	$n_o = 3.6$	$r_{41} = 1.5$	$\lambda \approx 546$ nm

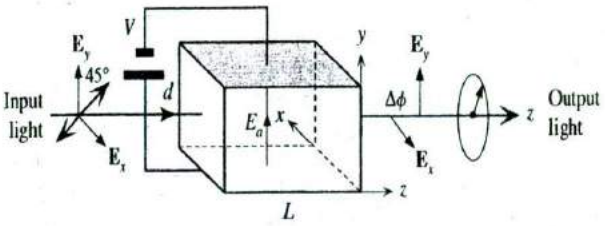


Figure 9.43 Transverse Pockels cell phase modulator. A linearly polarized input light into an electro-optic crystal emerges as a circularly polarized light.

It is clear that the control of the refractive index by an external applied field (and hence a voltage) is a distinct advantage that enables the phase change through a Pockels crystal to be controlled or modulated; such a **phase modulator** is called a *Pockels cell*. In the *longitudinal Pockels cell phase modulator* the applied field is in the direction of light propagation, whereas in the *transverse phase modulator* the applied field is transverse to the direction of light propagation.

Consider the transverse phase modulator in Figure 9.43. In this example, the applied electric field $E_a = V/d$ is applied parallel to the y direction, normal to the direction of light propagation along z. Suppose that the incident beam is linearly polarized (shown as **E**) say at 45° to the y axes. We can represent the incident light in terms of polarizations ($\mathbf{E}_x$ and $\mathbf{E}_y$) along the x and y axes. These components $\mathbf{E}_x$ and $\mathbf{E}_y$ experience refractive indices n_1' and n_2', respectively. Thus, when $\mathbf{E}_x$ traverses the distance L, its phase changes by ϕ_1,

$$\phi_1 = \frac{2\pi n_1'}{\lambda} L = \frac{2\pi L}{\lambda}\left(n_o + \frac{1}{2} n_o^3 r_{22} \frac{V}{d}\right)$$

When the component $\mathbf{E}_y$ traverses the distance L, its phase changes by ϕ_2, given by a similar expression except that r_{22} changes sign. Thus the phase change $\Delta\phi$ between the two field components is

Transverse Pockels effect

$$\Delta\phi = \phi_1 - \phi_2 = \frac{2\pi}{\lambda} n_o^3 r_{22} \frac{L}{d} V \qquad [9.87]$$

The applied voltage thus inserts an adjustable phase difference $\Delta\phi$ between the two field components. The polarization state of the output wave can therefore be controlled by the applied voltage and the Pockels cell is a **polarization modulator.** We can change the medium from a quarter-wave to a half-wave plate by simply adjusting V. The voltage $V = V_{\lambda/2}$, the **half-wave voltage,** corresponds to $\Delta\phi = \pi$ and generates a half-wave plate.

CD Selected Topics and Solved Problems

Selected Topics	Solved Problems
Real and Imaginary Dielectric Constant	Fresnel's Equations
Optical Dispersion and Absorption	Complex Refractive Index and Light Absorption
	Dispersion: Refractive Index versus Wavelength Behavior

DEFINING TERMS

Absorption is the loss in the power of electromagnetic radiation that is traveling in a medium. The loss is due to the conversion of light energy to other forms of energy, such as lattice vibrations (heat) during the polarization of the molecules of the medium, local vibrations of impurity ions, excitation of electrons from the valence band to the conduction band, and so on.

Activator is a luminescent center in a host crystal or glass in which it is excited, by some external excitation such as UV light; following excitation, the activator emits radiation to return to its ground state, or become de-excited.

Anisotropy (optical) refers to the fact that the refractive index n of a crystal depends on the direction of propagation of light and on the state of its polarization, that is, the direction of the electric field.

Antireflection coating is a thin dielectric layer coated on an optical device or component to reduce the reflection of light and increase the transmitted light intensity.

Attenuation is the decrease in the optical power (or irradiance) of a traveling wave in the direction of propagation due to absorption and scattering.

Attenuation coefficient α represents the spatial rate of attenuation of an EM wave along the direction of propagation. If P_o is the optical power at some location O, and if it is P at a distance L from O along the direction of propagation, then $P = P_o \exp(-\alpha L)$.

Birefringent crystals such as calcite are optically anisotropic which leads to an incident light beam becoming separated into ordinary and extraordinary waves with orthogonal polarizations; incident light becomes doubly refracted because these two waves experience different refractive indices n_o and n_e.

Brewster's angle or **polarization angle** (θ_p) is the angle of incidence that results in the reflected wave having no electric field in the plane of incidence (plane defined by the incident ray and the normal to the surface). The electric field oscillations in the reflected wave are in the plane perpendicular to the plane of incidence.

Circularly birefringent medium is a medium in which right and left circularly polarized waves propagate with different velocities and experience different refractive indices n_R and n_L.

Circularly polarized light is light where the magnitude of the field vector **E** remains constant but its tip at a given location on the direction of propagation traces out a circle by rotating either in a clockwise sense, *right circularly polarized,* with time, as observed by the

receiver of the wave, or in a counterclockwise sense, *left circularly polarized.*

Complex propagation constant $(k' - jk'')$ describes the propagation characteristics of an electromagnetic wave that is experiencing attenuation as it travels in a lossy medium. If $k = k' - jk''$ is the complex propagation constant, then the electric field component of a plane wave traveling in a lossy medium can be described by

$$E = E_o \exp(-k''z) \exp j(\omega t - k'z)$$

The amplitude decays exponentially while the wave propagates along z. The *real* k' part of the complex propagation constant (wavevector) describes the propagation characteristics, that is, the phase velocity $v = \omega/k'$. The *imaginary* k'' part describes the rate of attenuation along z.

Complex refractive index N with real part n and imaginary part K is defined as the ratio of the complex propagation constant k in a medium to propagation constant k_o in free space,

$$\mathrm{N} = n - jK = \frac{k}{k_o} = \left(\frac{1}{k_o}\right)(k' - jk'')$$

The real part n is simply called the refractive index, and K is called the *extinction coefficient.*

Critical angle (θ_c) is the angle of incidence that results in a refracted wave at 90° when the incident wave is traveling in a medium of lower refractive index and is incident at a boundary with a material with a higher refractive index.

Dielectric mirror is made from alternating high and low refractive index quarter-wave-thick multilayers such that constructive interference of partially reflected waves gives rise to a high degree of wavelength-selective reflectance.

Dispersion relation is a *relationship* between the refractive index n and the wavelength λ of the EM wave, $n = n(\lambda)$; the wavelength usually refers to the free-space wavelength. The relationship between the angular frequency ω and the propagation constant k, the ω–k curve, is also called the dispersion relation.

Dispersive medium has a refractive index n that depends on the wavelength; that is, n is not a constant.

Electro-optic effects refer to changes in the refractive index of a material induced by the application of an external electric field, which therefore "modulates" the optical properties; the applied field is not the electric field of any light wave, but a separate external field.

Extinction coefficient is the imaginary part of the complex refractive index N.

Fluorescence is luminescence that occurs over very short time scales, usually less than 10^{-8} seconds (or 10 ns). In fluorescence, the onset and decay of luminescent emission, due to the onset and cessation of excitation of the phosphor, is very short, appearing to be almost instantaneous.

Fresnel's equations describe the amplitude and phase relationships between the incident, reflected, and transmitted waves at a dielectric–dielectric interface in terms of the refractive indices of the two media and the angle of incidence.

Group index (N_g) represents the factor by which the group velocity of a group of waves in a dielectric medium is reduced with respect to propagation in free space, $N_g = c/v_g$ where v_g is the group velocity.

Group velocity (v_g) is the velocity at which energy, or information, is transported by a group of waves; v_g is determined by $d\omega/dk$ whereas phase velocity is determined by ω/k.

Instantaneous irradiance is the instantaneous flow of energy per unit time per unit area and is given by the instantaneous value of the Poynting vector **S**.

Irradiance (average) is the average flow of energy per unit time per unit area where averaging is typically carried out by the light detector (over many oscillation periods). Average irradiance can also be defined mathematically by the average value of the Poynting vector **S**. The *instantaneous irradiance* can only be measured if the power meter can respond more quickly than the oscillations of the electric field, and since this is in the optical frequencies range, all practical measurements invariably yield the average irradiance.

Kerr effect is a second-order effect in which the change in the refractive index n depends on the square of the electric field, that is, $\Delta n = a_2 E^2$, where a_2 is a material dependent constant.

Kramers–Kronig relations relate the real and imaginary parts of the relative permittivity. If we know the complete frequency dependence of the real part $\varepsilon_r'(\omega)$,

using the Kramer–Kronig relation, we can find the frequency dependence of the imaginary part $\varepsilon_r''(\omega)$.

Luminescence is the emission of light by a material, called a **phosphor,** due to the absorption and conversion of energy into electromagnetic radiation. Typically the emission of light occurs from certain dopant impurities or even defects, called **luminescent** or **luminescence centers** or **activators** purposefully introduced into a **host matrix,** which may be a crystal or glass, which can accept the activators. **Photoluminescence** involves excitation by photons (light). **Cathodoluminescence** is light emission when the excitation is the bombardment of the phosphor with energetic electrons as in TV cathode ray tubes. **Electroluminescence** is light emission due to the passage of an electric current as in the LED.

Optic axis is an axis in the crystal structure along which there is no double refraction for light propagation along this axis.

Optical activity is the rotation of the plane of polarization of plane polarized light by a substance such as quartz.

Optical indicatrix (Fresnel's ellipsoid) is a refractive index surface placed in the center of the principal axes x, y, and z of a crystal; the axis intercepts are n_1, n_2, and n_3. We can represent the optical properties of a crystal in terms of three refractive indices along three orthogonal axes, the *principal axes* of the crystal, x, y, and z.

Phase of a traveling wave is the quantity $(kx - \omega t)$ which determines the amplitude of the wave at position x and at time t given the propagation constant $k(= 2\pi/\lambda)$ and angular frequency ω. In three dimensions it is the quantity $(\mathbf{k} \cdot \mathbf{r} - \omega t)$ where $\mathbf{k}$ is the wavevector and $\mathbf{r}$ is the position vector.

Phase velocity is the rate at which a given phase on a traveling wave advances. It represents the velocity of a given phase rather than the velocity at which information is carried by the wave. Two consecutive peaks of a wave are separated by a wavelength λ, and it takes a time period $1/\nu$ for one peak to reach the next (or the time separation of two consecutive peaks at one location); then the phase velocity is defined as $v = \lambda\nu$.

Phosphor is a substance made of an activator and a host matrix (crystal or glass) that exhibits luminescence upon suitable excitation.

Phosphorescence is a slow luminescence process in which luminescent emission occurs well after the cessation of excitation, even after minutes or hours.

Pockels effect is a linear change in the refractive index n of a crystal due to an application of an external electric field E, other than the field of the light wave, that is, $\Delta n = a_1 E$, where a_1 is a constant that depends on the crystal structure.

Polarization of an EM wave describes the behavior of the electric field vector in the EM wave as it propagates through a medium. If the oscillations of the electric field at all times are contained within a well-defined line, then the EM wave is said to be *linearly polarized.* The field vibrations and the direction of propagation, *e.g.*, z direction, define a *plane of polarization* (plane of vibration), so linear polarization implies a wave that is plane-polarized.

Poynting vector (S) represents the energy flow per unit time per unit area in a direction determined by $\mathbf{E} \times \mathbf{B}$ (direction of propagation), $\mathbf{S} = v^2 \varepsilon_a \varepsilon_r \mathbf{E} \times \mathbf{B}$. Its magnitude, power flow per unit area, is called the irradiance.

Principal axes of the crystal, normally labeled, x, y, and z, are special axes along which the polarization vector and the electric field are parallel. Put differently, the electric displacement D and the electric field E vectors are parallel. The refractive indices along these x, y, and z axes are the principal indices n_1, n_2, and n_3, respectively, for electric field oscillations along these directions (not to be confused with the wave propagation direction).

Reflectance is the fraction of power in the reflected electromagnetic wave with respect to the incident power.

Reflection coefficient is the ratio of the amplitude of the reflected EM wave to that of the incident wave. It can be positive, negative, or a complex number which then represents a phase change.

Refraction is a change in the direction of a wave when it enters a medium with a different refractive index. A wave that is incident at a boundary between two media with different refractive indices experiences refraction and changes direction in passing from one to the other medium.

Refractive index n of an optical medium is the ratio of the velocity of light in a vacuum to its velocity in the medium $n = c/v$.

Retarding plates are optical devices that change the state of polarization of an incident light beam. For example, when a linearly polarized light enters a *quarter-wave plate*, it emerges from the device either as circularly or elliptically polarized light, depending on the angle of the incident electric field with respect to the optic axis of the retarder plate.

Scattering is a process by which the energy from a propagating EM wave is redirected as secondary EM waves in various directions away from the original direction of propagation. There are a number of scattering processes. In Rayleigh scattering, fluctuations in the refractive index, inhomogeneities, etc., lead to the scattering of light that decreases with the wavelength as λ^4.

Snell's law is a law that relates the angles of incidence and refraction when an EM wave traveling in one medium becomes refracted as it enters an adjacent medium. If light is traveling in a medium with index n_1 is incident on a medium of index n_2, and if the angles of incidence and refraction (transmission) are θ_i and θ_t, then according to Snell's law,

$$\frac{\sin\theta_t}{\sin\theta_i} = \frac{n_2}{n_1}$$

Specific rotatory power is defined as the amount of rotation of the optical field in a linearly polarized light per unit distance traveled in the optically active substance.

Stoke's shift in luminescence is the shift down in the frequency of the emitted radiation with respect to that of the exciting radiation.

Total internal reflection (TIR) is the total reflection of a wave traveling in a medium when it is incident at a boundary with another medium of lower refractive index. The angle of incidence must be greater than the critical angle θ_c which depends on the refractive indices $\sin\theta_c > n_2/n_1$.

Transmission coefficient is the ratio of the amplitude of the transmitted wave to that of the incident wave when the incident wave traveling in a medium meets a boundary with a different medium (different refractive index).

Transmittance is the fraction of transmitted intensity when a wave traveling in a medium is incident at a boundary with a different medium (different refractive index).

Wavefront is a surface where all the points have the same phase. A wavefront on a plane wave is an infinite plane perpendicular to the direction of propagation.

Wavenumber or **propagation constant** is defined as $2\pi/\lambda$ where λ is the wavelength. It is the phase shift in the wave over a distance of unit length.

Wavepacket is a group of waves with slightly different frequencies traveling together and forming a "group." This wavepacket travels with a group velocity v_g that depends on the slope of ω versus k characteristics of the wavepacket, *i.e.*, $v_g = d\omega/dk$.

Wavevector is a vector denoted as **k** that describes the direction of propagation of a wave and has the magnitude of the wavenumber, $k = 2\pi/\lambda$.

QUESTIONS AND PROBLEMS

9.1 **Refractive index and relative permittivity** Using $n = \sqrt{\varepsilon_r}$, calculate the refractive index n of the materials in the table given their low-frequency relative permittivities ε_r (LF). What is your conclusion?

	Material			
	a-Se	Ge	NaCl	MgO
ε_r (LF)	6.4	16.2	5.90	9.83
n (~ 1–5 μm)	2.45	4.0	1.54	1.71

9.2 **Refractive index and bandgap** Diamond, silicon, and germanium all have the same diamond unit cell. All three are covalently bonded solids. Their refractive indices (n) and energy bandgaps (E_g) are shown in the table. (*a*) Plot n versus E_g and (*b*) plot n^4 versus $1/E_g$. What is your conclusion? According to **Moss's rule**, very roughly,

$$n^4 E_g \approx K = \text{Constant}$$

Moss's rule

What is the value of K?

	Material		
	Diamond	**Silicon**	**Germanium**
Bandgap, E_g (eV)	5	1.1	0.66
n	2.4	3.46	4.0

***9.3** **Temperature coefficient of refractive index** Suppose that we could write the relationship between the refractive index n (at frequencies much less than ultraviolet light) and the bandgap E_g of a semiconductor as suggested by Hervé and Vandamme,

$$n^2 = 1 + \left(\frac{A}{E_g + B}\right)^2$$

where E_g is in eV, $A = 13.6$ eV, and $B = 3.4$ eV. (B depends on the incident photon energy.) Temperature dependence in n results from dE_g/dT and dB/dT. Show that the temperature coefficient of refractive index (TCRI) is given by.[24]

$$\text{TCRI} = \frac{1}{n}\cdot\frac{dn}{dT} = -\frac{(n^2-1)^{3/2}}{13.6n^2}\left[\frac{dE_g}{dT} + B'\right]$$

Hervé–Vandamme relationship

where B' is dB/dT. Given that $B' = 2.5 \times 10^{-5}$ eV K^{-1}, calculate TCRI for two semiconductors: Si with $n \approx 3.5$ and $dE_g/dT \approx -3 \times 10^{-4}$ eV K^{-1}, and AlAs with $n \approx 3.2$ and $dE_g/dT \approx -4 \times 10^{-4}$ eV K^{-1}.

9.4 **Sellmeier dispersion equation** Using the Sellmeier equation and the coefficients in Table 9.2, calculate the refractive index of fused silica (SiO_2) and germania (GeO_2) at 1550 nm. Which is larger, and why?

9.5 **Dispersion (n versus λ) in GaAs** By using the dispersion relation for GaAs, calculate the refractive index n and the group index N_g of GaAs at a wavelength of 1300 nm.

9.6 **Cauchy dispersion equation** Using the Cauchy coefficients and the general Cauchy equation, calculate the refractive index of a silicon crystal at 200 μm and at 2 μm, over two orders of magnitude wavelength change. What is your conclusion? Would you expect a significant chnge in n for $\hbar\omega > E_g$?

9.7 **Cauchy dispersion relation for zinc selenide** ZnSe is a II–VI semiconductor and a very useful optical material used in various applications such as optical windows (especially high-power laser windows), lenses, prisms, etc. It transmits over 0.50 to 19 μm. n in the 1–11 μm range described by a Cauchy expression of the form

$$n = 2.4365 + \frac{0.0485}{\lambda^2} + \frac{0.0061}{\lambda^4} - 0.0003\lambda^2$$

ZnSe dispersion relation

in which λ is in μm. What is ZnSe's refractive index n and group index N_g at 5 μm?

***9.8** **Dispersion (n versus λ)** Consider an atom in the presence of an oscillating electric field as in Figure 9.4. The applied field oscillates harmonically in the $+x$ and $-x$ directions and is given by $E = E_o \exp(j\omega t)$. The energy losses can be represented by a frictional force whose magnitude is proportional to the velocity

[24] P. J. L. Hervé and L. K. J. Vandamme, *J. Appl. Phys.*, 77, 5476, 1995 and references therein.

dx/dt. If γ is the proportionality constant per electron and per unit electron mass, then Newton's second law for Z electrons in the polarized atom is

$$Zm_e\frac{d^2x}{dt^2} = -ZeE_o\exp(j\omega t) - Zm_e\omega_o^2 x - Zm_e\gamma\frac{dx}{dt}$$

where $\omega_o = (\beta/Zm_e)^{1/2}$ is the **natural frequency** of the system composed of Z electrons and a $+Ze$ nucleus and β is a force constant for the restoring Coulombic force between the electrons and the nucleus. Show that the electronic polarizability α_e is

Electronic polarizability

$$\alpha_e = \frac{p_{\text{induced}}}{E} = \frac{Ze^2}{m_e(\omega_o^2 - \omega^2 + j\gamma\omega)}$$

What does a complex polarizability represent? Since α_e is a complex quantity, so is ε_r and hence the refractive index. By writing the complex refractive index $N = \sqrt{\varepsilon_r}$ where ε_r is related to α_e by the Clausius–Mossotti equation, show that

Complex refractive index

$$\frac{N^2 - 1}{N^2 + 2} = \frac{NZe^2}{3\varepsilon_o m_e(\omega_o^2 - \omega^2 + j\gamma\omega)}$$

where N is the number of atoms per unit volume. What are your conclusions?

9.9 **Dispersion and diamond** Consider applying the simple electronic polarizability and Clausius–Mossotti equations to diamond. Neglecting losses,

$$\alpha_e = \frac{Ze^2}{m_e(\omega_o^2 - \omega^2)}$$

and

Dispersion in diamond

$$\frac{\varepsilon_r - 1}{\varepsilon_r + 2} = \frac{NZe^2}{3\varepsilon_o m_e(\omega_o^2 - \omega^2)}$$

For diamond we can take $Z = 4$ (valence electrons only as these are the most responsive), $N = 1.8 \times 10^{29}$ atoms m^{-3}, $\varepsilon_r(\text{DC}) = 5.7$. Find ω_o and then find the refractive index at $\lambda = 0.5$ μm and 5 μm.

9.10 **Electric and magnetic fields in light** The intensity (irradiance) of the red laser beam from a He–Ne laser in air has been measured to be about 1 mW cm^{-2}. What are the magnitudes of the electric and magnetic fields? What are the magnitudes if this 1 mW cm^{-2} beam were in a glass medium with a refractive index $n = 1.45$ and still had the same intensity?

9.11 **Reflection of light from a less dense medium (internal reflection)** A ray of light which is traveling in a glass medium of refractive index $n_1 = 1.450$ becomes incident on a less dense glass medium of refractive index $n_2 = 1.430$. Suppose that the free-space wavelength (λ) of the light ray is 1 μm.

a. What should be the minimum incidence angle for TIR?

b. What is the phase change in the reflected wave when $\theta_i = 85°$ and when $\theta_i = 90°$?

c. What is the penetration depth of the evanescent wave into medium 2 when $\theta_i = 85°$ and when $\theta_i = 90°$?

9.12 **Internal and external reflection at normal incidence** Consider the reflection of light at normal incidence on a boundary between a GaAs crystal medium of refractive index 3.6 and air of refractive index 1.

a. If light is traveling from air to GaAs, what is the reflection coefficient and the intensity of the reflected light in terms of the incident light?

b. If light is traveling from GaAs to air, what is the reflection coefficient and the intensity of the reflected light in terms of the incident light?

9.13 **Antireflection coating**

a. Consider three dielectric media with flat and parallel boundaries with refractive indices n_1, n_2, and n_3. Show that for normal incidence the reflection coefficient between layers 1 and 2 is the same as that between layers 2 and 3 if $n_2 = \sqrt{n_1 n_3}$. What is the significance of this?

b. Consider a Si photodiode that is designed for operation at 900 nm. Given a choice of two possible antireflection coatings, SiO_2 with a refractive index of 1.5 and TiO_2 with a refractive index of 2.3, which would you use and what would be the thickness of the antireflection coating you chose? The refractive index of Si is 3.5.

9.14 **Optical fibers in communications** Optical fibers for long-haul applications usually have a core region that has a diameter of about 10 μm, and the whole fiber would be about 125 μm in diameter. The core and cladding refractive indices, n_1 and n_2, respectively, are normally only 0.3–0.5 percent different. Consider a fiber with n_1(core) = 1.4510, and n_2(cladding) = 1.4477, both at 1550 nm. What is the maximum angle that a light ray can make with the fiber axis if it is still to propagate along the fiber?

9.15 **Optical fibers in communications** Consider a short-haul optical fiber that has n_1(core) = 1.455 and n_2(cladding) = 1.440 at 870 nm. Assume the core-cladding interface behaves like the flat interface between two infinite media as in Figure 9.11. Consider a ray that is propagating that has an angle of incidence 85° at the core–cladding interface. Can this ray exercise total internal reflection? What would be its penetration depth into the cladding?

9.16 **Complex refractive index** Spectroscopic ellipsometry measurements on a silicon crystal at a wavelength of 620 nm show that the real and imaginary parts of the complex relative permittivity are 15.2254 and 0.172, respectively. Find the complex refractive index. What is the reflectance and absorption coefficient at this wavelength? What is the phase velocity?

Free carrier absorption

9.17 **Complex refractive index** Spectroscopic ellipsometry measurements on a germanium crystal at a photon energy of 1.5 eV show that the real and imaginary parts of the complex relative permittivity are 21.56 and 2.772, respectively. Find the complex refractive index. What is the reflectance and absorption coefficient at this wavelength? How do your calculations match with the experimental values of $n = 4.653$ and $K = 0.298$, $R = 0.419$ and $\alpha = 4.53 \times 10^6\ \text{m}^{-1}$?

9.18 An *n*-type germanium sample has a conductivity of about $300\ \Omega^{-1}\text{m}^{-1}$. Calculate the imaginary part ε_r'' of the relative permittivity at a wavelength of 20 μm. Find the attenuation coefficient α due to free carrier absorption. The refractive index of germanium at the specified wavelength is $n = 4$.

9.19 **Reststrahlen absorption in CdTe** Figure 9.22 shows the infrared extinction coefficient K of CdTe. Calculate the absorption coefficient α and the reflectance R of CdTe at 60 μm and 80 μm.

9.20 **Reststrahlen absorption in GaAs** Figure 9.22 shows the infrared extinction coefficient K of GaAs as a function of wavelength. Optical measurements show that K peaks at $\lambda = 37.1$ μm where $K \approx 11.6$ and $n \approx 6.6$. Calculate the absorption coefficient α and the reflectance R at this wavelength.

9.21 **Fundamental absorption** Consider the semiconductors in Figure 9.23, and those semiconductors listed in Table 9.3.

a. Which semiconductors can be candidates for a photodetector that can detect light in optical communications at 1550 nm?

b. For amorphous Si (a-Si), one definition of an *optical gap* is the photon energy that results in an optical absorption coefficient α of $10^4\ \text{cm}^{-1}$. What is the optical gap of a-Si in Figure 9.23?

c. Consider a solar cell from crystalline Si. What is the absorption depth of light at 1000 nm, and at 500 nm?

9.22 **Quartz half-wave plate** What are the possible thicknesses of a half-wave quartz plate for a wavelength $\lambda \approx 1.01$ μm given the extraordinary and ordinary refractive indices are $n_e = 1.534$ and $n_o = 1.543$, respectively?

9.23 **Pockels cell modulator** What should be the aspect ratio d/L for the transverse $LiNiO_3$ phase modulator in Figure 9.43 that will operate at a free-space wavelength of 1.3 μm and will provide a phase shift $\Delta\phi$ of π (half wavelength) between the two field components propagating through the crystal for an applied voltage of 20 V? The Pockels coefficient r_{22} is 3.2×10^{-12} m/V and $n_o = 2.2$.